

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Estimating global policy impacts in unprecedented contexts with interdisciplinary econometric approaches

Permalink

<https://escholarship.org/uc/item/040012x7>

Author

Bolliger, Ian Wesley

Publication Date

2020

Peer reviewed|Thesis/dissertation

Estimating global policy impacts in unprecedented contexts with interdisciplinary
econometric approaches

by

Ian Bolliger

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Energy and Resources

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Solomon Hsiang, Chair

Professor Daniel Kammen

Professor Benjamin Recht

Professor Robert Kopp

Summer 2020

Estimating global policy impacts in unprecedented contexts with interdisciplinary
econometric approaches

Copyright 2020
by
Ian Bolliger

Abstract

Estimating global policy impacts in unprecedented contexts with interdisciplinary
econometric approaches

by

Ian Bolliger

Doctor of Philosophy in Energy and Resources

University of California, Berkeley

Professor Solomon Hsiang, Chair

Global policy challenges, such as addressing climate change and improving global health, require coordinated international efforts. Enacting such policies is often hampered by difficulties in quantifying the public benefit brought about by action. Without this information, policy decisions may focus primarily on the costs of action and thus disproportionately harm the groups for which inaction enacts the most damage. While classical econometric approaches are designed to identify policy impacts, they fall short when globally representative data is missing or when present or future contexts have no clear historical precedent. In these situations, new approaches are needed to incorporate evolving global data sources, such as satellite imagery, and process-based understanding from relevant fields, such as climate science or epidemiology. This dissertation leverages recent computational advances to develop and apply scalable tools for incorporating these data and models into policy impact estimation. Specifically, the first chapter describes a novel method for quickly extracting socio-environmental measurements using satellite imagery and machine learning. The second and third chapters describe applications of interdisciplinary methods to (a) project climate-driven changes to hurricane damages and (b) estimate the effect of COVID-19 policies on the pandemic's spread. In general, the findings suggest that the previously unmeasured benefits of global action in these contexts are substantial, and they demonstrate how such interdisciplinary approaches can better inform policy-making for complex global challenges.

To my friends and family, who provide meaning, inspiration, and value for everything I do.

Contents

Contents	ii
1 Introduction	1
2 A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery	4
S.1 Methods summary	15
S.2 Data	23
S.3 Methods	30
S.4 Comparisons to other models	55
3 The Effect of Large-Scale Anti-Contagion Policies on the COVID-19 Pandemic	65
3.1 Introduction	66
3.2 Methods Summary	67
3.3 Results	68
3.4 Discussion	71
3.5 Methods	77
3.6 Data and Code Availability	86
3.7 End Notes	86
3.8 Extended Data	87
3.9 Supplementary Information	98
4 Past, Present, and Future Economic Risks from Hurricanes in the United States	149
4.1 Introduction	150
4.2 Modeling Hazard and Exposure from Historical Events	153
4.3 Estimating Wind and Coastal Flood Dose-Response Functions	155
4.4 Projecting Risk Under Climate Change	159
4.5 Discussion and Future Work	165
4.6 Methods	167

5	Implications and Future Work	185
5.1	Socioenvironmental prediction with satellite imagery and machine learning	186
5.2	Econometric-epidemiological prediction	186
5.3	Impacts of climate change on extreme event damages	187
	Bibliography	188

Acknowledgments

First, thank you to my family. Mom and dad, you've helped me navigate so many ups and downs (with a surprising number of them occurring over the last six years...) You never shy from delivering sharp advice, most of which I have foolheartedly discarded along the way only to pick up in later, wiser years.

Thank you to Rose, whose team I joined for the final stretch of this grad school experience and who has shown unending love and support. Our adventures have brought, and will continue to bring, joy to all corners of my life. For that I am incredibly grateful.

Thank you to the students, staff, and faculty of ERG and the Global Policy Lab. We've shared countless ideas, opinions, and debates as the world has changed dramatically around us. I eagerly await seeing where everyone's careers take them and hope that we can remain close even in distance. Thank you also to the students of the Earth System Science and Policy Lab. Though we interact far less regularly and live on different coasts, you very quickly accepted my digital presence and I'm very thankful to count you all among my colleagues and friends.

Thank you in particular to Kay Burns, without whom I'm quite sure no ERG student would have ever graduated. Your combination of kindness and assertion is something I have never before witnessed, and I'm glad to have been one of the many beneficiaries.

Thank you to my advisor Solomon Hsiang, whose dedication to impactful and robust research inspires and whose commitment to developing that quality in future researchers impresses all of us who are lucky enough to be your students.

Thank you to all of the other mentors and colleagues that inhabit the intellectual village that raised me. Thank you Dan Kammen for shepharding my early years at ERG, for demonstrating the value of an interdisciplinary education, and for encouraging me to dive headfirst into new opportunities. Thank you to Bob Kopp for your adept remote mentorship, which seems like a particularly valuable skill at this time. Thank you to Sally Thompson for your guidance and support through my qualifying exam. Thank you to the DS421 team for an experience that changed the course of my career. And thank you to the THIMBY team for the most rewarding and instructional experience in grad school. Nothing ever seemse quite as overwhelming since we took on that project...

Thank you to the University of California, Berkeley, the Climate Impact Lab, and the U.S. Department of Defense for supporting my research over the past several years.

Finally, thank you to all the other peers and mentors with whom I collaborated on projects large and small that do not appear in this dissertation: Ingvild Almas, Max Aufhammer, Tessa Bold, Jeff Deems, Ted Miguel, Noah Molotch, Shuhei Kitamura, Robert Pickmans, and many others. I am grateful for the research community that this experience has provided and hope that I can continue to lean on my friends, colleagues, and mentors as I try to live up to the impressive standard you all have set.

Chapter 1

Introduction

Addressing global problems, from climate change to poverty to international conflict, often requires coordinated policy and action across a large number of stakeholders. Numerous challenges hinder our ability to cooperate in solving such problems, including political and ideological differences and competing incentives. Often, these challenges are made particularly difficult by a lack of clear data on the effects that proposed action would bring. This ambiguity can facilitate status quo behavior because groups may argue that the “costs outweigh the benefits”. For example, addressing climate change necessitates investment in clean technologies in addition to large-scale structural change (UNEP 2019). Those who stand to lose from such changes may argue that their costs are larger than those of climate change itself.

The lack of quantitative information on the impacts of policy action can present a particular challenge to marginalized groups. By definition, the incentives of those who possess power in the current environment are likely to motivate maintenance of this status quo, even when implementing change would improve global welfare. Communities without power will thus be disproportionately harmed by the choice of inaction; however, without evidence to quantify this harm, it can be difficult to overcome the inertia of the status quo. In no way will an analysis of policy impact alone eliminate this canonical power dynamic. However, by providing tools with which to directly compare the state of the world in the presence and absence of policy, it can provide evidence to support efficient and effective action. In turn, this quantitative evidence can help overcome personal and political incentives that drive inaction. The metrics used in such impact analyses must themselves be appropriate for the decision at hand. However, if these metrics fully account for the welfare of marginalized groups, it can lend voice to communities whose needs are often underrepresented in policy-making.

Econometrics and causal statistics comprise a broad set of tools often employed to provide insight in these situations. By comparing current contexts to similar situations in the historical record, we can develop estimates of what a counterfactual world would look like were some action to be taken. This can help quantify the impacts of that policy with respect to metrics of societal welfare and has been employed in countless prospective and retrospective “impact analyses”. These approaches rely on the ability to construct some comparable

scenario in the historical record in which we can observe how the outcomes we care about differ across some dimension that could be affected by action. However, we increasingly deal with problems for which such analogs are not easily identified in traditional data sources.

This dissertation develops and applies interdisciplinary methods that enable the application of econometric approaches in these “low-data” contexts. My goal with this work is twofold. The first is the desire to provide actionable information to guide action on pressing global issues. The second is to develop a class of methodology that can be more broadly applied to new challenges that emerge and fit this criteria. Broadly, the classes of impact estimation problem that I seek to answer with the tools portrayed in this dissertation also fall into two classes.

The first class consists of “measurement problems”. These are problems for which a meaningful historical analog exists, but for which we lack the observational capacity to incorporate that information. A canonical example of this case is the effect of development programs on rural welfare. Collecting data from rural areas is challenging and expensive, and it thus hinders our ability to obtain the measurements that are necessary to understand the effects of a given program. Recent growth in earth observation technology provides a pathway for increasing our observational capacity. For example, multiple research efforts have sought to use satellite imagery to overcome data gaps in our knowledge of global poverty (Jean, Burke, et al. 2016; Perez et al. 2017). Presently, the generation of meaningful information from such efforts often requires a high degree of both domain expertise and computational sophistication. In Chapter 2, I develop a platform for facilitating rapid scaling of our ability to estimate socioenvironmental metrics. My goal is that this work can enable novel impact analyses in data poor regions.

The second class consists of “precedent problems”. These are problems for which we simply have never experienced an analogous event, and thus for which no amount of improved measurement could help identify the impact of a policy. A salient example is the early stages of the current COVID-19 pandemic. Modern society has never experienced a contagion like this, and thus we have had a limited understanding of how large-scale anti-contagion policies will affect the spread. Policies such as widespread school closures, for example, carry massive societal costs but could drastically limit mortality from the virus. A lack of precedent limits the applicability of traditional econometric approaches alone in identifying the trade-offs inherent in such decisions, yet approaches from other disciplines have also faced challenges. Many early epidemiological forecasts of disease burden, for example, have not proven accurate (IHME COVID-19 health service utilization forecasting team and Murray 2020). In these situations, process understanding can be merged with statistical analyses to help develop a more robust understanding of how actions will affect the overall burden of the pandemic. In Chapter 3 I demonstrate this approach, blending a retrospective econometric analysis with an epidemiological forecast model to quantify the infections that have been averted by anti-contagion policies.

Another example of this second class is the broad array of societal impacts brought about by climate change. Humanity has never lived in either a 1.5° or a 3° warmer world. Because we do not have these analogs to compare to, econometric tools alone will not be able to

estimate the avoided societal costs of limiting global warming to 1.5° . In this case, we must again leverage understanding of fundamental processes from other disciplines to augment a statistical analysis by generating estimates of these yet unrealized realities. Rather than epidemiology, here we borrow from earth system science. This interdisciplinary approach has led to the growth of climate econometrics (S. Hsiang 2016) as a discipline designed to address this exact challenge. While significant progress has been made on identifying the broad effect of increasing global temperatures, the societal impacts of changing extreme weather frequencies and intensities remains uncertain. In Chapter 4, I construct and apply a physical-econometric model to estimate the impact of climate change on 21st century hurricane damages in the U.S.

The three chapters of this dissertation share the goal of providing novel insights into key policy settings through the blending of modeling methodologies from multiple disciplines. In Chapter 2, I apply a unique machine learning algorithm to earth observation data to improve the quality of information provided by that data. In Chapter 3, I parameterize epidemiological forecasts with econometric outputs. And in Chapter 4 I blend a variety of tools from climate science, hydrodynamics, actuarial science, and econometrics. Despite this united goal, the chapters differ broadly in the context of the questions they seek to answer. Chapter 2 defines an approach to generating broadly applicable socioeconomic and environmental datasets; Chapter 3 answers a timely public health question; and Chapter 4 addresses a key unknown in our understanding of climate risks. To me, this breadth of context is an indicator of the value of these interdisciplinary statistical approaches in large-scale policy analysis.

A key finding across all of these studies is that process understanding is crucial to informing robust statistical analysis in the “low-data” yet “high-consequence” contexts I investigate. Even in Chapter 2, where I demonstrate a domain-agnostic and generalizable predictive model, the power of the method comes from its speed and the corresponding ability to thoroughly validate its performance. I find that it performs far better for some prediction tasks than others, likely indicative of the complexity of the underlying data generating processes. In the other two chapters, process modeling represents a critical step in producing estimates of the impact of global action.

An equally important finding is that statistical tools can unlock the use of process understanding in societally meaningful settings. For example, the methods I develop in Chapter 4 to empirically link modeled flooding to economic losses facilitate the use of recent advances in accurate, high-resolution hydrodynamic models. Without the tools introduced in that chapter, previous economics literature often relied on simpler statistical relationships that could not leverage contemporary advances in the physical sciences.

It is this broader ability to integrate statistical and process-driven analyses that I hope is demonstrated through this dissertation. Similarly, it is this approach that I hope to carry forward into future work that both builds from the modeling platforms I have developed and applies a similar intellectual framework to new and emerging global challenges.

Chapter 2

A Generalizable and Accessible Approach to Machine Learning with Global Satellite Imagery

Combining satellite imagery with machine learning (SIML) has the potential to address global challenges by remotely estimating socioeconomic and environmental conditions in data-poor regions, yet the resource requirements of SIML limit its accessibility and use. We show that a single encoding of satellite imagery can generalize across diverse prediction tasks (e.g. forest cover, house price, road length). Our method achieves accuracy competitive with deep neural networks at orders of magnitude lower computational cost, scales globally, delivers super-resolution predictions, and facilitates characterizations of uncertainty. Since image encodings are shared across tasks, they can be centrally computed and distributed to unlimited researchers, who need only fit a linear regression to their own ground truth data in order to achieve state-of-the-art SIML performance.

This chapter is adapted from (Rolf et al. 2020). The overall project was led by Esther Rolf, Jonathan Proctor, Tamma Carleton, and the author of this thesis. It was originally conceptualized by Solomon Hsiang and Benjamin Recht. Dr. Hsiang coordinated the writing of the submitted manuscript, along with substantial contributions from the aforementioned project leaders and Dr. Recht. In addition to overall project contributions, the author of this thesis managed and coordinated contributions to the project’s source code (<https://github.com/bolliger32/mosaiks>); performed data acquisition, processing, and analysis for the housing price prediction task; developed Figures 2, 3, S3, S6, S8, and S17; developed theory on the super-resolution applications; and adapted the analysis pipeline to run with a GPU on the reproducibility platform Code Ocean.

Introduction

Addressing complex global challenges—such as managing global climate changes, population movements, ecosystem transformations, or economic development—requires that many different researchers and decision-makers (hereafter *users*) have access to reliable, large-scale observations of many variables simultaneously. Planet-scale ground-based monitoring systems are generally prohibitively costly for this purpose, but satellite imagery presents a viable alternative for gathering globally comprehensive data, with over 700 earth observation satellites currently in orbit (Union of Concerned Scientists 2019). Further, application of machine learning is proving to be an effective approach for transforming these vast quantities of unstructured imagery data into structured estimates of ground conditions. For example, combining satellite imagery and machine learning (SIML) has enabled better characterization of forest cover (Hansen et al. 2013), land use (Inglada et al. 2017), poverty rates (Jean, Burke, et al. 2016) and population densities (Robinson, Hohman, and Dilkina 2017), thereby supporting research and decision-making. We refer to such predictions of an individual variable as a single *task*. Demand for SIML-based estimates is growing, as indicated by the large number of private service-providers specializing in predicting one or a small number of these tasks.

The resource requirements for deploying SIML technologies, however, limit their accessibility and usage. Satellite-based measurements are particularly under-utilized in low-income contexts, where the technical capacity to implement SIML may be low, but where such measurements would likely convey the greatest benefit (L. Yu et al. 2014; Haack and Ryerson 2016). For example, government agencies in low-income settings might want to understand local waterway pollution, illegal land uses, or mass migrations. SIML, however, remains largely out of reach to these and other potential users because current approaches require a major resource-intensive enterprise, involving a combination of task-specific domain knowledge, remote sensing and engineering expertise, access to imagery, customization and tuning of sophisticated machine learning architectures, and large computational resources.

To remove these barriers, we develop a new approach to SIML that enables non-experts to obtain state-of-the-art performance without manipulating imagery, using specialized computational resources, or developing a complex prediction procedure. We design a one-time, task-agnostic encoding that transforms each satellite image into a vector of variables (hereafter *features*). We then show that these features (x) perform well at predicting ground conditions (y) across diverse tasks, using only a linear regression implemented on a personal computer. Prior work has similarly sought an unsupervised encoding of satellite imagery (Romero, Gatta, and Camps-Valls 2016; Cheriyyadat 2014; Penatti et al. 2015; Jean, S. Wang, et al. 2019); however, to the best of our knowledge, we are the first to develop a single set of features that achieves performance competitive with deep-learning methods across a variety of tasks.

We focus here on the problem of predicting properties of small regions (e.g. average house price), using high-resolution daytime satellite imagery as the only input. We develop a simple yet high-performing system that is tailored to address the challenges and opportu-

nities specific to SIML applications, taking a fundamentally different approach from leading designs. We achieve large computational gains in model training and testing, relative to leading deep neural networks, through algorithmic simplifications that take advantage of the fact that satellite images are collected from a fixed distance and viewing angle and capture repeating patterns and objects. This contrasts with deep-learning approaches to SIML that use techniques originally developed for natural images (e.g. photos taken from handheld cameras), where inconsistency in many key factors, such as subject or camera perspective, require complex solutions that our results suggest are mostly unnecessary for SIML applications.

A key contribution of our method is the ability to encode rich image information in a single set of general-purpose features. By separating this feature construction from model-fitting, our approach dramatically increases computational speed for any given researcher and delivers large computational gains at the research-system level by reorganizing how imagery is processed and distributed. Traditionally, hundreds or thousands of researchers use the same images to solve different and unrelated tasks (e.g. Fig. 2.1A). Our approach allows common sources of imagery to be converted into one centralized set of features that can be accessed by many researchers, each solving different tasks. This isolates future users from the costly steps of obtaining, storing, manipulating, and processing imagery themselves. The magnitude of the resulting benefits grow with the size of the expanding SIML user community and the scale of global imagery data, which currently increases by more than 80TB/day (Littlepage 2017).

Multi-task Observation using Satellite Imagery & Kitchen Sinks

Our objective is to enable any user with basic resources to predict ground conditions using only satellite imagery and a limited sample of task-specific ground truth data which they possess. Our SIML system, “Multi-task Observation using Satellite Imagery and Kitchen Sinks” (MOSAICS, see Supplementary Materials S.1), makes SIML accessible and generalizable by separating the prediction procedure into two independent steps: a fixed “featurization step” which translates satellite imagery into succinct vector representations ($images \rightarrow x$), and a “regression step” which learns task-specific coefficients that map these features to outcomes for a given task ($x \rightarrow y$). The featurization step can be centrally executed once, producing one set of outputs that are used to solve many different tasks through repeated application of the regression step by multiple independent users (Fig. 2.1B). Because the regression step is computationally efficient, MOSAICS scales nearly costlessly across unlimited users and tasks.

The *accessibility* of our approach stems from the simplicity and computational efficiency of the regression step for potential users, given features which are already computed once and stored centrally (Fig. 2.1B). To generate SIML predictions, a user of MOSAICS (i) queries

these tabular data for a vector of K features for each of their N locations of interest; (ii) merges these features x with label data y , i.e. the user’s independently collected ground truth data; (iii) implements a linear regression of y on x to obtain coefficients β – below, we use ridge regression; (iv) uses coefficients β and features x to predict labels $\hat{y}$ in new locations where imagery and features are available but ground truth data are not.

The *generalizability* of our approach means that a single mathematical summary of satellite imagery (x) performs well across many prediction tasks ($y_1, y_2, \dots$) without any task-specific modification to the procedure. The success of this generalizability relies on how images are encoded as features. We design a featurization function by building on the theoretically grounded machine learning concept of “random kitchen sinks” (Rahimi and Benjamin Recht 2008b), which we apply to satellite imagery by constructing “random convolutional features” (RCFs) (Fig. 2.1C, Supplementary Materials S.1). RCFs are suitable for the structure of satellite imagery and have established performance encoding genetic sequences, (Morrow et al. 2017) classifying photographs, (Coates and Andrew Y. Ng 2012) and predicting solar flares (Jonas et al. 2018) (see Supplementary Materials Section S.3.3). RCFs capture a flexible measure of similarity between every sub-image across every pair of images without using contextual or task-specific information. The regression step in MOSAIKS then treats these features x as an overcomplete basis for predicting any y , which may be a nonlinear function of image elements (see Supplementary Materials Section S.1).

In contrast to alternative approaches, MOSAIKS does not require training or using the output of a deep neural network, and encoding images into features requires no labels. Nonetheless, MOSAIKS achieves competitive performance at a large computational advantage that grows linearly with the number of SIML users and tasks, due to shared computation and storage. Below, we show that MOSAIKS achieves a practical level of generalization in real world contexts.

Results

We design a battery of experiments to test whether and under what settings MOSAIKS can provide access to high-performing, computationally-efficient, global-scale SIML predictions. Specifically, we 1) demonstrate generalization across tasks, and compare MOSAIKS’s performance and cost to existing state-of-the-art SIML models; 2) assess its performance when data are limited and when predicting far from observed labels; 3) scale the analysis to make global predictions and try recreating the results of a national survey; and 4) detail additional properties of MOSAIKS, such as the ability to make super-resolution predictions.

Generalization across tasks

We first test whether MOSAIKS achieves a practical level of generalization by applying it to a diverse set of pre-selected tasks in the United States (US). While many applications of interest for SIML are in remote and/or data-limited environments where ground-truth

may be unavailable or inaccurate, systematic evaluation and validation of SIML methods are most reliable in well-observed and data-rich environments (Blumenstock 2018).

Multi-task performance of MOSAIKS in the US We sample daytime images using the Google Static Maps API from across the continental US ($N = 100,000$), each covering $\sim 1\text{km} \times 1\text{km}$ (256-by-256 pixels) (Supplementary Materials Sections S.3.1-S.3.2). We first implement the featurization step, passing these images through MOSAIKS’ feature extraction algorithm to produce $K = 8,192$ features per image (Supplementary Materials Section S.3.3). Using only the resulting matrix of features ($\mathbf{X}$), we then repeatedly implement the regression step by solving a cross-validated ridge regression for each task and predict forest cover ($R^2 = 0.91$), elevation ($R^2 = 0.68$), population density ($R^2 = 0.72$), nighttime lights ($R^2 = 0.85$), average income ($R^2 = 0.45$), total road length ($R^2 = 0.53$), and average house price ($R^2 = 0.52$)¹ in a holdout test sample (Fig. 2.2, Table S2, Supplementary Materials Sections S.3.4-S.3.6). Computing the feature matrix $\mathbf{X}$ from imagery took less than 2 hours on a cloud computing node (Amazon EC2 p3.2xlarge instance, Tesla V100 GPU). Subsequently, solving a cross-validated ridge regression for each task took 6.8 minutes to compute on a local workstation with ten cores (Intel Xeon CPU E5-2630) (Supplementary Materials Section S.4.2). These seven outcomes are not strongly correlated with one another (Fig. S2) and no attempted tasks in this experiment are omitted. These results indicate that MOSAIKS is skillful for a wide range of possible applications without changing the procedure or features and without task-specific expertise.

Comparison to state-of-the-art SIML approaches We contextualize this performance by comparing MOSAIKS to existing deep-learning based SIML approaches. First, we retrain end-to-end a commonly-used deep convolutional neural network (CNN) architecture (He et al. 2016) (ResNet-18) using identical imagery and labels for the seven tasks above. This training took 7.9 hours per task on a cloud computing node (Amazon EC2 p3.xlarge instance, Tesla V100 GPU). We find that MOSAIKS exhibits predictive accuracy competitive with the CNN for all seven tasks (mean $R_{CNN}^2 - R_{MOSAIKS}^2 = 0.04$; smallest $R_{CNN}^2 - R_{MOSAIKS}^2 = -0.03$ for housing; largest $R_{CNN}^2 - R_{MOSAIKS}^2 = 0.12$ for elevation) in addition to being approximately 250 to $10,000\times$ faster to train, depending on whether the regression step is performed on a laptop (2018 Macbook Pro) or on the same cloud computing node used to train the CNN (Fig. 2.3A, Supplementary Materials Section S.4.1 and Table S5).

Second, we apply “transfer learning” (Pan and Qiang Yang 2010) using the ResNet-152 CNN pre-trained on natural images to featurize the same satellite images. We then apply ridge regression to the CNN-derived features. The speed of this approach is similar to MOSAIKS, but its performance is dramatically lower on all seven tasks (Fig. 2.3A, Supplementary Materials Section S.4.1).

¹Performance observed for housing using our published data will be higher ($R^2 = 0.60$) because privacy concerns mandate the withholding of a subset of this data (see Supplementary Materials Section S.2.1).

Third, we compare MOSAIKS to an approach from prior studies (Jean, Burke, et al. 2016; Xie et al. 2016; Head et al. 2017) where a deep CNN (VGG16 (Simonyan and Zisserman 2015) pretrained on the ImageNet dataset) is trained end-to-end on night lights and then each task is solved via transfer learning (Supplementary Materials Section S.4.1). We apply MOSAIKS to the imagery from Rwanda, Haiti, and Nepal used in ref. (Head et al. 2017) to solve all eleven development-oriented tasks they analyze. We find MOSAIKS matches prior performance across tasks in Rwanda and Haiti, and has slightly lower performance (average $\Delta R^2 = 0.08$) on tasks in Nepal (Fig. S17). The regression step of this transfer learning approach and MOSAIKS are similarly fast, but the transfer learning approach requires country-specific retraining of the CNN, limiting its accessibility and reducing its generalizability.

Together, these three experiments illustrate that with a single set of task-independent features, MOSAIKS predicts outcomes across a diverse set of tasks, with performance and speed that favorably compare to existing SIML approaches. However, throughout this set of experiments, we find that some sources of variation in labels are not recovered by MOSAIKS. For example, extremely high elevations ($>3,000\text{m}$) are not reliably distinguished from high elevations ($2,400\text{-}3,000\text{m}$) that appear visually similar (Fig. S9). Additionally, roughly half the variation in incomes and housing prices is unresolved, presumably because they depend on factors not observable from orbit, such as tax policies or school districts (Fig. 2.2). We also show that patterns of predictability across tasks are strikingly similar across MOSAIKS and alternative SIML approaches (Figs. S16 and S17). Together, these findings are consistent with the hypothesis that there exists some performance ceiling for each task, due to some factors not being observable from satellite imagery.

Evaluation of model sensitivity

There is growing recognition that understanding the accuracy, precision, and limits of SIML predictions is important, since consequential decisions increasingly depend on these outputs, such as which households should receive financial assistance (Athey 2017; Blumensstock 2018). However, historically, the high costs of training deep-learning models have generally prevented the stress-testing and bench-marking that would ensure accuracy and constrain uncertainty. To characterize the performance of MOSAIKS we test its sensitivity to the number of features (K) and training observations (N), as well as the extent of spatial extrapolation.

Changes to training data Unlike some featurization methods, there is no measure of “importance” for individual features in MOSAIKS so the computational complexity of the regression step can be manipulated by simply including more or fewer features. Repeatedly re-solving the linear regression step in MOSAIKS with a varied number of features indicates that increasing K above 1,000 features provides minor predictive gains (Fig. 2.3B). A majority of the observable signal in the baseline experiment using $K = 8,192$ is recovered using $K = 200$ (min 55% for income, max 89% for nighttime lights), reducing each 65,536-

pixel tri-band image to just 200 features ($\sim 250\times$ data compression). Similarly, re-solving MOSAIKS predictions with a different number of training observations demonstrates that models trained with fewer samples may still exhibit high accuracy (Fig. 2.3B). A majority of the available signal is recovered for many outcomes using only $N = 500$ (55% for road length to 87% for forest cover), with the exception of income (28%) and housing price (26%) tasks, which require larger samples. Together, these experiments suggest that users with computational, data acquisition, or data storage constraints can easily tailor MOSAIKS to match available resources and can reliably estimate the performance impact of these alterations (Supplementary Material Section S.3.7).

Spatial cross-validation To systematically evaluate the ability of MOSAIKS to make accurate predictions in large contiguous areas where labels are not available, we conduct a spatial cross-validation experiment by partitioning the US into a checkerboard pattern (Fig. 2.3C), training on the “black squares” and testing on the “white squares” (Supplementary Materials Section S.3.8). Increasing the width of squares (δ) in the checkerboard increases the average distances between train and test observations, simulating increasingly large spatial extrapolations. We find that for three of seven tasks (forest cover, population density, and nighttime lights), performance declines minimally regardless of distance (maximum R^2 decline of 10% at $\delta = 16^\circ$ for population density). For income, road length, and housing price, performance falls moderately at small degrees of spatial extrapolation (19%, 33%, and 35% decline at $\delta = 4^\circ$, respectively), but largely stabilizes thereafter. Note that the poor performance of road length predictions is possibly due to missing labels and data quality (Supplementary Materials Section S.2.1 and Fig. S1). Finally, elevation exhibits steady decline with increasing distances between training and testing data (49% decline at $\delta = 16^\circ$).

To contextualize this performance, we compare MOSAIKS to spatial interpolation of observations, a widely used approach to fill in regions of missing data (Supplementary Materials Section S.3.8). Using the same samples, MOSAIKS substantially outperforms spatial interpolation (Fig. 2.3C, grey dashed lines) across all tasks except for elevation, where interpolation performs almost perfectly over small ranges ($\delta = 0.5^\circ : R^2 = 0.95$), and housing price, where interpolation slightly outperforms MOSAIKS at small ranges. For both, interpolation performance converges to that of MOSAIKS over larger distances. Thus, in addition to generalizing across tasks, MOSAIKS generalizes out-of-sample across space, outperforming spatial interpolation of ground-truth in 5 of 7 tasks.

The above sensitivity tests are enabled by the speed and simplicity of training MOSAIKS. These computational gains also enable quantification of uncertainty in model performance within each diagnostic test. As demonstrated by the shaded bands in Figs. 2.3B-C, uncertainty in MOSAIKS performance due to variation in splits of training-validation data remains modest under most conditions.

Applications

Having evaluated MOSAIKS systematically in the data-rich US, we test its performance at planetary scale and its ability to recreate results from a national survey.

Global observation We test the ability of MOSAIKS to scale globally using the four tasks for which global labels are readily available. Using a random sub-sample of global land locations (training and validation: $N = 338,781$, test: $N = 84,692$; Supplementary Materials Section S.3.10), we construct the first planet-scale, multi-task estimates using a single set of label-independent features ($K = 2048$, Fig. 2.4A), predicting the distribution of forest cover ($R^2 = 0.85$), elevation ($R^2 = 0.45$), population density ($R^2 = 0.62$), and nighttime lights ($R^2 = 0.49$). Note that inconsistent image and label quality across the globe are likely partially responsible for lowering performance relative to the US-only experiments above (Supplementary Materials Section S.3.10).

National survey “field test” It has been widely suggested that SIML could be used by resource-constrained governments to reduce the cost of surveying their citizens (Jean, Burke, et al. 2016; Head et al. 2017; Reed et al. 2018; Bedi, Coudouel, and Simler 2007; De Sherbinin et al. 2017). To demonstrate MOSAIKS’s performance in this theoretical use-case, we simulate a ‘field test’ with the goal of recreating results from an existing nationally representative survey. Using the pre-computed features from the first US experiment above, we generate predictions for 12 pre-selected questions in the 2015 American Community Survey (ACS) conducted by the US Census Bureau (U.S. Census Bureau 2015). We obtain R^2 values ranging from 0.06 (*percent household income spent on rent*, an outlier) to 0.52 (*building age*), with an average R^2 of 0.34 across 12 tasks (Fig. 2.4B).

Compared to a baseline of “no ground survey,” or a costly survey extension, these results suggest that MOSAIKS predictions could provide useful information to a decision-maker for almost all tasks at low cost; noting that, in contrast, the ACS costs $> \$200$ million to deploy annually (U.S. Census Bureau 2021). However, some variables (e.g. *percent household income spent on rent*) may continue to be retrievable only via ground survey.

Extensions

The design of MOSAIKS naturally provides two additional useful properties.

Incorporating multiple sensors. Available satellites exhibit a diversity of properties (e.g. wavelength, timing of sampling) that can be used to improve SIML predictions (Tsagkatakis et al. 2019). While most SIML approaches, including the above analysis, use a single sensor, the design of MOSAIKS allows seamless integration of data from additional satellites because the regression step is linear in the features. To demonstrate this, we include nighttime lights as a second data source in the analysis of survey data from Rwanda, Haiti, and Nepal discussed above (Supplementary Materials S.4.1). In all 36 tasks, predictions either

improved or were unchanged when nighttime imagery was added to daytime imagery in the model (average $\Delta R^2 = 0.03$). This approach naturally optimizes how data from all sensors are used without requiring that users possess expertise on each technology.

Super-resolution predictions. Many use cases would benefit from SIML predictions at finer resolution than is available in training data (Tsagkatakis et al. 2019). Here we show that MOSAIKS can estimate the relative contribution of sub-regions within an image to overall image-level labels, even though only aggregated image-level labels are used in training (See Fig. 2.4C and Fig. S12). Such “super-resolution” prediction follows from the functional form of the featurization and linear regression steps in MOSAIKS, allowing it to be analytically derived for labels that represent nearly linear combinations of ground-level conditions (Supplementary Materials Section S.3.9 and Fig. S11). We numerically assess super-resolution predictions of MOSAIKS for the forest cover task, since raw label data are available at much finer resolution than our image labels. Provided only a single label per image, MOSAIKS recovers substantial within-image signal when predicting forest cover in 4 to 1,024 sub-images per label (within-image $R^2 = 0.54$ -0.32, see Fig. S13 and Supplementary Materials Section S.3.9).

Discussion

We develop a new approach to SIML that achieves practical generalization across tasks while exhibiting performance that is competitive with deep learning models optimized for a single task. Crucial to planet-scale analyses, MOSAIKS requires orders of magnitude less computation time to solve a new task than CNN-based approaches and it allows 1km-by-1km image data to be compressed ~ 6 -500 times before storage/transmission (Supplementary Materials Section S.1). Such compression is a deterministic operation that could theoretically be implemented in satellite hardware. We hope these computational gains, paired with the relative simplicity of using MOSAIKS, will democratize access to global-scale SIML technology and accelerate its application to solving pressing global challenges. We hypothesize that there exist hundreds of variables observable from orbit whose application could improve human well-being if measurements were made accessible.

While we have shown that in many cases MOSAIKS is a faster and simpler alternative to existing deep learning methods, there remain contexts in which custom-designed SIML pipelines will continue to play a key role in research and decision-making. Existing ground-based surveys will also remain important. In both cases we expect MOSAIKS can complement these systems, especially in resource constrained settings. For example, MOSAIKS can provide fast assessments to guide slower SIML systems or extend the range and resolution of ground-based surveys.

As real-world policy actions increasingly depend on SIML predictions, it is crucial to understand the accuracy, precision and sensitivity of these measurements. The low cost and

high speed of re-training MOSAIKS enables unprecedented stress tests that can support robust SIML-based decision systems.

Here, we tested the sensitivity of MOSAIKS to model parameters, number of training points, and degree of spatial extrapolation, and expect that many more tests can be developed and implemented to analyze model performance and prediction accuracies in context.

To aid systematic bench-marking and comparison of SIML architectures, the labels and features used in this study are made publicly available; to our knowledge this represents the largest multi-label benchmark dataset for SIML regression tasks.

By distilling SIML to a pipeline with simple and mathematically interpretable components, MOSAIKS facilitates development of methodologies for additional SIML use cases and enhanced performance. For example, the ability of MOSAIKS to achieve super-resolution is easily derived analytically (Supplementary Materials Section S.3.9). Furthermore, while we have focused here on tri-band daytime imagery, we showed that MOSAIKS can seamlessly integrate data from multiple sensors through simple concatenation, extracting useful information from each source to maximize performance. We conjecture that integrating new diverse data, from both satellite and non-satellite sources, may substantially increase predictive accuracy for tasks not entirely resolved by daytime imagery alone.

We hope that MOSAIKS lays the foundation for an accessible and democratized system of global information sharing, where imagery from all available global sensors is continuously encoded as features and stored in a single table of data, which is then distributed and used planet-wide. As a step in this direction, we make a global cross-section of features publicly available using 2019 imagery from Planet Labs, Inc. Such a unified global system may enhance our collective ability to observe and understand the world, a necessary condition for tackling pressing global challenges.

Code and data availability

Code, data, a configured computing environment, and free cloud computing for this analysis is provided via Code Ocean. The editor will be provided with a link to the Code Ocean “capsule”, which will be distributed to reviewers.

All data used in this analysis is from free, publicly available sources and is available for download other than the house price data. House price data is provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). More information on accessing the data can be found at <http://www.zillow.com/ztrax>. The results and opinions are those of the author(s) and do not reflect the position of Zillow Group. The house price dataset we release publicly is a subset of that which used in the analysis, where grid cells containing <30 observations of recent property sales are removed to preserve privacy. Instructions for downloading the replication data are included in the Readme file within the project’s Code Ocean capsule.

Acknowledgements

We thank Patrick Baylis, Joshua Blumenstock, Jennifer Burney, Hannah Druckenmiller, Jonathan Kadish, Alyssa Morrow, James Rising, Geoffrey Schiebinger, and participants in seminars at UC Berkeley, University of Chicago, Harvard, American Geophysical Union, The World Bank, The United Nations Development Program & Environment Program, Planet Inc., The Potsdam Institute for Climate Impact Research, and The Workshop in Environmental Economics and Data Science for helpful comments and suggestions. We acknowledge funding from the NSF Graduate Research Fellowship Program (Grant DGE 1752814), the US Environmental Protection Agency Science To Achieve Results Fellowship Program (Grant FP91780401), the NSF Research Traineeship Program Data Science for the 21st Century, the Harvard Center for the Environment, the Harvard Data Science Initiative, the Sloan Foundation, and a gift from the Rhodium Group. The authors declare no conflicts of interest.

Supplementary Materials

The primary goal of our analysis is to develop, evaluate and contextualize the performance of MOSAIKS. In the following four supplementary sections we first summarize the methods used in this evaluation and then describe the data, the experiments conducted, and how MOSAIKS compares to other approaches in the literature in greater depth. We also describe the intuition behind and the mechanics of MOSAIKS’s algorithms in greater detail.

S.1 Methods summary

Here we provide additional information on our implementation of MOSAIKS and experimental procedures, as well as a description of the theoretical foundation underlying MOSAIKS. Full details on the methodology behind MOSAIKS can be found throughout Section S.3.

Implementation of MOSAIKS

We begin with a set of images $\{\mathbf{I}_\ell\}_{\ell=1}^N$, each of which is centered at locations indexed by $\ell = \{1, \dots, N\}$. MOSAIKS generates task-agnostic feature vectors $\mathbf{x}(\mathbf{I}_\ell)$ for each satellite image $\mathbf{I}_\ell$ by convolving an $M \times M \times S$ “patch”, $\mathbf{P}_k$, across the entire image. M is the width and height of the patch in units of pixels and S is number of spectral bands. In each step of the convolution, the inner product of the patch and an $M \times M \times S$ sub-image region is taken, and a ReLU activation function with bias $b_k = 1$ is applied. Each patch is a randomly sampled sub-image from the set of training images $\{\mathbf{I}_\ell\}_{\ell=1}^N$ (Fig. S5). In our main analysis, we use patches of width and height $M = 3$ (Fig. S6) and $S = 3$ bands (red, green, and blue). To create a single summary metric for the image-patch pair, these inner product values are then averaged across the entire image, generating the k th feature $\mathbf{x}_k(\mathbf{I}_\ell)$, derived from patch $\mathbf{P}_k$. The dimension of the resulting feature space is equal to K , the number of patches used, and in all of our main analyses we employ $K = 8,192$ (i.e. 2^{13}). Both images and patches are whitened according to a standard image preprocessing procedure before convolution (Section S.3.3).

In practice, this one-time featurization can be centrally computed and then features $\mathbf{x}_k(\mathbf{I}_\ell)$ distributed to users in tabular form. A user need only (i) obtain and link the subset of these features that match spatially with their own labels and then (ii) solve *linear* regressions of the labels on the features to learn *nonlinear* mappings from the original image pixel values to the labels (the nonlinearity of the mapping between image pixels and labels stems from the nonlinearity of the ReLU activation function). We show strong performance across seven different tasks using ridge regression to train the relationship between labels y_ℓ and features $\mathbf{x}_k(\mathbf{I}_\ell)$ in this second step, although future work may demonstrate that other fitting procedures yield similar or better results for particular tasks.

Implementation of this one-time unsupervised featurization takes about the same time to compute as a single forward pass of a CNN. With $K = 8,912$ features, featurization

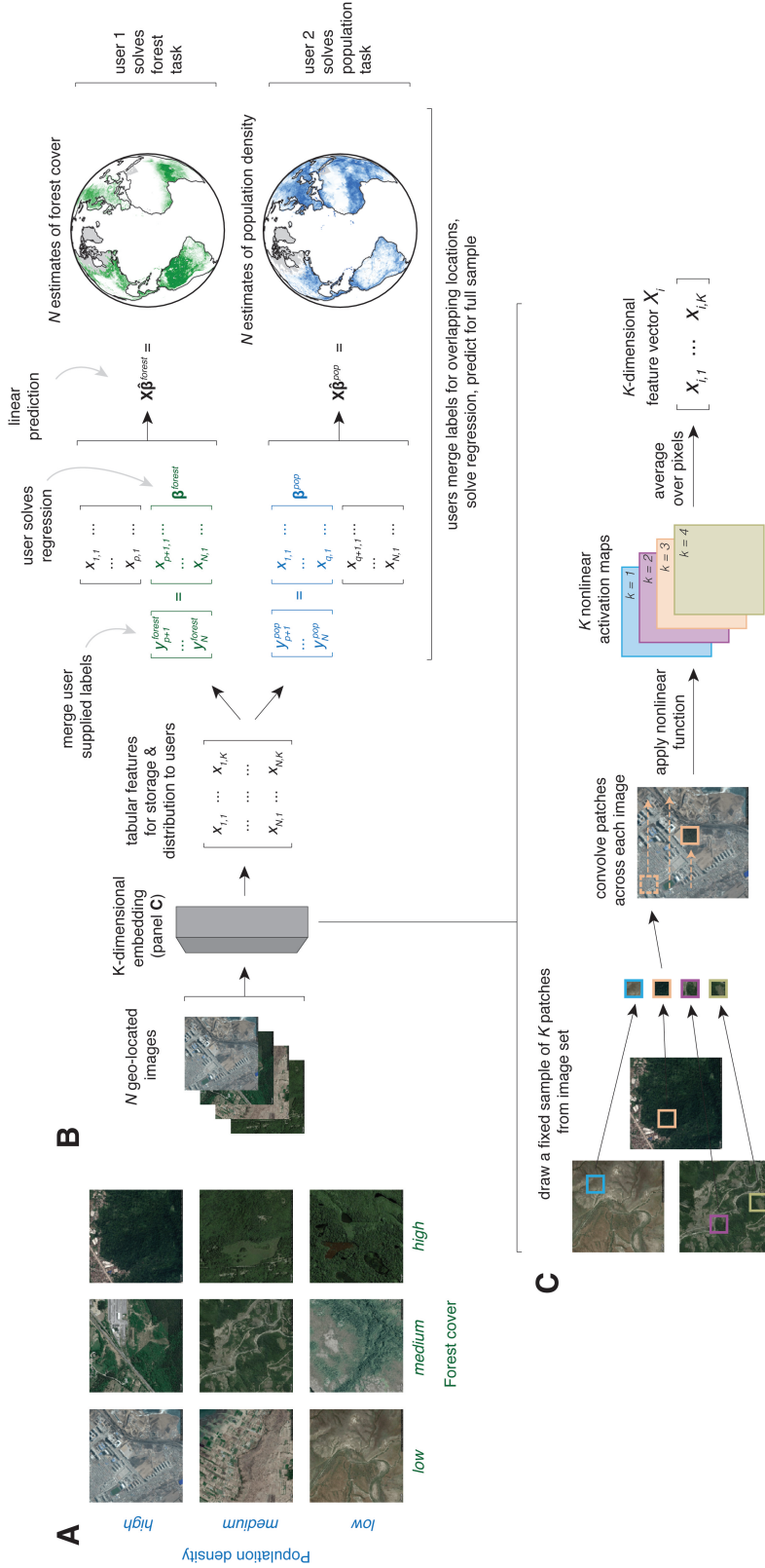


Figure 2.1: A generalizable approach to combining satellite imagery with machine learning (SIML) without users handling images. MOSAICS is designed to solve an unlimited number of tasks at planet-scale quickly. After a one-time unsupervised image featurization using random convolutional features, MOSAICS centrally stores and distributes task-agnostic features to users, each of whom generates predictions in a new context. (A) Satellite imagery is shared across multiple potential tasks. For example, nine images from the US sample are ordered based on population density and forest cover, both of which have distinct identifying features that are observable in each image. (B) Schematic of the MOSAICS process. N images are transformed using random convolutional features into a compressed and highly descriptive K -dimensional feature vector before recomputation. Once features are computed, they can be stored in tabular form (matrix $\mathbf{X}$) and used for unlimited tasks without recomputation. Users interested in a new task (s) merge their own labels (y^s) to features for training. Here, *User 1* has forest cover labels for locations $p+1$ to N and *User 2* has population density labels for locations 1 to q . Each user then solves a single linear regression for β^s . Linear prediction using β^s and the full sample of MOSAICS features $\mathbf{X}$ then generates SIML estimates for label values at all locations. Generalizability allows different users to solve different tasks using an identical procedure and the same table of features—differing only in the user-supplied label data for training. Each task can be solved by a user on a desktop computer in minutes without users ever manipulating the imagery. (C) Illustration of the one-time unsupervised computation of random convolutional features (Methods and Supplementary Materials Section S.3.3). K patches are randomly sampled from across the N images. Each patch is convolved over each image, generating a nonlinear activation map for each patch. Activation maps are averaged over pixels to generate a single K -dimensional feature vector for each image.

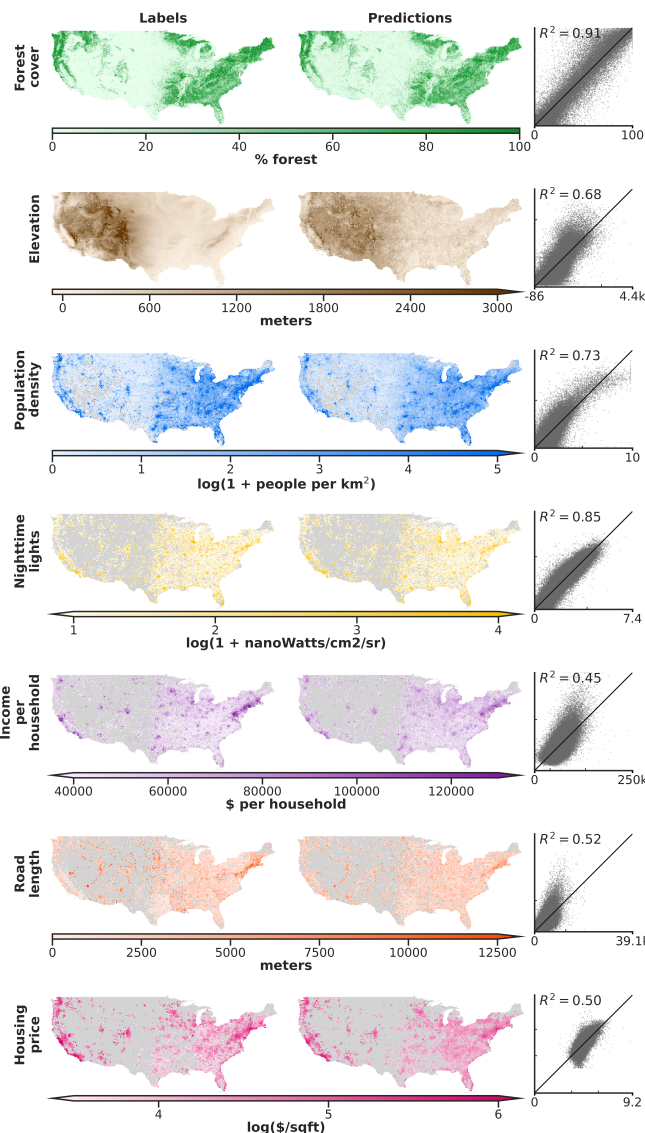


Figure 2.2: 1km $\times$ 1km resolution prediction of many tasks across the continental US using daytime images processed once, before tasks were chosen. 100,000 daytime images were each converted to 8,192 features and stored. Seven tasks were then selected based on coverage and diversity and predictions were generated for each task using the same procedure. Left maps: 80,000 observations used for training and validation, aggregated up to 20km $\times$ 20km cells for display. Right maps: concatenated validation set estimates from 5-fold cross-validation for the same 80,000 grid cells (observations are never used to generate their own prediction), identically aggregated for display. Scatters: Validation set estimates (vertical axis) vs. ground-truth (horizontal axis); each point is a $\sim$ 1km $\times$ 1km grid cell. Black line is at 45°. Test set and validation set performance are essentially identical (Table S2); validation set values are shown for display purposes only, as there are more observations. The tasks in the top three rows are uniformly sampled across space, the tasks in the bottom four rows are sampled using population weights (Supplementary Materials Section S.3.1); grey areas were not sampled in the experiment.

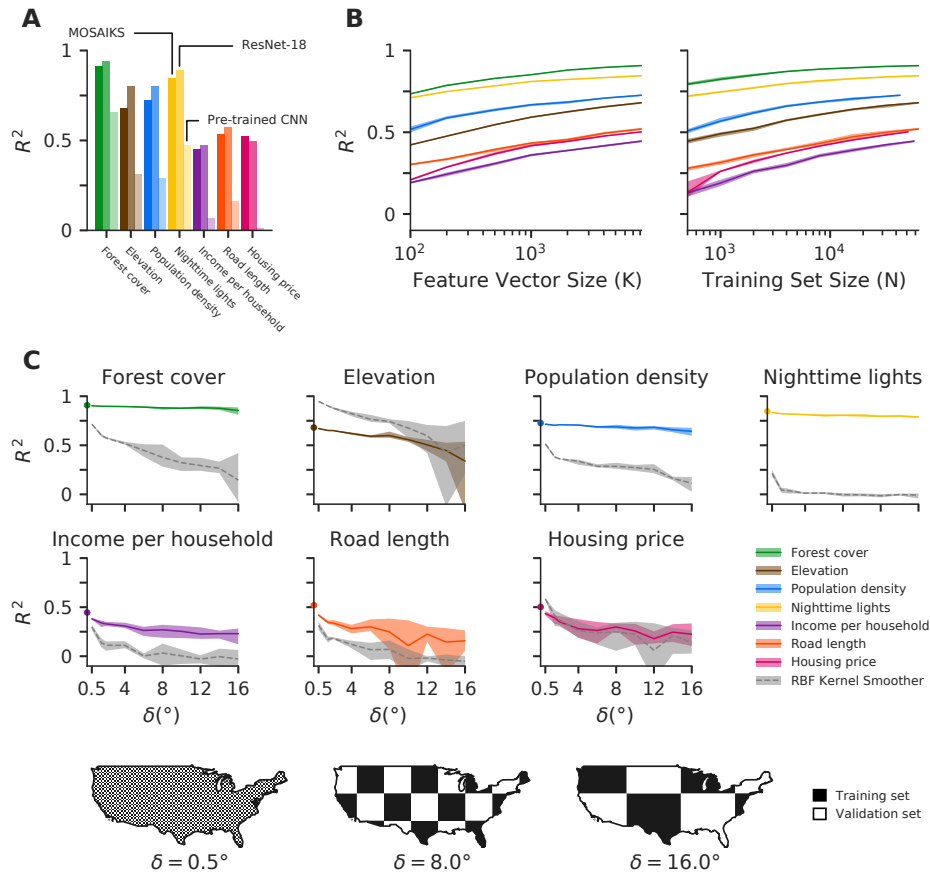


Figure 2.3: **Prediction accuracy relative to a convolutional neural network and transfer learning, using smaller K and N , and over large contiguous regions with no ground truth data.** (A) Task-specific MOSAIKS test-set performance (dark bars) in contrast to: an 18-layer variant of the ResNet Architecture (ResNet-18) trained end-to-end for each task (middle bars); and an unsupervised featurization using the last hidden layer of a 152-layer ResNet variant pre-trained on natural imagery and applied using ridge regression (lightest bars). See Supplementary Materials Section S.4.1 for details. (B) Validation set R^2 performance for all seven tasks while varying the number of random convolutional features K and holding $N = 64,000$ (left) and while varying N and holding $K = 8,192$ (right). Shaded bands indicate the range of predictive skill across 5 folds. Lines indicate average accuracy across folds. (C) Evaluation of performance over regions of increasing size that are excluded from training sample. Data are split using a “checkerboard” partition, where the width and height of each square is δ (measured in degrees). Example partitions with $\delta = 0.5^\circ$, 8° , 16° are shown in maps. For a given δ , training occurs using data sampled from “black squares” and performance is evaluated in “white squares.” Plots show colored lines representing average performance of MOSAIKS in the US across δ values for each task. Benchmark performance from Fig. 2.2 are indicated as circles at $\delta = 0$. Grey dashed lines indicate corresponding performance using only spatial interpolation with an optimized radial basis function kernel instead of MOSAIKS (Supplementary Materials Section S.3.8). To moderate the influence of the exact placement of square edges, training and test sets are resampled four times for each δ with the checkerboard position re-initialized using offset vertices (see Supplementary Materials Section S.3.8 and Fig. S10). The ranges of performance are plotted as colored or grey bands.

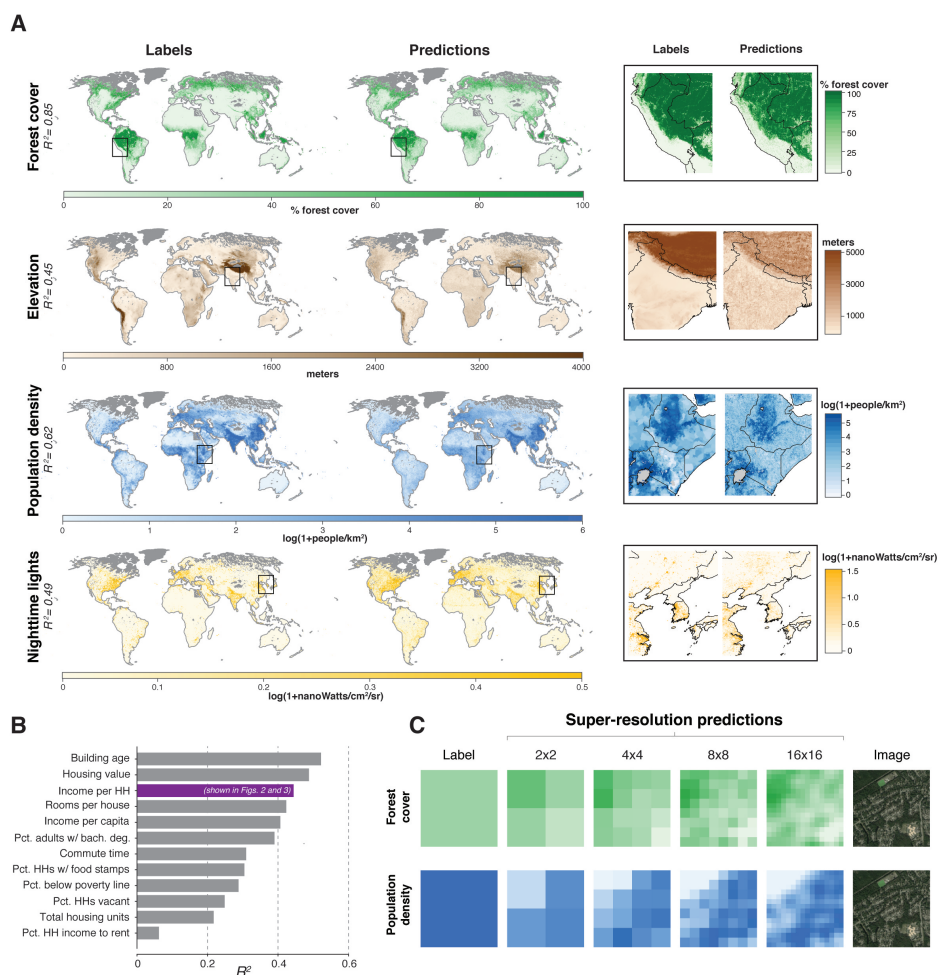


Figure 2.4: A single featurization of imagery predicts multiple variables at planet-scale, predicts results from a national survey, and achieves super-resolution. (A) Training data (left maps) and estimates using a single featurization of daytime imagery (right maps). Insets (far right) marked by black squares in global maps. Training sample is a uniform random sampling of 1,000,000 land grid cells, 498,063 for which imagery were available and could be matched to task labels. Out-of-sample predictions are constructed using 5-fold cross-validation. For display purposes only, maps depict $\sim 50\text{km} \times 50\text{km}$ average values (ground truth and predictions at $\sim 1\text{km} \times 1\text{km}$). (B) Test-set performance in the US shown for 12 variables from the 2015 American Community Survey (ACS) conducted by the US Census Bureau (U.S. Census Bureau 2015). Income per household (in purple) is also shown in Figs. 2.2 and 2.3, and was selected as an outcome for the analysis in these figures before this ACS experiment was run. (C) Both labels and features in MOSAIKS are linear combinations of sub-image ground-level conditions, allowing optimized regression weights to be applied to imagery of any spatial extent (Supplementary Materials Section S.3.9). MOSAIKS thus achieves super-resolution by generating skillful estimates at spatial resolutions finer than the labels used for training. Shown are example super-resolution estimates at 2×2 , 4×4 , 8×8 , and 16×16 the original 1×1 label resolution (See Fig. S12 for additional examples). Systematic evaluation of within-image R^2 across the entire sample is reported in Supplementary Materials Section S.3.9 for the forest cover task.

results in a roughly 6 to 1 compression of stored and transmitted imagery data in the cases we study.² Notably, storage and computational cost can be traded off with performance by using more or fewer features from each image (Fig. 2.3B). Since features are random, there is no natural value for K that is specifically preferable.

Experimental procedures

Task selection and data Tasks were selected based on diversity and data availability, with the goal of evaluating the generalizability of MOSAIKS (Section S.2.1). Results for all tasks evaluated are reported in the paper. We align image and label data by projecting imagery and label information onto a $\sim 1\text{km} \times 1\text{km}$ grid, which was designed to ensure zero spatial overlap between observations (Sections S.3.1 and S.3.2). Images are obtained from the Google Static Maps API (Section S.2.2) (*Maps Static API*), and labels for the seven tasks are obtained from refs. (Hansen et al. 2013; Amazon Web Services 2018; Center for International Earth Science Information Network (CIESIN) 2016; NOAA National Centers for Environmental Information 2019; U.S. Census Bureau 2015; U.S. Census Bureau 2016; Zillow 2017). Details on data are described in Table S1 and Section S.2.

US experiments From this grid we sample 20,000 hold-out test cells and 80,000 training and validation cells from within the continental US (Section S.3.4). To span meaningful variation in all seven tasks, we generate two of these 100,000-sample data sets according to different sampling methods. First, we sample uniformly at random across space for the forest cover, elevation, and population density, tasks which exhibit rich variation across the US. Second, we sample via a population-weighted scheme for nighttime lights, income, road length, and housing price, tasks for which meaningful variation lies within populated areas of the US. Some sample sizes are slightly reduced due to missing label data ($N = 91,377$ for income, 80,420 for housing price, and 67,968 for population density). We model labels whose distribution is approximately log-normal using a log transformation (Section S.3.5 and Table S3).

Because fitting a linear model is computationally cheap, relative to many other SIML approaches, it is feasible to conduct numerous sensitivity tests of predictive skill. We present cross-validation results from a random sample, while also systematically evaluating the behavior of the model with respect to: (a) geographic distance between training and testing samples, i.e. spatial cross-validation, (b) the dimension K of the feature space, and (c) the size N of the training set (Fig. 2.3, Sections S.3.7 and S.3.8). We represent uncertainty in each sensitivity test by showing variance in predictive performance across different training and validation sets. We also benchmark model performance and computational expense against an 18-layer variant of the ResNet Architecture, a common deep network architecture

²This is calculated as: $(256 * 256 * 3) / (8192 * 4) = 6 \times$ compression, where $256 * 256 * 3$ integer values per image are compressed into 8192 float32 features, each of which takes $4 \times$ the storage of an integer. Using 100 features gives a $500 \times$ compression.

that has been used in satellite based learning tasks (Perez et al. 2017), trained end-to-end for each task and a transfer learning approach (Pan and Qiang Yang 2010) utilizing an unsupervised featurization based on the last hidden layer of a 152-layer ResNet variant pre-trained on natural imagery and applied using ridge regression (Sections S.4.1 and S.4.2).

Global experiment To demonstrate performance at scale, we apply the same approach used within the data-rich US context to global imagery and labels. We employ a target sample of $N = 1,000,000$, which drops to a realized sample of $N = 423,476$ due to missing imagery and label data outside the US (Fig. 2.4). We generate predictions for all tasks with labels that are available globally (forest cover, elevation, population density, and nighttime lights) (Section S.3.10).

Super-resolution experiment Predictions at super-resolution (i.e. higher resolution than that of the labels used to train the model), shown in Fig. 2.4B, are generated for forest cover and population density by multiplying the trained ridge regression weights by the un-pooled feature values for each sub-image and applying a Gaussian filter to smooth the resulting predictions (Section S.3.9). Additional examples of super-resolution performance are shown in Fig. S12. We quantitatively assess super-resolution performance (Fig. S13) using forest cover, as raw forest cover data are available at substantially finer resolution than our common $\sim 1\text{km} \times 1\text{km}$ grid. Performance is evaluated by computing the fraction of variance (R^2) within each image that is captured by MOSAIKS, across the entire sample.

Theoretical foundations

MOSAIKS is motivated by the goal of enabling generalizable and skillful SIML predictions. It achieves this by embedding images in a basis that is both descriptive (i.e. models trained using this single basis achieve high skill across diverse labels) and efficient (i.e. such skill is achieved using a relatively low-dimensional basis). The approach for this embedding relies on the theory of “random kitchen sinks” (Rahimi and Benjamin Recht 2008b), a method for feature generation that enables the linear approximation of arbitrary well-behaved functions. This is akin to the use of polynomial features or discrete Fourier transforms for function approximation generally, such as functions of one dimension. When users apply these features in linear regression, they identify linear weightings of these basis vectors important for predicting a specific set of labels. With inputs of high dimension, such as the satellite images we consider, it has been shown experimentally (Morrow et al. 2017; Coates and Andrew Y. Ng 2012; Jonas et al. 2018) and theoretically (Rahimi and Benjamin Recht 2008a) that a randomly selected subspace of the basis often performs as well as the entire basis for prediction problems.

Convolutional random kitchen sinks Random kitchen sinks approximate arbitrary functions by creating a finite series of features generated by passing the input variables z

through a set of K nonlinear functions $g(z; \Theta_k)$, each parameterized by draws of a random vector Θ . The realized vectors Θ_k are drawn independently from a pre-specified distributions for each of $k = 1 \dots K$ features. Given an expressive enough function g and infinite K , such a featurization would be a universal function approximator (Rahimi and Benjamin Recht 2008a). In our case, such a function g would encode interactions between all subsets of pixels in an image. Unfortunately, for an image of size $256 \times 256 \times 3$, there are $2^{256 \times 256 \times 3}$ such subsets. Therefore, the fully-expressive approach is inefficient in generating predictive skill with reasonably concise K because each feature encodes more pixel interactions than are empirically useful.

To adapt random kitchen sinks for satellite imagery, we use convolutional random features, making the simplifying assumption that most information contained within satellite imagery is represented in *local* image structure. Random convolutional features have been shown to provide good predictive performance across a variety of tasks from predicting DNA binding sites (Morrow et al. 2017) and solar flares (Jonas et al. 2018) to clustering photographs (Coates and Andrew Y. Ng 2012) (kitchen sinks have also been used in a non-convolutional approach to classify individual pixels of hyper-spectral satellite data (Pérez-Suay et al. 2017)). Applied to satellite images, random convolutional features reduce the number of effective parameters in the function by considering only local spatial relationships between pixels. This results in a highly expressive, yet computationally tractable, model for prediction.

Specifically, we create each Θ_k by extracting a small sub-image patch from a randomly selected image within our image set, as described above. These patches are selected independently, and in advance, of any of the label data. The convolution of each patch across the satellite image being featurized captures information from the entire $\mathbb{R}^{256 \times 256 \times 3}$ image using only $3 * M^2$ free parameters for each k . Creating and subsequently averaging over the activation map (after a ReLU nonlinearity) defines our instantiation of the kitchen sinks function $g(z; \Theta_k)$ as $g(\mathbf{I}_\ell; \mathbf{P}_k, b_k) = \mathbf{x}_k(\mathbf{I}_\ell)$, where b_k is a scalar bias term. Our choice of this functional form is guided by both the structural properties of satellite imagery and the nature of common SIML prediction tasks, and it is validated by the performance demonstrated across tasks.

Relevant structural properties of satellite imagery and SIML tasks Three particular properties provide the the motivation for our choice of a convolution and average-pool mapping to define g .

First, we hypothesize that convolutions of small patches will be sufficient to capture nearly all of the relevant spatial information encoded in images because objects of interest (e.g. a car or a tree) tend to be contained in a small sub-region of the image. This is particularly true in satellite imagery, which has a much lower spatial resolution than most natural imagery (Fig. S6).

Second, we expect a single layer of convolutions to perform well because satellite images

are taken from a constant perspective (from above the subject) at a constant distance and are (often) orthorectified to remove the effects of image perspective and terrain. Together, these characteristics mean that a given object will tend to appear the same when captured in different images. This allows for MOSAIKS’s relatively simple, translation invariant featurization scheme to achieve high performance, and avoids the need for more complex architectures designed to provide robustness to variation in object size and orientation.

Third, we average-pool the convolution outputs because most labels in the types of problems we study can be approximately decomposed into a sum of sub-image characteristics. For example, forest cover is measured by the percent of total image area covered in forest, which can equivalently be measured by averaging the percent forest cover across sub-regions of the image. Labels that are strictly averages, totals, or counts of sub-image values (such as forest cover, road length, population density, elevation, and night lights) will all exhibit this decomposition. While this is not strictly true of all SIML tasks, for example income and average housing price, we demonstrate that MOSAIKS still recovers strong predictive skill on these tasks. This suggests that some components of the observed variance in these labels may still be decomposable in this way, likely because they are well-approximated by functions of sums of observable objects.

Additional interpretations The full MOSAIKS platform, encompassing both featurization and linear prediction, bears similarity to a few related approaches. Namely, it can be interpreted as a computationally feasible approximation of kernel ridge regression for a fully convolutional kernel or, alternatively, as a two-layer CNN with an incredibly wide hidden layer generated with untrained filters. A discussion of these interpretations and how they can help to understand MOSAIKS’s predictive skill can be found in Section S.3.3.

S.2 Data

This section describes the datasets we use to construct our ground truth labels across all seven of our tasks: forest cover, elevation, population density, nighttime lights, income, road length, and housing price. In addition, we describe the imagery used in the analysis. In Section S.3.2 we detail our method for linking the labeled data for each outcome to the imagery (Fig. S4).

In evaluating the ability of MOSAIKS to generalize, we are interested in its ability to recover different types of variables, including: (i) variables that are averages of sub-image properties, (ii) variables that not directly observable through daytime imagery but are a function of visible objects in the image, such as nighttime lights, and (iii) variables that are an underlying factor that determines what material appears in the image, such as elevation. Labels may also be a combinations of (i)-(iii), such as housing price or household income. An advantage of MOSAIKS is that it solves all these cases without any alteration of method. In the main text, we use the the same set of image features to predict all seven outcomes and, in principle, this set of features can be used to predict an unlimited number of outcomes

(Section S.3.3, so long as the outcomes and the images are aligned as described in Section S.3.2).

For each task, we obtain an up-to-date and geographically complete publicly available datasource to match with the images. Most of these data are based on measurements from 2010 - 2015, though our data on population density draws from sources that date back as far as 2005 in order to achieve global coverage. Our imagery data, from the Google Static Maps API (Section S.2.2), was mostly acquired in 2018, though in some cases images may be a few years older.

Task	Units	Native resolution	Data source
Forest cover	% forest cover	$\sim 30\text{m} \times 30\text{m}$	(Hansen et al. 2013)
Elevation	meters	$\sim 611.5\text{m} \times 611.5\text{m}$	(Amazon Web Services 2018)
Population density	people per sq. km.	$\sim 1\text{km} \times 1\text{km}$	(Center for International Earth Science Information Network (CIESIN) 2016)
Nighttime lights	nanoWatts/cm ² /sr	$\sim 500\text{m} \times 500\text{m}$	(NOAA National Centers for Environmental Information 2019)
Income	USD per household	census block group	(U.S. Census Bureau 2015)
Road length	meters	polyline	(U.S. Census Bureau 2016)
Housing price	USD per sq. ft.	geocoded point data	(Zillow 2017)

Table S1: **Data sources for all tasks.** Note that for all raster data sets (forest cover, elevation, population density, and nighttime lights) stated resolutions apply to grid cells located at the equator; raster size in Euclidean distance will vary with latitude.

S.2.1 Labels

Tasks were chosen to represent outcomes of classes (i)-(iii) above subject to the condition that high resolution and up-to-date label data is available across the US. Below we describe these data sources. See Section S.3.2 and Fig. S4 for a description of how we assign raw label data to images.

Forest cover To measure forest cover, we use globally comprehensive raster data from ref. (Hansen et al. 2013), which is designed to accurately measure forest cover in 2010. This dataset is commonly used to measure forest cover when ground-based measurements are not available (Alkama and Cescatti 2016; Carlson et al. 2018). Forest in these data is defined as vegetation greater than 5m in height, and measurements of forest cover are given at a raw resolution of roughly 30m by 30m. These estimates of annual maximum forest cover are derived from a model based on Landsat imagery captured during the growing season.

Specifically, the authors train a pixel-level bagged decision tree using three types of features: “(i) reflectance values representing maximum, minimum and selected percentile values (10, 25, 50, 75 and 90% percentiles); (ii) mean reflectance values for observations between selected percentiles (for the max-10%, 10-25%, 25-50%, 50-75%, 75-90%, 90%-max, min-max, 10-90%, and 25-75% intervals); and (iii) slope of linear regression of band reflectance value versus image date.” These estimates of forest cover were derived using different spectral bands than we observe in our imagery, and using information about how surface reflectance changes over the growing season, which we did not observe. This gives us confidence that we are indeed learning to map visual, static, high-resolution imagery to forest cover, rather than simply recovering the model used in ref. (Hansen et al. 2013).³

Elevation We use data on elevation provided by Mapzen, and accessed via the Amazon Web Services (AWS) Terrain Tile service. These Mapzen terrain tiles provide global elevation coverage in raster format. The underlying data behind the Mapzen tiles comes from the Shuttle Radar Topography Mission (SRTM) at NASA’s Jet Propulsion Laboratory (JPL), in addition to other open data projects.

These data can be accessed through AWS at different zoom levels, which range from 1 to 14 and, along with latitude, determine the resolution of the resulting raster. To align with the resolution of our satellite imagery, we use zoom level 8, which leads to a raw resolution of 611.5 meters at the equator.⁴

Population density We use data on population density from the Gridded Population of the World (GPW) dataset (Center for International Earth Science Information Network (CIESIN) 2016). The GPW data estimates population on a global 30 arc-second (roughly 1 km at the equator) grid using population census tables and geographic boundaries. It compiles, grids, and temporally extrapolates population data from 13.5 million administrative units. It draws primarily from the 2010 Population and Housing Censuses, which collected data between 2005 and 2014. GPW data in the US comes from the 2010 census.⁵

Nighttime lights We use luminosity data generated from nighttime satellite imagery, which is provided by the Earth Observations Group at the National Oceanic and Atmospheric Administration (NOAA) and the National Geophysical Data Center (NGDC). The values we use are Version 1.3 annual composites representing the average radiance captured from satellite images taken at night by the Visible Infrared Imaging Radiometer Suite (VIIRS). We use values from 2015, the most recent annual composite available.

This composite is created after the Day/Night VIIRS band is filtered to remove the effects of stray light, lightening, lunar illumination, lights from aurora, fires, boats, and background

³These data can be accessed at:
<https://landcover.usgs.gov/glc/TreeCoverDescriptionAndDownloads.php>.

⁴We accessed these data via the R function `get_aws_terrain` from the `elevatr` package. Code and documentation can be found here: <https://www.github.com/jhollist/elevatr>.

⁵These data can be accessed at <http://sedac.ciesin.columbia.edu/data/collection/gpw-v4>

light. Cloud cover is removed using the VIIRS Cloud Mask product. These values are provided across the globe from a latitude of 75N to 65S at a resolution of 15 arc-seconds. The radiance units are $\text{nW cm}^{-2} \text{sr}^{-1}$ (nanowatts per square centimeter per steradian).

Like forest cover, these labels are themselves derived from satellite imagery. However, because they capture luminosity at night, while our satellite imagery is taken during the day, the labels for luminosity and the imagery used to predict luminosity represent independent data sources. Our ability to predict nighttime lights depends on how well objects visible during the day are indicative of light emissions at night.⁶

Income We use the American Community Survey (ACS) 5-year estimates of median annual household income in 2015. These data are publicly available at the census block group level, of which there are 211,267 in the US, including Puerto Rico. On average, block groups are around 38 km^2 , though block groups are smaller in more densely populated areas.⁷

Road length We use road network data from the United States Geological Survey (USGS) National Transportation Dataset, which is based on TIGER/Line data provided by US Census Bureau in 2016. Shapefiles for each state provide the road locations and types, including highways, local neighborhood roads, rural roads, city streets, unpaved dirt trails, ramps, service drives, and private roads. Road types are indicated by a 5-digit code Feature Class Code which is assigned by the Census Bureau.⁸ The variable we predict is road length (in meters), which is computed as the total length of all types of roads that are recorded in a given grid cell.

The Census Bureau database is created and corrected via a combination of partner supplied data, aerial images, and fieldwork. The spatial accuracy of linear features of roads and coordinates vary by source materials used. The accuracy also differs by region, causing cases in which some regions lack recordings of certain road types, the most common one being private roads and dirt trails. For example, private roads are rarely recorded in Indiana and some regions in Ohio (Fig. S1A), despite satellite images that suggest they are present (Fig S1B).⁹

Housing price We estimate housing price per square foot using sale price and assessed square footage values for residential buildings. Data are provided by Zillow through the Zillow Transaction and Assessment Dataset (ZTRAX). This dataset aggregates transaction and assessment data across the United States, combining reported values from states and counties with widely varying regulations and standards. Thus, significant data cleaning is

⁶These data can be accessed at https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html#NTL_2015.

⁷These data are accessible using the `acs` package in R (Glenn 2019), table number B19013.

⁸<https://www.census.gov/geo/reference/mtfcc.html>

⁹The data can be accessed at: <https://prd-tnm.s3.amazonaws.com/index.html?prefix=StagedProducts/Tran/Shape/>

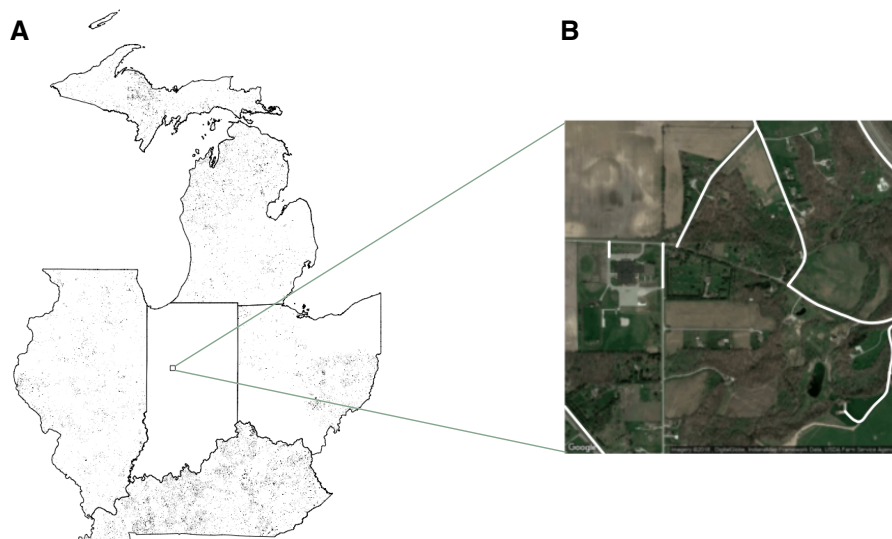


Figure S1: **Quality of ground truth road data varies by region.** (A) Private roads in the northern Midwest recorded in the USGS National Transportation Dataset. The conspicuous lack of recorded private roads in Indiana and sections of Ohio suggests that road data quality in certain regions may be lacking. (B) Overlaying recorded roads of all types (shown in white) over a single satellite image in Indiana, demonstrates that some roads that are easily visible from satellite imagery are missing in the available data that we use to construct labels.

required. Furthermore, because some states do not require mandatory disclosure of the sale price, we currently have limited data for the following states: Idaho, Indiana, Kansas, Mississippi, Missouri, Montana, New Mexico, North Dakota, South Dakota, Texas, Utah, and Wyoming. To address data quality issues, we develop a quality assurance and quality control (QA/QC) approach that is based on approaches employed in previous work (Moulton and Wentland 2018; Gindelsky, Moulton, and Wentland 2019; Union of Concerned Scientists 2018) but adapted for our case.

ZTRAX contains data on the majority of buildings in the United States, initially comprising 374 million detailed records of transactions across more than 2,750 counties. The data is organized into two components - *transaction data* and *assessment data*. These two datasets are linked, allowing us to merge the latest sale price of a property to the latest assessment data. To minimize the effect of nation-wide trends in housing price that would be unobservable from our cross-sectional satellite imagery, we limit our dataset to sales occurring in 2010 or later. Further, we restrict our analysis to buildings coded as “residential” or “residential income - multi-family” and drop any sale that was coded as an intra-family transfer. To obtain a square footage value, we follow the example in Zillow Research’s GitHub repository (*zillow-research/ztrax*) and take the maximum reported square footage for a given improvement, and then sum over all improvements on a given property.

To reduce the number of potentially miscoded outliers at the bottom end of the distribution of sale price and property size, we drop any remaining sales that fall under \$10,000 USD,

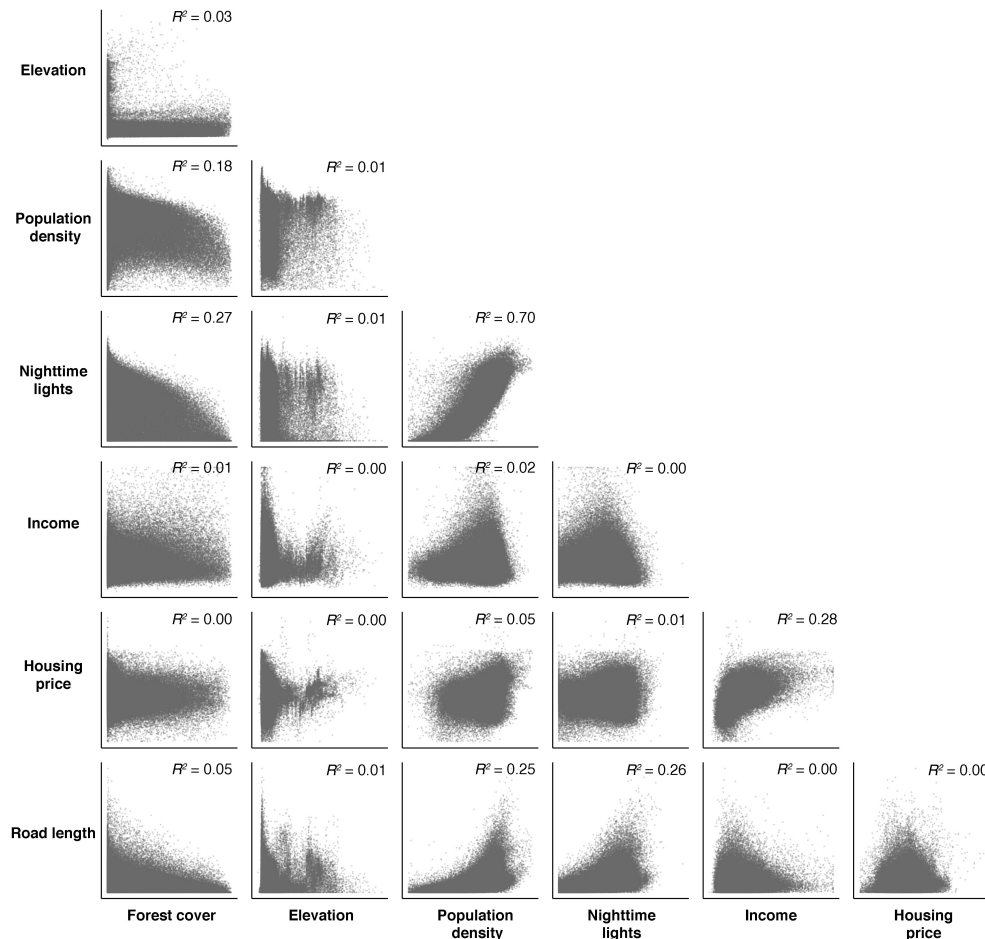


Figure S2: **Correlation of labels across tasks.** Each figure shows a scatter plot of labeled outcomes for one of our seven tasks against another. All points come from a population-weighted random sampling of grid cells (as described in Section S.3.1) across the US. Scatters and R^2 values are shown across approximately 100,000 grid cell labels, depending on the data availability for each task.

any properties that fall under 100 sq. ft., and any \$/sq. ft. values under \$10. To address outliers on the high end of the distribution, we take this restricted sample and further cut our dataset at the 99th percentile of \$/sq. ft. by state. Afterwards, we select the most recent recorded sale price for each property (divided by the most recent assessed square footage). We then average across all of the remaining units within each grid cell to comprise our final dataset of housing price per square foot.

To protect potentially identifiable information, our public data release contains housing price labels only for grid cells that contain 30 or more sales meeting the aforementioned criteria. This reduces the size of the dataset from $N = 80,420$ to $N = 52,355$ and makes the model performance obtainable by users better than that stated in the main text. For example, the public dataset will yield a test set R^2 of 0.60, rather than 0.52 (Table S2).

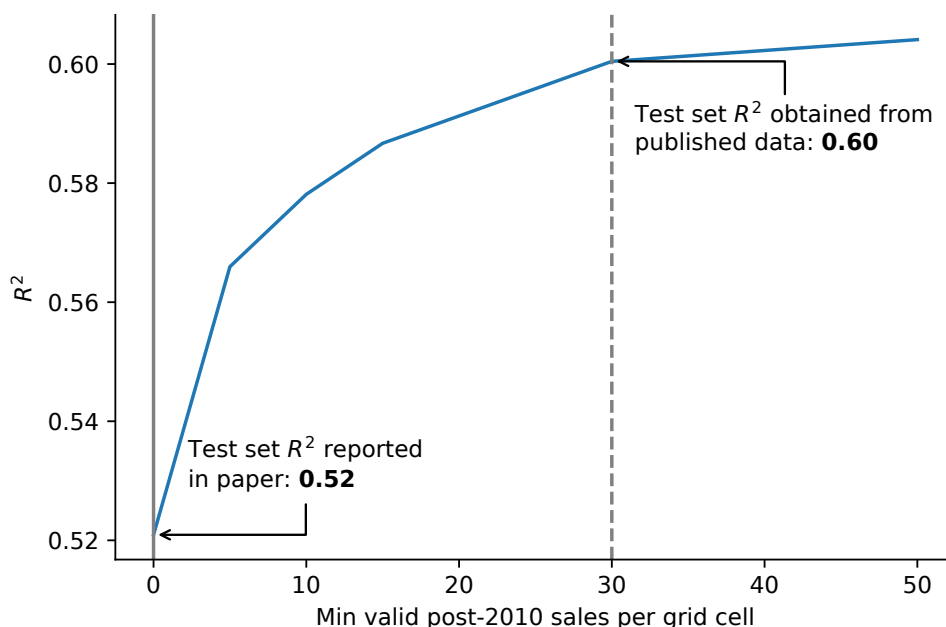


Figure S3: **Test set performance when restricting the dataset of mean housing price labels.** Curve shows the R^2 obtained in the test set for the housing price task in our main prediction experiment (Table S2), when removing data with low numbers of valid, recent sales of buildings within the associated grid cell. The dashed line indicates the restriction applied to the publicly available dataset.

This could be due to the fact that the *average housing price* label we train on is noisier when estimated in a grid cell with few valid sales prices. It could also be because the *average housing price* of areas with few recent sales may be inherently harder to predict via satellite imagery than that of areas with a greater number of recent sales. Fig. S3 empirically demonstrates the performance effect of removing grid cells with few recent sales.

Correlation of outcomes across tasks The seven tasks described above were selected in order to evaluate the performance of MOSAIKS across many diverse contexts. Figure S2 evaluates the extent to which this was achieved, by plotting label values against one another. A few of the labels are moderately correlated, most notably population density and nighttime lights, but in general there is substantial orthogonal variation across these seven tasks.

S.2.2 Imagery

We use satellite imagery from Google Static Maps API (*Maps Static API*), zoom level 16 (see Fig. 2.1A for examples). This gives roughly $1\text{km} \times 1\text{km}$ images which are 640×640 pixels across and 3 dimensions deep (red, green, and blue spectral bands). We coarsen these images

to $256 \times 256 \times 3$ prior to featurizing, meaning that our models are trained on images with roughly 4m resolution. These images can be composites of several satellite images – sources include the Landsat, Sentinel, SPOT, Pleiades, WorldView and QuickBird satellites.¹⁰ Prior to downloading, images were geo-rectified and pre-processed to remove cloud occlusions.¹¹

S.3 Methods

This section describes the methods that we use to define samples (Section S.3.1), to construct labels (Section S.3.2), and to construct features (Section S.3.3) for each image. It then describes how we separate data for training and evaluation (Section S.3.4), train models (Section S.3.5), test predictive skill (Section S.3.6), test sensitivity to the dataset size (Section S.3.7) and test model extrapolation performance (Section S.3.8). Next, we describe tests of model performance at sub-label or “super” resolution as well as at the global scale (Sections S.3.9 and S.3.10).

S.3.1 Grid definition and sampling strategy

Grid definition: To evaluate the generalizability of MOSAIKS performance across tasks we need a standardized unit of observation to link raw labels for all tasks and imagery. To do this, we construct a single global grid onto which we project both satellite imagery and labeled data. We design the grid to match our source of satellite imagery to ensure adjacent images do not overlap. Each element of the grid, i.e. each “grid cell,” was designed to be a square in physical space. Because the earth is a sphere, the angular extent of grid cells changes across latitudes.¹²

Sampling strategy: For our primary experiment in the continental US we subsample sets of 100,000 observations, roughly 1.25% of the grid cells in the continental US, using two distinct sampling strategies.¹³ First, we sample uniformly-at-random (UAR) from all grid cells within the continental US. This sampling strategy is most appropriate for tasks like forest cover, where there is meaningful variation in most regions of the country. Second, we implement a population-weighted (POP) sampling strategy. To generate this sample, each grid cell is weighted by population density values taken from Version 4 of the Gridded Population of the World dataset, which provides a raster of population density estimates for

¹⁰In some cases aerial photography is also integrated into images.

¹¹More information is available at: <https://developers.google.com/maps/documentation/maps-static/dev-guide>.

¹²For the continental US (spanning 25 to 50 degrees latitude and -125 to -66 longitude), the grid cells are 0.0138 degrees in width (1.39 km) at the southern edge of the grid, and 0.0138 degrees in width (0.98 km) at the northern edge of the grid. The grid cells are 0.012 degrees in height (1.39 km) at the southern edge of the grid, and 0.0089 degrees in height (.98 km) at the northern edge of the grid.

¹³We discard marine grid cells, but do not discard grid cells that are composed only of lakes or smaller inland bodies of water.

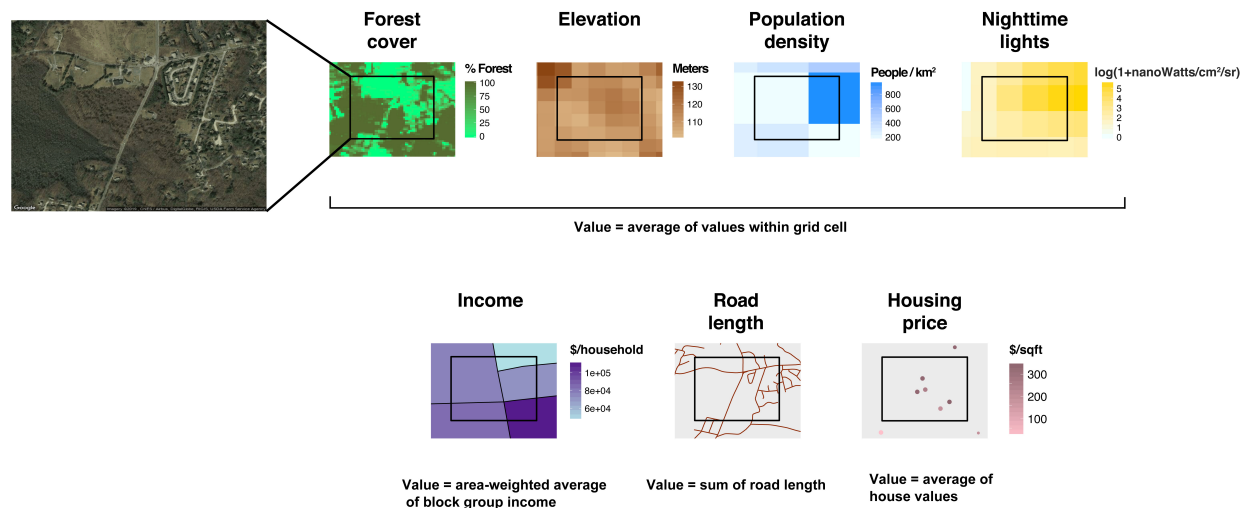


Figure S4: **Calculation of grid cell labels from raw data.** We calculate labels by spatially overlaying our grid cells and raw labeled data. We calculate labels as the average of raw label values that fall within the grid cell, except for roads where we calculate the label as the sum of road length within the grid cell.

the year 2015.¹⁴ This weighted sampling strategy is most applicable to tasks like housing price, where the most meaningful variation lies in more populated regions of the US. We use the UAR grid when sampling population density to avoid any issues that might arise from sampling a task using the same variable as sampling weights. In both the UAR and POP samples, we randomly sample just once; all results in the paper are displayed using the same two subsets of the full grid. Note that these sub-sampled grid cells, by construction, are each covered by exactly one satellite image without having to process data over the entire US.

In our main results, we use the UAR sample for the forest cover, elevation, and population density tasks. We use the POP sample for nighttime lights, income, road length, and housing price. See Section S.3.10 for a discussion of how we extend this grid and sampling procedure to the global scale.

S.3.2 Assigning labeled data to sampled imagery

To assign labels to each grid cell, we spatially overlay our raw labeled data and our custom grid. The native format and spatial resolution of the labeled data vary across the tasks studied, necessitating different aggregation or disaggregation procedures for each task. Here, we describe the approach taken in each task (Fig. S4).

The raw forest cover, elevation, population density and nighttime lights data are provided natively as rasters with higher spatial resolution than our custom grid. For these tasks, we perform aggregation by calculating the mean of all labeled pixels with centroids that fall

¹⁴These data are available at <http://sedac.ciesin.columbia.edu/data/collection/gpw-v4/sets/bro>

within the imagery grid cell. The resulting labels indicate mean forest cover, mean elevation, mean population density, and mean nighttime lights across the image grid cell.

Our road length data are provided as high-resolution spatial line segments. To aggregate these data to the image grid cell, we calculate the sum of road length segments within each image. The resulting labels indicate the total length of recorded roads that fall within an image grid cell.

Our housing price data are available as individual geocoded house sales. We aggregate these geocoded prices to the image grid cell by taking the average housing price per square foot across all sale prices that fall within the extent of the image. The resulting labels indicate the average housing price per square foot across all observed houses within a grid cell.

Our income data are provided at the block-group level (see Section S.2.1 for details). In some parts of the U.S., these block-groups are larger in total area than our image grid cells. However, in other regions, block-groups are smaller than our image grid cells. To treat both cases consistently, we aggregate incomes to the grid cell level by taking the weighted average of block-group incomes, where the weights are the area of intersection between the image grid cell and the block-group polygons. These weights are normalized to unity for each grid cell. The resulting labels indicate the area-weighted average median income across the grid cell.

Future users of a production-scale version of MOSAIKS would employ label data of arbitrary format and resolution. The above approaches provide guidelines for how to match various forms of label data to the pre-computed image feature grid, but other methods may be used. In the simplest case, for example, sparse point data could be directly matched to the nearest grid cell centroid.

S.3.3 Featurization of satellite imagery

Notation In our context, the input variable z is a set of satellite images $\mathbf{I}$, each corresponding to a physical location, ℓ . We use brackets to denote indexing into images, with colons denoting sub-regions of images (e.g. $\mathbf{I}_\ell[i, j]$ is the $(i, j)^{th}$ pixel of image $\mathbf{I}_\ell$, $\mathbf{I}_\ell[i : i + M, j : j + M]$ is the square sub-image of size $M \times M$ starting at pixel (i, j) .) Because images have a third dimension (spectral bands), a colon $\mathbf{I}_\ell[i, j, :]$ denotes all bands at pixel (i, j) . Indexing into non-image objects is denoted with subscripts (e.g. the k^{th} element of vector $\mathbf{x}$ is denoted as $\mathbf{x}_k$ and the k^{th} patch in a set of patches $\mathbf{P}$ is denoted as $\mathbf{P}_k$). We denote inner products with angular brackets $\langle \cdot, \cdot \rangle$ and the convolution operator with $*$.

Connection to the kitchen sinks framework The *random kitchen sink* featurization used in MOSAIKS relies on a nonlinear mapping $g(z; \Theta_k)$, where z is an input variable and Θ_k is a randomly drawn vector. Here, we describe the implementation details of this featurization in the context of satellite imagery. Connecting our implementation and notation to the framework of random kitchen sinks, the random variables Θ_k are instantiated as the values of a random patch $\mathbf{P}_k$ and the bias b_k . The input variable z is an image $\mathbf{I}_\ell$, and

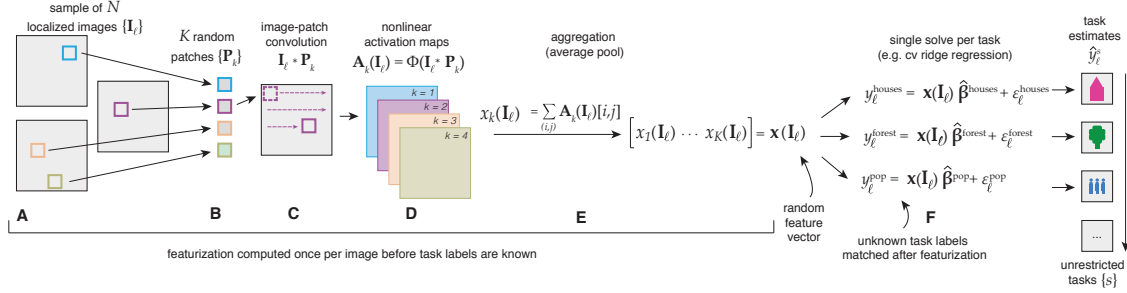


Figure S5: **MOSAIKS process from featurization to multi-task prediction.** Given a large sample of N satellite images (A), a random sample of K patches (B) are drawn. (C) These K random patches P_k are convolved over each image I_ℓ and (D) passed through a nonlinear function $\phi(\cdot)$ to generate K activation maps. (E) Pixel-specific activations are pooled across each image to generate one set of $N \times K$ features that are stored and distributed to all users. (F) The same random feature vector x is used in cross-validated ridge regression across many distinct tasks, after labeled and geo-referenced data y_ℓ is matched to features from each image I_ℓ (as shown in Fig. 1B of the main text). (G) Models trained via ridge regression can be used to generate predictions across unrestricted tasks for any location with satellite imagery.

$g(z; \Theta_k)$ represents the convolution of the patch over the image, followed by addition of the bias b_k and application of an element-wise ReLU function and an average pool, as described in the Methods of the main article and detailed below.

Methodological Details Fig. S5 depicts our featurization process. As described in Section S.2.1 and S.3.1, we begin with two sets (uniform and population-weighted samples) of $N = 100,000$ satellite images, each of which measures $640 \times 640 \times 3$ pixels (the third dimension represents the visible red, green, and blue spectral bands). We then coarsen the images to $256 \times 256 \times 3$ pixels to reduce computation. Next, we draw $K/2 = 4,096$ small sub-image “patches” of size $M \times M \times 3$ uniformly at random from the 80,000 images that comprise our training and validation set, and calculate the negative of each patch to get another 4,096 patches (Fig. S5A, S5B). Our chosen specification sets $M = 3$, so that each patch P_k is of dimension $3 \times 3 \times 3$ (see Fig. S6 for performance in experiments using different patch sizes).

We then “whiten” each patch by zero components analysis (ZCA), a common pre-processing routine in image processing (Krizhevsky and G. Hinton 2009). ZCA whitening pre-multiplies each patch by a transformation such that the resulting empirical covariance matrix of the whitened patches is the identity matrix. We then convolve each patch P_k over each of the N images (Fig. S5C) to obtain a set of $254 \times 254 \times 1$ pixel matrices for each image I_ℓ ¹⁵. During the convolutions each $3 \times 3 \times 3$ sub-image $I_\ell[i : i + M, j : j + M, :]$ is also whitened according to the same whitening matrix as is applied to the patches.¹⁶ We then

¹⁵To improve efficiency of the featurization process, our implementation calculates the inner product of patch and image only for the original $K/2$ patches. We then create an additional $K/2$ values equal to the negative of each of the original inner products.

¹⁶In practice, we apply the whitening operator as a right multiplication to the original 8192×27 whitened patch matrix in order to reduce computation.

apply a pixel-wise nonlinearity operator Φ to each resulting matrix to obtain K *nonlinear* activation maps $\mathbf{A}_k(\mathbf{I}_\ell) = \Phi(\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k)$ for each image $\mathbf{I}_\ell$ (Fig. S5D) so that the $(i, j)^{th}$ pixel of the k^{th} activation map is defined as

$$\mathbf{A}_k(\mathbf{I}_\ell)[i, j] = \Phi(\langle \mathbf{I}_\ell[i : i + M, j : j + M, :], \mathbf{P}_k \rangle + b_k),$$

where b_k is a bias term from the constant bias matrix $\mathbf{b}_k$, in which every element is equal to $b_k = 1$. We use $\Phi(\mathbf{I}_\ell; \mathbf{P}_k, \mathbf{b}_k) = \text{ReLU}(\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k) := \max\{\mathbf{P}_k * \mathbf{I}_\ell + \mathbf{b}_k, 0\}$ as the nonlinear operator. We then aggregate across the image by taking the average of the nonlinear activation maps (Fig. S5E). The combination of the nonlinear operator $\Phi(\cdot)$ and average pooling composes the function $g(\cdot)$ above, and creates a scalar value for each patch k and image ℓ pair:

$$\mathbf{x}_k(\mathbf{I}_\ell) = \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i, j] \quad (2.1)$$

Stacking these scalars across all K patches provides the resulting K -dimensional feature vector, $\mathbf{x}(\mathbf{I}_\ell) := [\mathbf{x}_1(\mathbf{I}_\ell) \ \mathbf{x}_2(\mathbf{I}_\ell) \ \dots \ \mathbf{x}_K(\mathbf{I}_\ell)] \in \mathbb{R}^K$. This featurization thus embeds the original image $\mathbf{I}_\ell$ into a K -dimensional feature space, which can then be mapped to many different outcomes using task-specific models (s) implemented by researchers (r): $y_\ell^{s,r} = \mathbf{x}(\mathbf{I}_\ell)\beta^{s,r} + \epsilon_\ell^{s,r}$, as illustrated in Fig. S5F. This linear relationship between labels and features may express a relationship between labels and image pixels that is highly nonlinear because the features themselves are nonlinear with respect to the images.

Patch size and sampling We approximate the idealized complete convolutional basis, which contains features for all patch sizes, with the simpler truncated basis where we use only a single patch size. Throughout our main analysis, we use a $3 \times 3 \times 3$ patch size for $\mathbf{P}_k$. While larger patches may, in principle, enable the detection of image features with larger spatial structure, we find that, in practice, patch size $M = 3$ performs best across all seven tasks (Fig. S6). This finding suggests that most information contained within satellite imagery of this resolution can be represented by local-level image structure, and that the inclusion of “non-local” relationships reduces the efficiency of the function approximator by introducing more degrees of freedom. This empirical finding is consistent with previous applications of kitchen sink features (Simonyan and Zisserman 2015).

We draw patches randomly from the empirical distribution of $M \times M \times 3$ patches from our training data set of satellite images. Drawing patches from the empirical distribution, rather than generating them randomly, allows us to sample efficiently from the distribution of sub-images we will encounter in the sample. This patch selection process is almost identical to the filter selection methods described in refs. (Coates, H. Lee, and Andrew Y Ng 2011; Benjamin Recht et al. 2019; Agarwal et al. 2014). It may be valuable for future research to explore whether MOSAICS performance and computational efficiency could be improved through patch selection algorithms. For example, one goal in selecting patches-based features is to

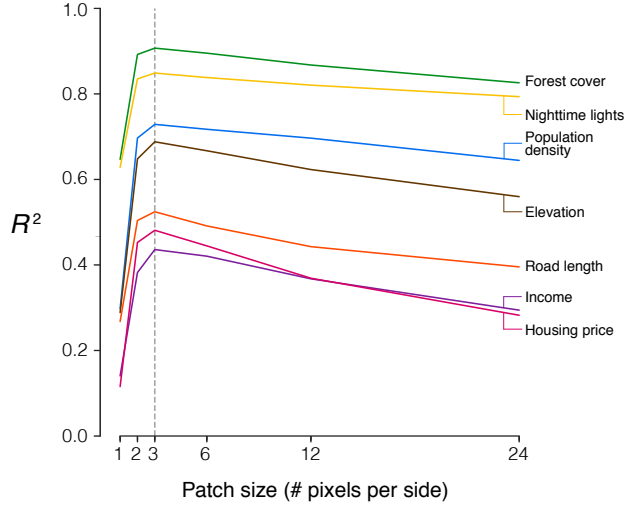


Figure S6: **Performance by patch size.** Featurization in MOSAIKS relies on convolving an $M \times M \times 3$ patch $\mathbf{P}_k$ across satellite images. M indicates the width in pixels of each sub-image patch, and the third dimension indexes the 3 spectral bands used throughout the analysis in this paper (an analogous approach can be applied to hyperspectral data). This figure shows, for each task, test set R^2 for patch sizes $M = 1, 2, 3, 6, 12$ and 24 , using $K = 2,048$ features for each M . The dotted gray line indicates the benchmark model used throughout the paper, with $M = 3$.

promote relative sparsity in the resulting patch-based features, as in ref. (Romero, Gatta, and Camps-Valls 2016). However, any attempt to tailor patch selection or featurization to a particular task of interest requires sacrificing the generalizability of this task-agnostic featurization. It remains an open question whether a non-randomly selected set of basis patches could potentially achieve similar (or greater) performance than what we present here when applied to arbitrary new tasks.

Alternative interpretations relating MOSAIKS to kernels and CNNs The methods summary describes how MOSAIKS’s convolutional random features enable nonparametric approximations of nonlinear functions through an embedding in a rich basis that expresses local spatial relationships. Here, we provide two alternative interpretations of the approach, the first relating to kernel methods and the second relating to convolutional neural networks. We believe these interpretations can provide useful lenses to consider why MOSAIKS works, and may also be helpful to researchers thinking about related problems.

First, one could interpret the design of MOSAIKS as if we were attempting to design a computationally tractable approximation to implementing a ridge regression using a convolutional kernel and the kernel trick. Under this interpretation, one could arrive at the same design of MOSAIKS using the following logic: (i) Design a kernel that allows us to describe the “similarity” of every image to every other image in the sample. (ii) For any new task, we want to use a kernel regression to predict the unobserved labels of new images based on their similarity to all other images—specifically, predicted labels would be a weighted sum of all

observed labels using weights determined by this kernel-based measure of image similarity, i.e. the kernel trick. (iii) Unfortunately, calculating such a kernel exactly would be computationally intractable on a data set as large as the one we use, so instead use convolutional kitchen sinks (i.e. the featurization in MOSAIKS) to approximate the desired kernel regression. This last step follows from prior work demonstrating two concepts. First, random features can approximate the lifted feature space induced by well-known kernels (Rahimi and Ben Recht 2007) as the number of random features increases. Second, convolutions of random patches drawn from joint Gaussian distributions has been proven to approximate, in the limit, a kernel in which every sub-image from one image is compared with every sub-image from another using an arc-cosine distance function (Daniely, Frostig, and Singer 2016). Thus, convolutions with random patches should, in the the limit, approximate a kernel that compares every sub-image with every other sub-image in the sample. However, because our distribution of patches is drawn from training imagery, rather than from Gaussian distributions, there is not an analytical expression that is known for the kernel being approximated by MOSAIKS in the limit.

The above logic would arrive at a design essentially the same as MOSAIKS, although it is not our preferred motivation or interpretation of why MOSAIKS works because it is a more complicated rationale than is needed. Ref. (Rahimi and Benjamin Recht 2008b) showed that the existence of an associated kernel is not necessary for performance using kitchen sinks. Rather, it is simply the embedding of an input in a descriptive basis that provides the predictive skill, the insight that motivates our preferred—and we think simpler—interpretation presented in the main text. Nevertheless, the interpretation of MOSAIKS in the context of kernels motivates one way to understand the mechanism through which MOSAIKS achieves predictive skill at low computational cost. Namely, it enables the approximation of a non-parametric kernel regression, using some (unknown) fully convolutional kernel that is sufficiently rich to represent meaningful similarity between images but costly enough to prohibit a direct application of the kernel trick.

An additional way to contextualize MOSAIKS is in terms of its computational elements. In particular, MOSAIKS uses image convolutions and nonlinear activation operations common to convolutional neural networks (CNNs) (Alber et al. 2017). Indeed, MOSAIKS is mathematically identical to the architecture one would arrive at if one designed a very shallow and very wide CNN without using backpropagation and instead using random filters. Specifically, MOSAIKS could be viewed as a two-layer CNN that has an 8,192-neuron wide hidden layer with untrained weights that are randomly initialized by drawing from sub-images in the sample, and that uses an average-pool over the entire image. In contrast to the conventional CNN approach of optimizing weights (via backpropagation), the random initialization with no subsequent optimization significantly reduces training time and avoids numerical challenges associated with non-convex optimization procedures (such as vanishing gradients). Thus, in the main text, we do not frame MOSAIKS as a CNN because MOSAIKS does not exploit the primary benefits of a deep CNN, since MOSAIKS lacks intermediate layers and does not implement backpropagation. Nonetheless, some readers may find this description more intuitive, and, as mentioned in the main article, we believe that the high

performance of MOSAIKS might motivate the design of CNN architectures that share some of these computational elements.

Because deep CNNs are a state-of-the-art tool for SIML tasks, we provide further comparisons of MOSAIKS performance and cost relative to this benchmark in Sections S.4.1 and S.4.2, respectively.

S.3.4 Data separation practices and cross-validation

We split our data into a 20% holdout test sample and an 80% training and validation sample. Within the training and validation sample, we perform 5-fold cross validation in our primary analysis, splitting the training and validation sample into 5 sets of 80% training data (64% of full sample) and 20% validation data (16% of full sample), such that the validation sets are disjoint.

Creating the holdout test set Before any of the label data are touched, we remove a hold-out test set that is chosen uniformly at random from the entire sample, consisting of 20% of the original data. The analysis and diagnostic procedures that follow use only the remaining 80% of the observations. The held-out test set is only used once, for the purposes of comparison to the validation set performance in Table S2. It is important to keep these data untouched until this point to ensure that our final performance results do not suffer from over-fitting.

Tuning hyperparameters We choose the optimal λ in Eq. (2.3) for each outcome through 5-fold cross-validation over the training and validation sample. Specifically, λ is chosen to maximize average performance (R^2) across 5 folds, from a list of candidate values.¹⁷ For tasks with the same sampling scheme (i.e. UAR versus population-weighted sampling), the folds are consistent across tasks, so that each of the five folds comprises the same set of locations across the tasks.

Using cross-validation to measure model robustness In addition to being a principled way of selecting hyperparameters, cross-validation gives us a notion of how robust our model is to changes in the training and validation samples. Since each of the 5 validation sets is disjoint and randomly selected, the empirical spread of performance across folds gives us a notion of the variability of our model when applied to new data sets from the same distribution. Understanding this variation is one way of understanding the performance of our model; it gives us a notion of variance of aggregated performance (e.g. R^2 over the entire sample, for a given set of hyperparameters). A useful aspect of MOSAIKS’s low computational cost of model training, however, is that it enables researchers to calculate the variance of individual predictions by bootstrapping.

¹⁷We choose these candidate values so as to ensure the chosen optimal λ is not the minimum or maximum of all λ s supplied.

S.3.5 Training and testing the model

In our primary model (results shown in Fig. 2.2 of the main text) we solve for grid cell labels as a linear function of random convolutional features using a ridge regression model and a cross-validation procedure. To obtain training and validation sets, we follow the data separation practices outlined in Sec. S.3.4, and drop any observations with missing values. The resulting combined training and validation set sizes are $N = 80,000$ for forest cover, 80,000 for elevation, 54,375 for population density, 80,000 for nighttime lights, 73,102 for income, 80,000 for road length, and 80,420 for housing price.

Population density, nighttime lights, and housing price have label distributions that are approximately log-normal (Fig. S7), so we take a log transformation of the labels. We add 1 before logging to avoid dropping labels with an initial value of zero (see Section S.3.6 for performance in logs and levels for all tasks).¹⁸

With these labels and features in hand, we regress each outcome y_ℓ^s for each task s on features $\mathbf{x}_\ell$ as follows:

$$y_\ell^s = \mathbf{x}(\mathbf{I}_\ell)\boldsymbol{\beta}^s + \epsilon_\ell^s \quad (2.2)$$

We solve for $\boldsymbol{\beta}^s$ by minimizing the sum of squared errors plus an l_2 regularization term:

$$\min_{\boldsymbol{\beta}^s} \frac{1}{2} \|\mathbf{y}_\ell^s - \mathbf{x}(\mathbf{I}_\ell)\boldsymbol{\beta}^s\|_2^2 + \frac{\lambda^s}{2} \|\boldsymbol{\beta}^s\|_2^2 \quad (2.3)$$

We use ridge regression across all outcomes to demonstrate the generalizability of using a single set of image features across many simple regression models. Further, this standardized methodology facilitates comparison of performance and sensitivity across tasks. We note that other modeling choices could potentially improve fit (e.g. using a hurdle model for zero-inflated outcome distributions such as road length); we leave such task-specific explorations for future research.

In visual display of results and calculation of performance metrics such as R^2 , we clip our predictions for each task at the minimum and maximum values observed in the labeled data.

The resulting weights (i.e. regression coefficients) $\hat{\boldsymbol{\beta}}^s$ obtained from estimation of Eq. (2.2) indicate, along with the variance of the features, which features k (derived from random patch $\mathbf{P}_k$) capture meaningful information for prediction in each task. Fig. S8 demonstrates that the recovered weights are stable across cross-validation folds within a task. The first two columns show standardized weights that are estimated from disjoint training and validation splits for the same task.¹⁹ Values corresponding to each axis are the regression weights estimated when the corresponding fold composes the training set. High R^2 values indicate a

¹⁸Since housing price per square foot is always positive, for that variable we use just a log transformation.

¹⁹For consistency across comparisons, R^2 is calculated on standardized regression weights, which have been demeaned and divided by their standard deviations. The number of random features is set to $K = 1,024$ for visual display purposes.

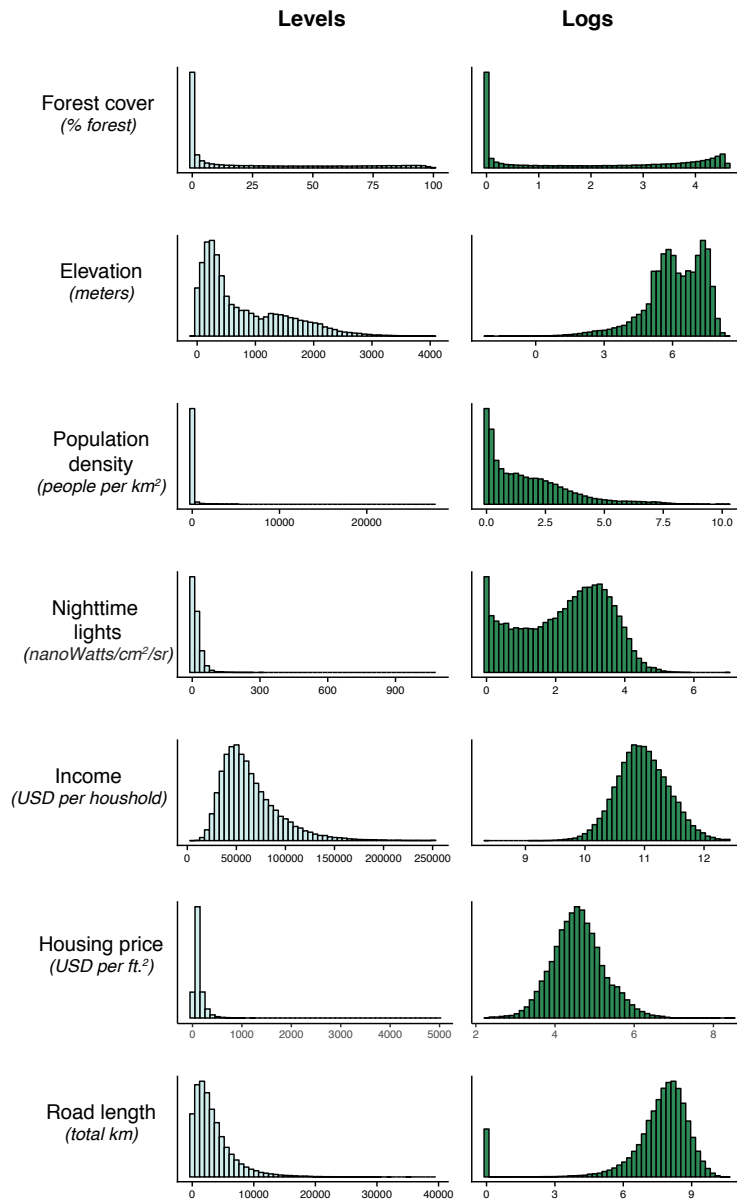


Figure S7: **Distribution of outcome variables in levels and logs.** Histograms show the distribution of each outcome variable over all sampled image grid cells (approximately 100,000 observations, depending on data availability). Forest cover, elevation, and population density are sampled uniform at random across the continental US, while all other variables are randomly sampled with population weighting. The first column shows the distribution in levels, and the second in logs. For elevation, population density, nighttime lights, and road length, logs were taken after adding 1 to the raw values, given the propensity of zero values in these outcomes.

strong correlation between regression weights from *different folds within a single task* (forest cover, elevation, and population density are shown), demonstrating that similar linear com-

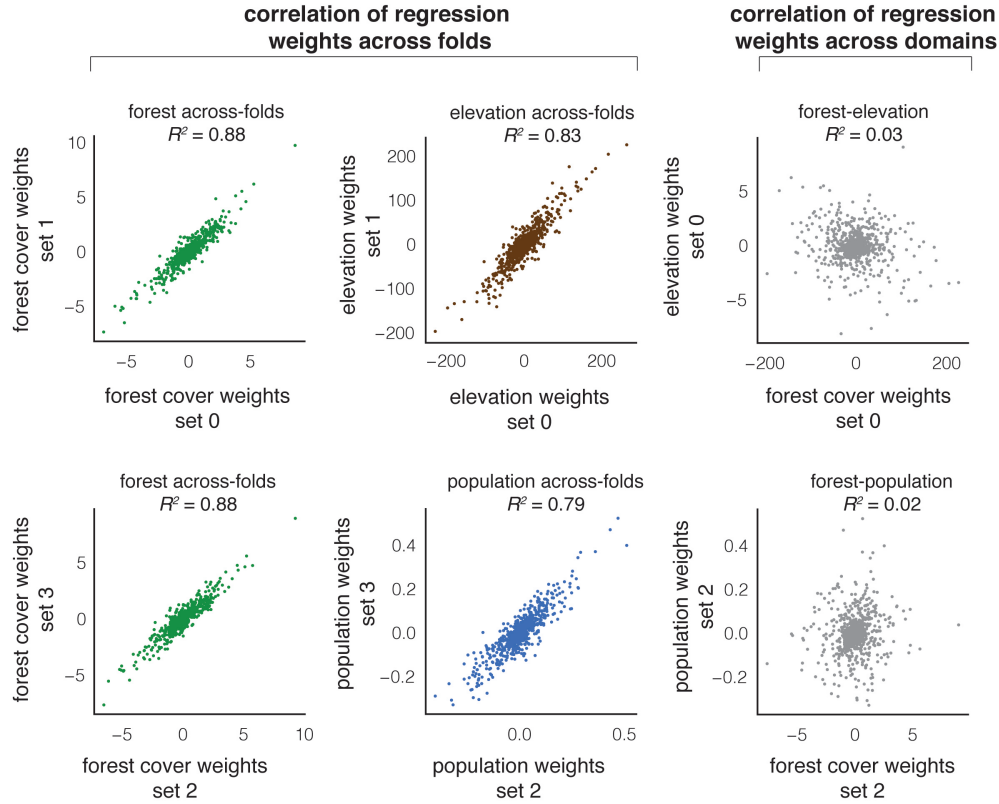


Figure S8: **Regression weights across folds within a task vs. across tasks within a fold.** All scatterplots indicate regression weights for forest cover, elevation and/or population density. Each point depicts estimated coefficient values for the k th feature (β_k^s) when trained on either different samples or different labels. In the across-fold examples (first two columns), we learn weights for disjoint training and validation splits for the same task via cross-validation in which one fold acts as the training set and the other as the validation set. Values corresponding to each axis are the regression weights when that fold is the training set (e.g. the top left scatter shows $\{\beta_k^{forest1}, \beta_k^{forest0}\}$), and indicate a strong correlation across regression weights from different folds. In the across-task examples (last column), regression weights are shown for the same training and validation sets for two distinct tasks (e.g. the top right scatter shows $\{\beta_k^{elevation0}, \beta_k^{forest0}\}$). We see that there is virtually no correlation in regression weights across tasks, demonstrating that predictions across tasks lie in orthogonal subspaces of the feature space. Across all examples here, we set the number of random features to $K = 1,024$.

binations of features are selected by the regression model, even when the sample of training images changes. This suggests that specific sets of patches consistently contain valuable information in predicting outcomes for a specific task. However, different combinations of patches are useful for different tasks, and we find no correlation in the weights recovered *between tasks*. For example, in the last column of the figure, we show that regression weights that are recovered for forest cover and elevation (top right) are essentially orthogonal as are regression weights recovered for forest cover and population (lower right). In these two plots, regression weights are shown for the same training and validation sets, but for two distinct

tasks. Sets of features that are relevant for prediction in one task appear to be irrelevant for another, as there is virtually no correlation in regression weights.

Intuition The consistency of weights recovered in MOSAIKS across folds within a task, and the orthogonality of weights recovered within a fold but across tasks, provides some intuition for why MOSAIKS provides consistent results and generalizes across a very large (potentially infinite) number of potential tasks. The rich featurization $\mathbf{x}(\mathbf{I}_\ell)$ locates image $\mathbf{I}_\ell$ in a very high-dimensional (K -dimensional) feature space. Solving for β^s in Eq. (2.2) then identifies the K -dimensional vector β^s that points in the direction of most rapid ascent (the gradient vector) for labels y^s , when the position of images $\mathbf{x}(\mathbf{I}_\ell)$ are projected onto this vector. Because the feature space is so large — our baseline model has an 8,192-dimensional feature space — there are a vast number of orthogonal gradient vectors that can be drawn through this space along which images can be organized for different tasks. The left and center panels of Fig. S8 illustrate that similar K -dimensional gradient vectors β^s are selected when solving for the same task but using different samples (each point depicts an element of the vector β^s). The right panels show that for different tasks, the gradient vectors are orthogonal and point in completely unrelated directions in the feature space. This orthogonality means that predictions $\hat{y}^s$ for different tasks will be independent of one another, even though both are constructed as linear combinations of the same set of features.

S.3.6 Primary model test set performance, robustness to functional form, and spatial distribution of errors

Here, we describe how we test for overfitting to the training and validation set in our primary model, test for primary model performance robustness to alternative functional forms, and characterize the spatial distribution of primary model error.

Performance in a holdout test set To test for overfitting, we evaluate the performance of our primary model on a randomly sampled 20% holdout set. These data were never used for model selection and were only touched at the end of our analysis to check for overfitting. To conduct this test, for each outcome, we use cross-validation within the training set to determine the outcome-specific optimal λ . We then retrain the model on the full training set using this optimal λ , and evaluate this model on the holdout test set. We find that performance in the test set is nearly identical to that of the validation set (Table S2), which indicates that our models were not overfit to the data. For some performance metrics, such as the maps in the main text, we present validation set performance (instead of the test set) because the sample is larger and the performance is unchanged.

Robustness of model to alternative functional forms Throughout the main text, we report primary model performance in each task from a model estimated with labels that are either logged (e.g. population density), or in levels (e.g. forest cover). The decision regarding

<i>Task</i>	Cross-validation R^2	Test set R^2
Forest cover	0.91	0.91
Elevation	0.68	0.68
Population density	0.73	0.72
Nighttime lights	0.85	0.85
Income	0.45	0.45
Road length	0.52	0.53
Housing price	0.50	0.52

Table S2: **Model performance in the hold out test set.** For each outcome, we use 5-fold cross-validation within the training/validation set using 80% of our labeled data to optimally select task-specific hyperparameters in ridge regression (i.e. λ). We then retrain each model on the full training set using this optimal λ . Performance on the validation set (column 1) is compared to that of the held out test set (column 2).

functional form for each task was made based on the underlying distribution of labels across our image grid cells. Many outcomes, such as housing prices, display exceptionally skewed distributions that approximate log-normality (see Fig. S7). For these outcomes, we take the natural log of the image grid cell values in model training and testing. Table S3 shows model performance for all tasks under both the levels and logs functional forms.²⁰ Tasks with highly skewed distributions, such as population density, housing price per square foot, and nighttime lights have substantially higher performance (R^2 increases by 10-64%) after being logged. Tasks whose labels display much less skew in levels, such as road length, income, and elevation show small to modestly reduced performance (4-21%) when their outcomes are modeled in logs.

Spatial distribution of errors Fig. S9 shows the distribution of errors over space, for the model predictions presented in Fig. 2.2. The model systematically over-predicts low values and under-predicts high values across all tasks. This is likely due to our choice of ridge regression, which favors predictions that tend toward the mean due to the ℓ_2 penalty. The structured correlation of errors across space suggests that there is substantial room for model improvement, potentially from including task specific knowledge. For example, our models of housing price and elevation could likely, respectively, be improved by adding in information about school districts –to address clustering of house price error in parts of big cities – or location – to help identify large areas of high elevation such as the Rocky Mountains. We recognize that there exists substantial room for task-specific model performance, which we leave for future research. Further, discontinuities in the error structure over political

²⁰In tasks where negative values or zeros are present (e.g. forest cover, elevation, and nighttime lights), we drop negative values and add one to zero values before taking logs for this test.

Forest cover	0.90	0.91
Elevation	0.58	0.68
Population density	0.73	0.56
Nighttime lights	0.85	0.77
Income	0.43	0.45
Road length	0.41	0.52
Housing price	0.50	0.44

Table S3: **Model performance across tasks and functional forms.** All R^2 values indicate performance using the optimal hyperparameter λ after 5-fold cross-validation. In the log model, the outcome variable is defined as the natural logarithm of the original labeled data (e.g. natural log of the average forest cover over an image gridcell). In the levels model, the outcome variable is simply the level of the aggregated labeled data, as defined in Section S.3.2. Values in bold are reported in the main text.

boundaries can help identify inconsistency in label quality. For example, the sharp increase in road length prediction error moving across the border from Louisiana to Texas suggests that the raw data labeling in these two states may differ methodologically, which introduces error into the label, and in turn, the model.

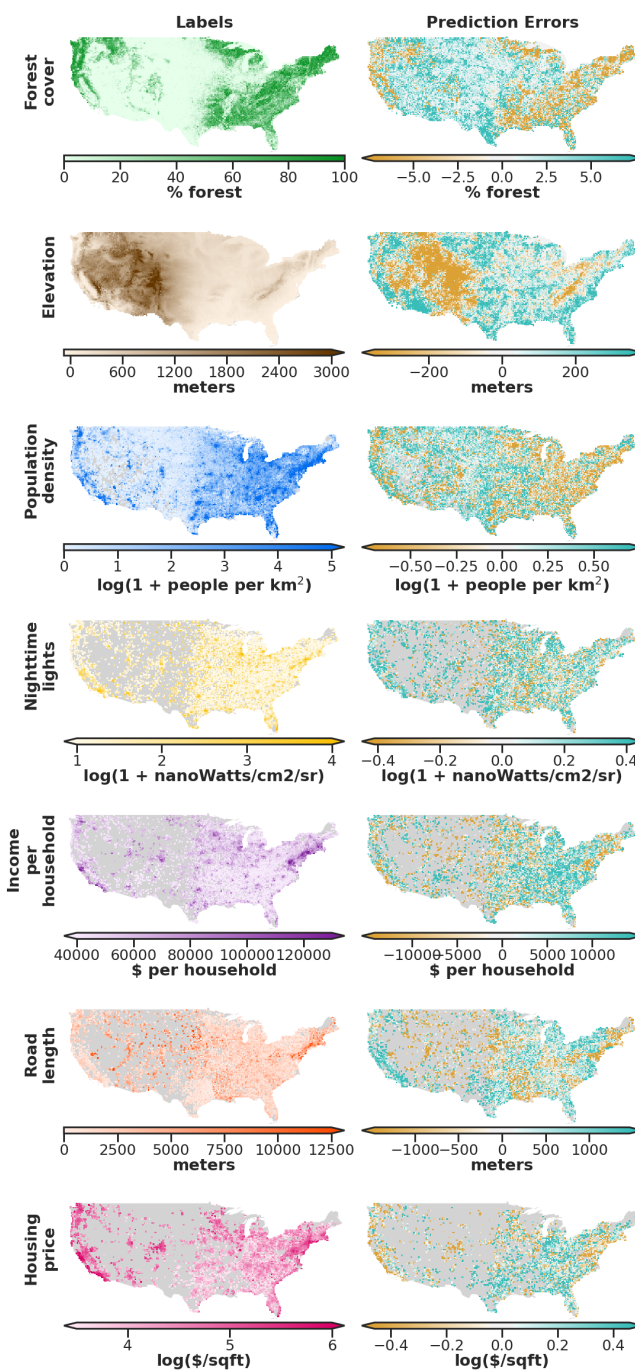


Figure S9: **Labels and prediction errors over space for each task.** Left maps: $\sim 80,000$ observations used for training and validation, aggregated up to $20\text{km} \times 20\text{km}$ cells for display (precise number of observations varies by task based on data availability; see Section S.3.5). Right maps: prediction errors from concatenated validation set estimates from 5-fold cross-validation for the same $\sim 80,000$ grid cells, identically aggregated for display.

S.3.7 Altering the number of features and training set size

To better understand factors that could improve the primary model, we test the sensitivity of its performance to the number of features and the training set size (results shown in Fig. 2.3 in the main text). Understanding the returns to additional features and observations enables better optimization of model performance given cost constraints.

Since features in MOSAIKS are generated randomly, there is no theoretical reason to select a specific number of features. To test the sensitivity of the primary model performance to the number of features, we train a model identically to our primary specification (Section S.3.5) except that we vary the number of features across the values $\{100, 200, 500, 1000, 2000, 4096, 8192\}$ (Fig. 2.3A). For each set of features and each task, we conduct 5-fold cross-validation to recover the optimal hyperparameter λ .

Notably, using only 100 features recovers a substantial amount of the variation across tasks. Of the tasks, the least variation is recovered for income (R^2 using 100 features is 81% of R^2 using 8,192 features) and the most variation is retained in nighttime lights (R^2 using 100 features is 96% of R^2 using 8,192 features). This suggests that in computation or memory-limited settings, fewer features could be used with only minor losses in performance. On the other hand, even with 8,192 features, performance does not fully flatten out (on a logarithmic scale). This suggests that performance could be improved further by increasing the number of features past $K = 8,192$. At the limit of our testing, a doubling of K from 4,096 to 8,192 led to a largest performance increase of 0.026 R^2 for income and a smallest of 0.010 R^2 for forest cover.

To test the sensitivity of primary model performance to the number of training samples, we train a model identical to our primary specification (with 8,192 features) except with a varying size of training set (from 500 to 64,000 images) (Fig. 2.3A).²¹ In cases where the training set has fewer than 64,000 total observations due to missing data (e.g. population density, income, road length and housing price), we use the full training data set to construct our largest training sample.

Similarly to increasing the number of features, increasing the training set size increases model performance with diminishing marginal returns. Notably, models trained on only 500 observations recover at minimum 56% (road length) of performance relative to $N = 64,000$ and at maximum 87% (forest cover), excluding income and housing price, which require larger samples to attain performance. This suggests that, for all but the most difficult SIML tasks, MOSAIKS may be useful even when label collection is very costly. For the tasks with the best R^2 performance (forest cover, nighttime lights), performance plateaus out as the number of training observations approaches 64,000. However, for the remaining five tasks, these results show that more training data could substantially increase performance further. The range of performance gain from increasing $N = 32,000$ to 64,000 is bounded below by forest cover (.005 R^2) and above by road length (.027 R^2).

²¹The same per-fold validation sets are used for each iteration of this analysis as well as for the primary analysis and for the test of model performance sensitivity to the number of features.

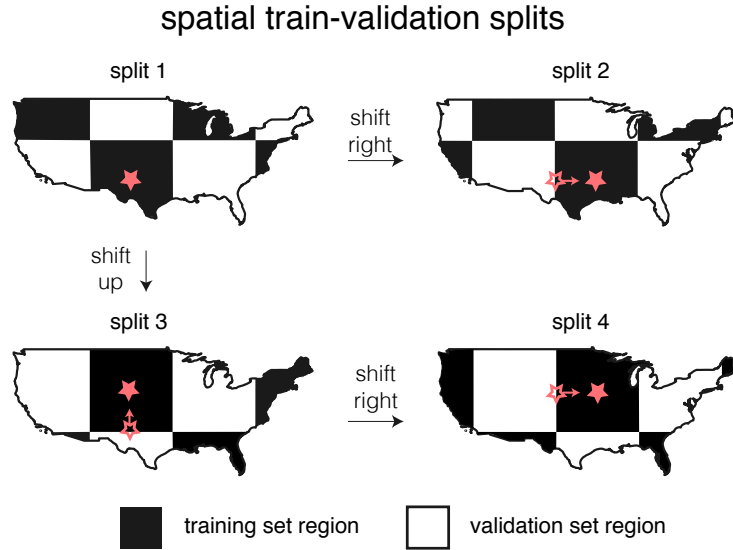


Figure S10: **Illustration of the procedure to systematically shift train and validation sets in space when assessing the performance of MOSAIKSover regions with no ground-truth data.** To assess the ability of MOSAIKS to generate meaningful predictions when extrapolating across large spatial distances, we conduct a “checkerboard” experiment (Section S.3.8, Fig. 3B-C of the main text) in which the training set (“black squares”) and validation set (“white squares”) are separated by increasingly large distances. The length of a square in each experiment is δ , measured in degrees. This figure demonstrates the four different train/validation splits that are created by shifting a given spatial checkerboard (split 1) by $\delta/2$ to the right (split 2), $\delta/2$ up (split 3), and both simultaneously (split 4).

S.3.8 Testing generalizability across space and comparison to kernel-based interpolation.

To understand the ability of our model to predict outcomes in large contiguous regions with no ground truth, we design an experiment where we evaluate models using training and validation sets that are increasingly far away from each other in space. Specifically, we iteratively create a grid over the US with a side length of δ degrees and then use this grid to divide the training and validation dataset ($N = 80,000$) into spatially disjoint sets of roughly equal size. We create these disjoint sets by assigning observations that lie in every other box within the grid to the train set and test set, respectively, creating a checkerboard pattern with the train set and test set, as shown in Fig. 2.3B. We vary the width δ of each square in the grid range across the values of $\{0.5, 1.5, 2, 4, 6, 8, 10, 12, 14, 16\}$ degrees (roughly 40 to 1400 km) in sequential runs of the experiment. As δ increases, validation set observations become on average farther away from the training set points. This distance makes prediction on the validation set more difficult, because observations in the validation set are now likely to be less similar to those in the training set. We learn the model on the training set using ridge regression. To assess the stability of this performance, we offset the checkerboard and re-run the above analysis four times – once in the original location and then three more

times – shifting the grid up, right, and both up and right by half the width of the grid (see Fig. S10). The ℓ_2 regularization term, λ , is selected to maximize average performance in the four validation sets, as we would select in a standard cross-validation procedure.

The performance plotted in Fig. 2.3C is the performance on the the resulting validation sets. We find that across most tasks, performance degrades only slightly as the distance between training observations and testing observations increases. This suggests that MOSAIKS is indeed learning image-label mappings that transfer across spatial regions.

Comparison of MOSAIKS to kernel-based spatial interpolation In these experiments we demonstrate that MOSAIKS outperforms spatial interpolation (or extrapolation, depending on geometry) – a commonly used simple technique to fill in missing data (Fig. 2.3C). This suggests that MOSAIKS, and SIML generally, exploits the spectral and structural content of information within an image to generate predictions at national scale that extend beyond what can be captured by geographic location alone.

We compare MOSAIKS to kernel-based spatial interpolation using a Gaussian Radial Basis Function (RBF) kernel, a simple and general widely used approach. In this approach, the value for a point in the validation set at location $\ell_v \in \mathbb{R}^2$ is predicted to be a weighted sum of the values of all the points in the training set ℓ_t , as follows:

$$\hat{y}_v^s = \frac{\sum_{\ell_t \in [\text{Train}]} y_t^s w(\ell_t, \ell_v)}{\sum_{\ell_t \in [\text{Train}]} w(\ell_t, \ell_v)}; \quad w(\ell_t, \ell_v) = e^{-\frac{1}{2\sigma^2} \|\ell_t - \ell_v\|^2}$$

Here, w is the weight assigned to each observation in the training set based on kernel values that are indexed to distance, such that w decreases as the distance between the point being predicted and the point in the training set increases. We select σ – the parameter that determines the rate at which w degrades with distance – to maximize average performance on the validation set across all four spatially-offset runs, similar to how we tune λ in the spatial extrapolation experiment described above. The optimal value of the bandwidth parameter σ will depend on the task at hand, as well as the average distance from points in the validation set to points in the training set. To ensure comparability, spatial interpolation based predictions and performance are computed for the exact same samples as used for MOSAIKS in each checkerboard partition.

S.3.9 Super-resolution

As discussed in the methods summary, the featurization method in MOSAIKS exploits the fact that many image-level outcomes of interest are linearly decomposable across sub-image regions. This is done by creating image-level features that are averages of statistics from all sub-image regions. Because these features are ultimately used in linear regression, a natural property of this approach is that weights estimated in this linear regression can be

used not only to generate predictions of outcome variables at the image-scale, but also at the scale of any sub-image region. As satellite imagery are available at increasingly high spatial resolution, this “super-resolution” property is both practical and powerful, enabling researchers to generate novel predictions at higher resolution than available ground truth data.

This section gives mathematical justification for a simple method to use MOSAIKS to predict outcomes of interest at a finer resolution than available labeled data. We display the super-resolution properties of MOSAIKS visually, and quantitatively document the empirical performance of this super-resolution approach.

Why MOSAIKS naturally achieves super-resolution for label predictions Given an image-label pair $\{\mathbf{I}_\ell, y_\ell^s\}$, the goal of super-resolution is to resolve which sub-regions of the image $\mathbf{I}_\ell$ contribute to high or low values of y_ℓ^s . Recall that for image $\mathbf{I}_\ell$, feature vector $\mathbf{x}(\mathbf{I}_\ell)$ is a K dimensional vector, where each scalar element $\mathbf{x}_k(\mathbf{I}_\ell)$ of $\mathbf{x}(\mathbf{I}_\ell)$ is an average across the pixels of the image of the values obtained by convolving sub-regions of the image with patch $\mathbf{P}_k$. As in Section S.3.3, denote by $\mathbf{X}$ the full random feature matrix in $\mathbb{R}^{N \times K}$, so that $\mathbf{X}_{\ell k}$ denotes the k^{th} element of the feature vector describing image $\mathbf{I}_\ell$. By Eq. (2.1), we can decompose the feature elements as:

$$\mathbf{X}_{\ell k} := \mathbf{x}_k(\mathbf{I}_\ell) = \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i, j]$$

where $\mathbf{A}_k$ is the activation map associated with patch $\mathbf{P}_k$. Since we are using a linear model to form predicted values, we can trace these values back to subregions of the original image. When we perform a linear regression for task s , the resulting regression weights are a vector $\hat{\beta}^s \in \mathbb{R}^K$ such that the scalar $\hat{\beta}_k^s$ describes the relative weight of feature k in the image-scale predictions. The prediction of outcome s using image $\mathbf{I}_\ell$ thus decomposes as:

$$\begin{aligned} \hat{y}_\ell^s &= \mathbf{X}_\ell \hat{\beta}^s \\ &= \sum_{k=1}^K \mathbf{X}_{\ell k} \cdot \hat{\beta}_k^s \\ &= \sum_{k=1}^K \left(\frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \mathbf{A}_k(\mathbf{I}_\ell)[i, j] \right) \cdot \hat{\beta}_k^s \\ &= \frac{1}{254^2} \sum_{i=1}^{254} \sum_{j=1}^{254} \underbrace{\left(\sum_{k=1}^K \hat{\beta}_k^s \cdot (\mathbf{A}_k(\mathbf{I}_\ell)[i, j]) \right)}_{\text{super-resolution prediction}} \end{aligned}$$

where the third line follows from substituting $\mathbf{X}_{\ell k}$ according to Eq. (2.1). Therefore, we can associate with each pixel indexed by (i, j) a predicted super-resolution value:

$$\hat{y}_{\ell, (i, j)}^s = \sum_{k=1}^K \hat{\beta}_k^s \cdot (\mathbf{A}_k(\mathbf{I}_\ell)[i, j]) \quad (2.4)$$

which is that pixel’s predicted label value, and thus its contribution to the overall predicted image-level label value $\hat{y}_\ell$ for $\mathbf{I}_\ell$. We use a Gaussian filter to smooth these per-pixel predictions to enforce spatial consistency and reduce variance of the high-resolution predictions, using a kernel bandwidth of $\sigma = 16$ pixels. These smoothed pixel-level predictions can be average-pooled to larger sub-image scales as shown in Fig. 2.4B. The procedure to construct super-resolution predictions, and a comparison of it to the procedure to construct image-level predictions, is illustrated in Fig. S11.

Fig. S12 demonstrates empirical performance of Eq. (2.4) using ten examples of this approach at super-resolutions on both the forest cover and population density outcomes. The ten images were randomly selected from the union of observations with forest cover $> 10\%$ and population density > 100 people/km² to ensure that all images considered had a non-negligible value for each variable.²²

In our formulation, super-resolution predictions are easily estimable during featurization. Consider again the per-pixel contributions of Eq. (2.4). An alternative way to express this is

$$\hat{y}_{\ell,(i,j)}^s = \left(\sum_{k=1}^K \hat{\beta}_k^s \cdot \mathbf{A}_k(\mathbf{I}_\ell) \right) [i, j]$$

That is, super-resolution estimates are just a linear combination of the activation maps $\mathbf{A}_k(\mathbf{I}_\ell)$ weighted by $\hat{\beta}_k^s$ (see Fig. S11). Every time we featurize a new image $\mathbf{I}_\ell$, we must perform the step of computing the K activation maps $\{\mathbf{A}_k(\mathbf{I}_\ell)\}_{k=1}^K$ (Fig. S5 D). Therefore, if we already have a suitable regression weight vector $\hat{\beta}^s$ for task s , for any new images $\mathbf{I}_\ell$ that we featurize, we can compute the super-resolution predictions $\sum_{k=1}^K \hat{\beta}_k^s \cdot \mathbf{A}_k(\mathbf{I}_\ell)$ as weighted combinations of the activation maps at negligible additional cost, prior to pooling, in the existing featurization pipeline.

²²To ensure that weights decomposed as a sum, as in Eq. (2.4), we used level values (i.e. not log-transformed) for population density labels in Fig. S12.

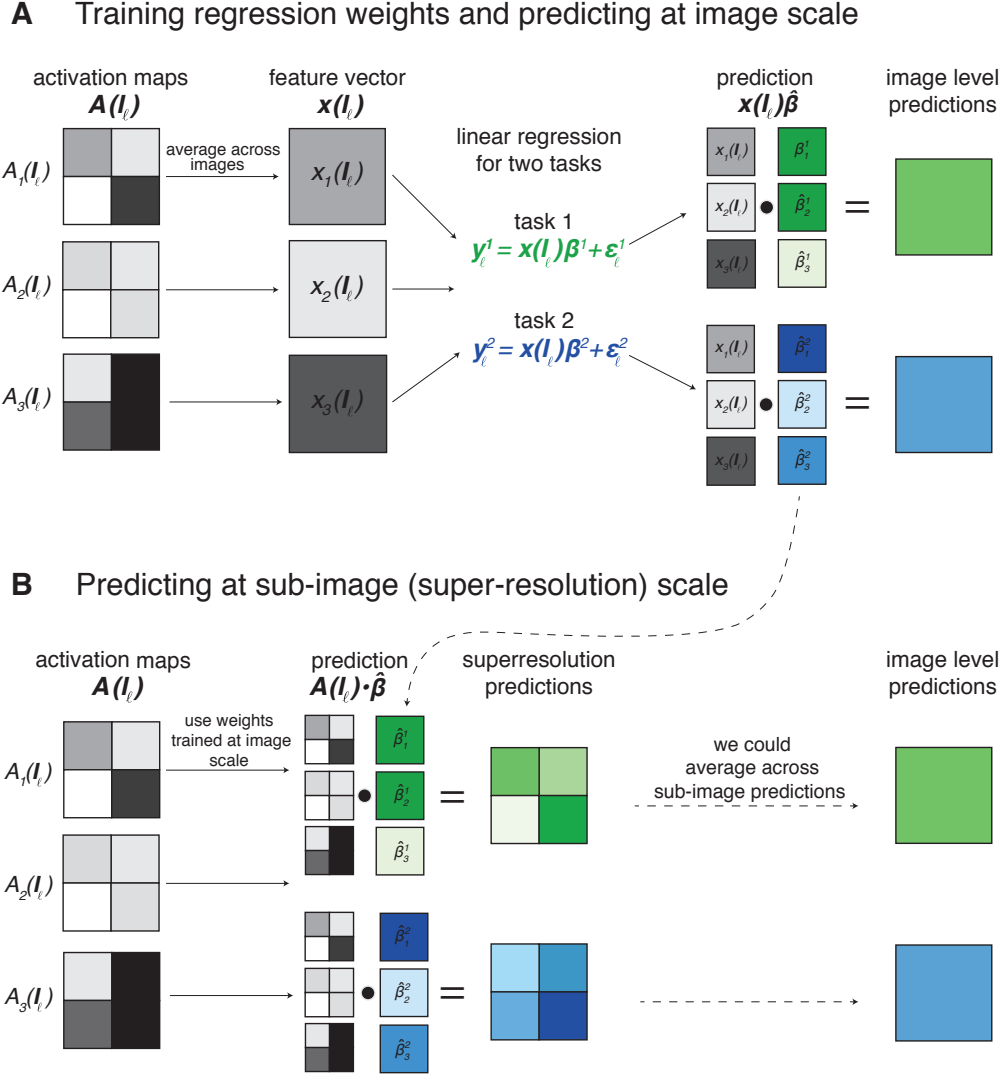


Figure S11: **Illustration of the procedure to construct predictions at image resolution and super-resolution.** Panel A illustrates the standard MOSAICS prediction pipeline. After convolution with random patches, nonlinear activation maps $\mathbf{A}_k(\mathbf{I}_\ell)$ are averaged across images to construct a set of image-level features $\mathbf{x}_k(\mathbf{I}_\ell)$ used in linear regression to generate predictions at image-scale (Section S.3.3). Panel B illustrates how the weights trained using labels and features at image-scale in panel A can be used to generate predictions at resolutions higher than the images and labeled data, achieving predictions at super-resolution. The scalar product of the entire activation map $\mathbf{A}_k(\mathbf{I}_\ell)$ and the estimated weights vector $\hat{\beta}$ generates super-resolution predictions at any desired sub-image scale larger than pixel-level. The last column of panel B illustrates the fact that super-resolution predictions, when averaged across an image, are identical to predictions generated from the standard process in panel A.

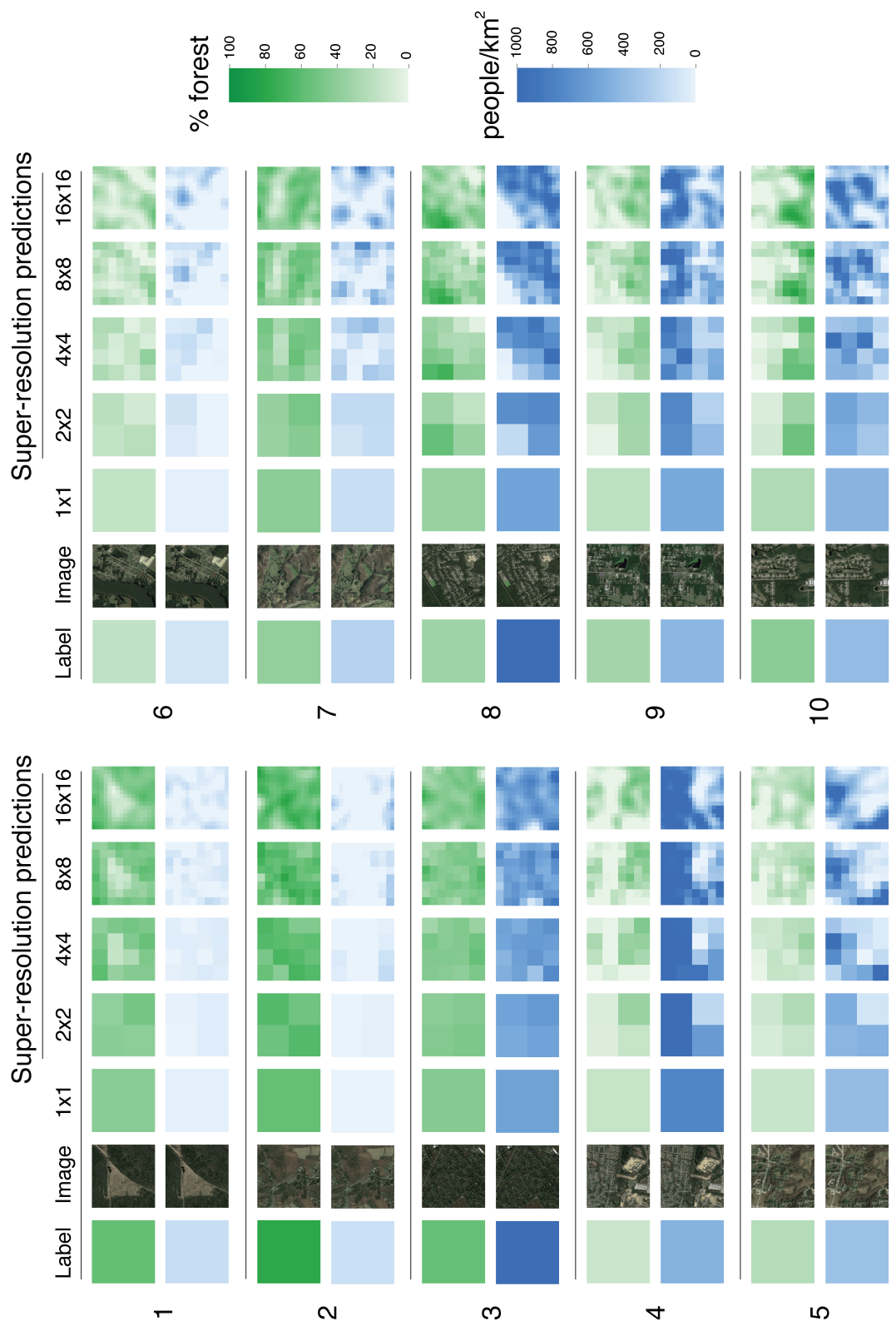


Figure S12: **Super-resolution performance across ten randomly selected images.** Each set of images indicate the image-level labels (column 1), the image itself (column 2) and predicted outcomes from MOSAICS at increasing levels of super-resolution (columns 3-7). These ten examples were selected uniformly at random from images in which our labels indicated at least 10% total forest cover and at least 100 people/km².

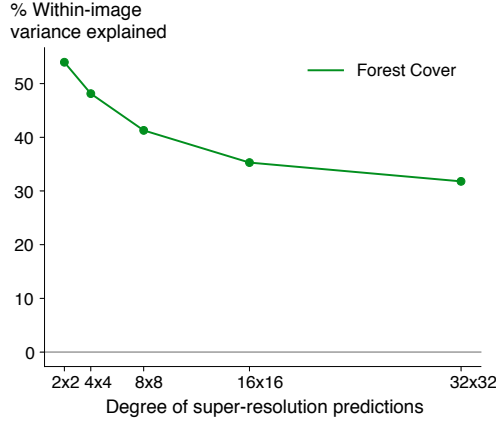


Figure S13: Systematic evaluation of *within-image* R^2 recovered in the forest cover task.

Evaluating super-resolution performance

To systematically evaluate the ability of MOSAIKS to accurately predict outcomes at super-resolution, we evaluate the within-image label variation that MOSAIKS’s super-resolution predictions accurately explain. We use forest cover for this test because the raw label resolution is substantially finer than the grid cell used to construct labels (see Section S.3.1 and Fig. S4), so we are able to attach “true” labels to super-resolution predictions within each image. In our main analysis, we construct grid cell forest cover labels by averaging fine-resolution raw forest cover data (see Section S.3.2). Here we leverage the fine resolution of the raw data to compare super-resolution performance of a model trained on aggregated labels but tested on high-resolution raw forest cover data.

Specifically, we learn regression weights $\hat{\beta}^s$ using a ridge regression applied to image-level labels from the full U.S. UAR sample ($N = 100,000$). We do so for multiple regularization parameters, λ , then make super-resolution predictions on a “validation” set of 1,000 images²³. We use the R^2 score for 32x downscaled predictions (32x32 predictions per image) to choose optimal values for λ and σ (the Gaussian filter length scale).²⁴ Next, we use the weights derived from our image-level ridge regression, along with the corresponding optimal λ and σ , to make super-resolution predictions for 16,000 additional images drawn randomly from the full set of 100,000 (excluding the 1,000 used for choosing hyperparameters). Lastly, we aggregate these pixel-level predictions to coarser sub-image scales, where increasing aggregation (lower super-resolution factor) reduces noise in the predictions at the cost of lower resolution.

We assess the performance of super-resolution at a variety of scales by calculating the percent of the variance of the raw within-image forest cover labels that can be explained by the super-resolution predictions at each scale. For example, to assess the performance of

²³Note that none of the pixel-level values in this validation set are used in the ridge regression, but the corresponding image-level labels are in-sample.

²⁴The optimal $\lambda = 1e5$ is higher than that chosen to optimize image-level predictions (Fig. 2.2), likely due to increased noise in sub-image predictions.

2×2 super-resolution predictions, we average predictions from the 254×254 super-resolution predictions by quadrants, resulting in four predicted values (twice the original resolution).²⁵ We perform the same per-quadrant average for the raw fine-resolution forest cover labels. We demean both the within-image predictions and labels to eliminate across-image variation, thereby focusing this test on the ability of the predictions to explain residual within-image variation. We then concatenate these within-image predictions and labels across the $N = 16,000$ images, so that the resulting R^2 value reported is the percent of super-resolution label variance explained by super-resolution predictions, across $64,000 = 16,000 \cdot 2^2$ label-prediction pairs.

The resulting performance of super-resolution predictions at different scales is shown in Fig. S13 for width scales of 2×2 , 4×4 , 8×8 , 16×16 , and 32×32 . We test up to $w = 32$ because the native width of the forest cover labels ($\sim 30\text{m}$) is just under $1/32$ the width of the original image ($\sim 1\text{km}$). Super-resolution predictions are trained only on the aggregate label at the image-level. Nonetheless, as Fig. S13 shows, we are able to explain over 50% of the within-image label variations at 2×2 super-resolution, and over 30% of the variation using 32×32 super-resolution grids.

Comparisons to other within-image prediction algorithms The most similar approach in the literature to MOSAIKS’s super-resolution predictions are methods specifically designed for pixel-level classification, or *semantic labelling* of satellite imagery (Firat, Can, and Vural 2014; Volpi and Tuia 2017). However, these approaches make use of sub-image labels for training, as opposed to our setting, where only one label per image (per task) is provided. Such semantic labelling approaches tend to use a downsample-then-upsample approach inspired by auto-encoders (Vincent et al. 2010) to learn lower-dimensional latent representations which are then up sampled to image-size prediction maps from which per-pixel classifications can be made. The upsampling procedure introduces more parameters to be tuned during model training, as well as additional computational cost in producing predictions. We again contrast this complex machinery with the simplicity of MOSAIKS’s approach, which calculates super-resolution predictions as a weighted sum of activation maps.

Conditions where super-resolution is most easily interpretable The linear decomposition of Eq. (2.4) holds when using labels that represent the average or sum of values within a grid cell, such as forest cover, elevation, population density, nighttime lights, income, or road length. However, it does not hold exactly when values are transformed nonlinearly after aggregation (e.g. $\log(\sum y) \neq \sum \log(y)$).²⁶ In these cases, the interpretation of super-resolution estimates requires care. Another case in which the interpretation of the sub-image predictions is difficult is when an image-level characteristic is not directly the sum of sub-image parcels. For instance, when predicting mean housing price in a grid cell, a manicured park might contribute to a higher value, yet that component of the image does not, in it-

²⁵For the analysis, we clip the images and predictions to 224×224 pixels so they are evenly divisible by a 32×32 super-resolution factor.

²⁶This issue could be addressed – in the case of logged variables – if one obtained a geometric mean image-level outcome rather than an arithmetic mean.

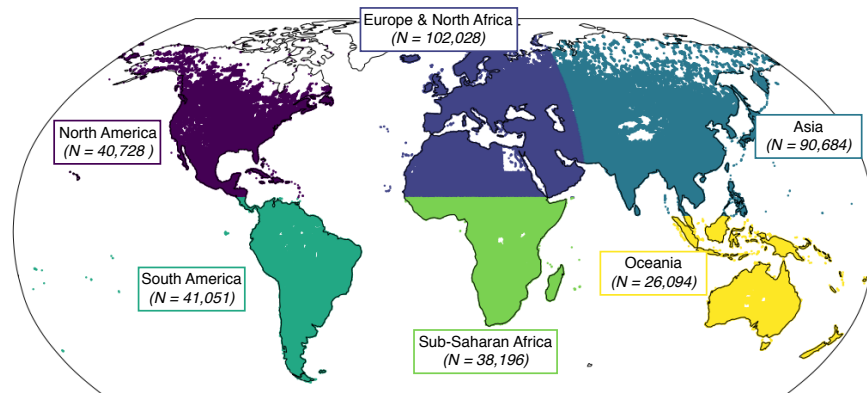


Figure S14: **Continent samples used to solve four tasks at global scale.** MOSAIKS predictions at global scale are generated from six separate cross-validated ridge regressions using random convolutional features. Each continent model is trained on 80% of the sample size shown (N).

self, have any associated housing price. In this case, we would interpret the sub-image predictions as “contributions to grid cell mean housing price” rather than the more natural interpretation as simply “a finer resolution prediction of housing price.”

S.3.10 Global model

For our global analysis, we create a global grid, composed of roughly 420 million cells just over 1km^2 in size, using an identical structure to that described in Section S.3.1 for the US. To obtain observations for our global analysis, we sub-sample 1,000,000 cells from this grid, sampling UAR from non-marine grid cells. This relatively sparse sampling of global data is due to the cost of obtaining imagery data.

One of the difficulties in sub-sampling from the global grid is that there are many grid cells where no Google imagery is available (there are negligibly few missing images in the US grid). After discarding grid cells with missing imagery from our original sample of 1,000,000 observations, we are left with $N = 498,063$ valid observations. After removing observations for which labeled data are missing for any of the tasks we analyze at global scale (forest cover, elevation, population density, and nighttime lights), we are left with $N = 423,476$ observations, which we use to train/validate (80%, $N = 338,781$) and test (20% $N = 84,692$) the model.

When generating features ($K = 2,048$) for our global model, we conduct featurization as described in Section S.3.3. Note that to create the global features we use patches drawn randomly from the *global* sample of images, not just from within the US.

When training the global model, we follow the approach outlined in Sec. S.3.5, solving for grid cell labels as a linear function of the random convolutional features using ridge regression and cross-validation to tune the regularization parameter λ . However, recovered regression weights are likely to differ across regions of the globe due to heterogeneity in

image quality and in visual signal of task labels or their derivatives. Therefore, we divide our global sample into six continental regions before solving each task. The continents (and sample sizes used for training and testing each continent-specific model) are shown in Fig. S14.

Modeling heterogeneity using the continents shown in Figure S14 leads to meaningful gains in performance over ignoring continent effects. As shown in the main text, the approach accounting for heterogeneity generates R^2 values of 0.85, 0.45, 0.62, and 0.49, for forest cover, elevation, population density, and nighttime lights, respectively (Fig. 2.4). In contrast, a global model that pools all observations across the globe and solves for a single linear function of random convolutional features generates R^2 values of 0.80, 0.26, 0.48, and 0.41, for the same tasks.

S.3.11 Generalizing to other ACS variables

Here we demonstrate the ability of MOSAIKS to generalize rapidly across a range of new variables by replicating our primary analysis (i.e. that in Fig. 2.2) for 12 variables from the American Community Survey, the source we use in the main text to measure income across the continental US. This survey is conducted annually across the US, tracking a diverse range of socioeconomic outcomes, from housing information to income and education. For this exercise, we select variables from the ACS that span a range of diverse outcomes and which seem likely to have at least some visible signal in daytime satellite imagery. We report performance for all tested variables.

We calculate grid cell level labels from census block group level ACS data using the same method used for ACS income label construction outlined in Sec. S.3.2. The resulting labels represent the area-weighted average value of the outcome across the grid cell. The ACS variables we predict are listed and described in Table S4.

Patterns in performance across tasks could be explained by the hypothesis that some outcomes exhibit more visible features, such as age and value of housing, while other outcomes exhibit less clear visible signal, such as the percent of household income dedicated to rent.

This exercise shows the ease of generalizing MOSAIKS to new contexts. In total, training these twelve predictive models took less than 45 minutes on a workstation with ten cores (IntelXeon CPU E5-263) (Table S5). Given the similarity in performance between MOSAIKS and other state of the art approaches documented in Sec. S.4.1, MOSAIKS offers a relatively quick and easy way to determine how predictable a variable might be from high resolution visible satellite imagery.

S.4 Comparisons to other models

Here, we compare the predictive performance and computational cost of MOSAIKS to other approaches in the literature.

Name	Code	Description
Travel time to work	B08303	“Travel time (minutes) to work refers to the total number of minutes that it usually took the worker to get from home to work during the reference week. The elapsed time includes time spent waiting for public transportation, picking up passengers in carpools, and time spent in other activities related to getting to work.”
Percent Bachelor’s Degree	B15003	Calculated as the number of people over 25 with only bachelor’s degrees (i.e. not masters or doctorate) divided by the total number of people over 25.
Median Household Income	B19013	Median household income in the past 12 Months (2015 inflation-adjusted dollars)
Per Capita Income	B19301	Per capita income in the Past 12 Months (2015 inflation-adjusted dollars)
Percent below poverty level	C17002	Calculated as the number of people 15 years or older whose income fell below the poverty level divided by the total number of people 15 years or older.
Percent food stamp/snap	B22010	Percent household received food stamps/snap in the past 12 months.
Median income	B25071	Gross rent as a percentage of household income in the past 12 months (dollars)
Number of housing units	B25001	“A housing unit may be a house, an apartment, a mobile home, a group of rooms or a single room that is occupied (or, if vacant, intended for occupancy) as separate living quarters. Separate living quarters are those in which the occupants live separately from any other individuals in the building and which have direct access from outside the building or through a common hall. Both occupied and vacant housing units are included in the housing unit inventory. Boats, recreational vehicles (RVs), vans, tents, railroad cars, and the like are included only if they are occupied as someone’s current place of residence.”
Percent vacant	B25002	“A housing unit is vacant if no one is living in it at the time of interview.”
Structure age	B25035	Data reported is the median year structure built. We calculate structure age as 2015 – median year structure built.
Number of rooms	B25017	“For each unit, rooms include living rooms, dining rooms, kitchens, bedrooms, finished recreation rooms, enclosed porches suitable for year-round use, and lodger’s rooms. Excluded are strip or pullman kitchens, bathrooms, open porches, balconies, halls or foyers, half-rooms, utility rooms, unfinished attics or basements, or other unfinished space used for storage.”
Median house value	B25077	For owner-occupied housing units.

Table S4: Description of variables from the American Community Survey (ACS) used in the analysis (Fig. 2.4). Quoted descriptions of variables are from: <https://censusreporter.org/topics/table-codes/>.

S.4.1 Benchmarking performance

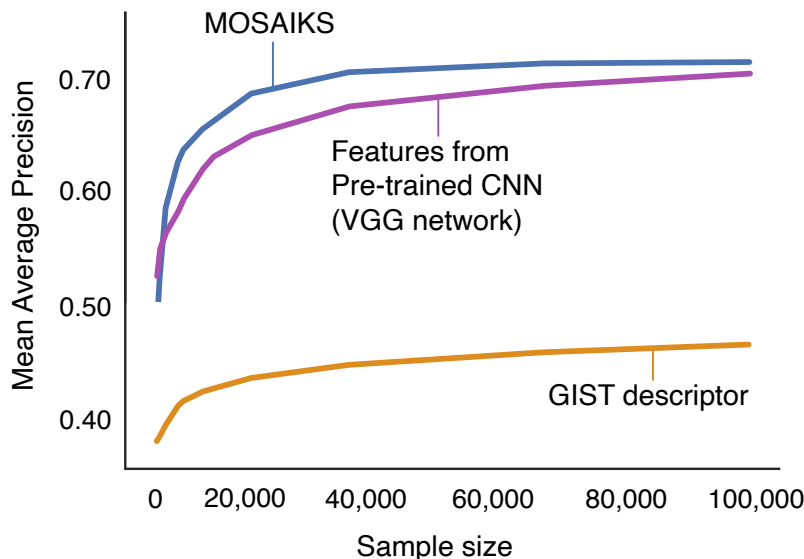


Figure S15: **Comparison of out-of-sample performance across feature extraction techniques as a function of sample size.** Mean Average Precision is shown for the out-of-sample prediction of housing price class (low, medium, high) for all single-family home sales after 2010 in Arizona, as a function of training sample size. Three feature extraction techniques are compared: MOSAIKS(blue), a pre-trained CNN (VGG, purple), and the GIST descriptor (orange). Figure reproduced from (Bolliger et al. 2017).

Convolutional neural networks (CNNs) have become the default “gold standard” in many image recognition tasks (Krizhevsky, Sutskever, and G. E. Hinton 2012), and are increasingly used in remote sensing applications (Jean, Burke, et al. 2016; Volpi and Tuia 2017; Gechter et al. 2018; Perez et al. 2017; Robinson, Hohman, and Dilkina 2017; Penatti et al. 2015; Jean, S. Wang, et al. 2019; Zhong et al. 2017; W. Hu et al. 2019; Maggiori et al. 2017). Simultaneously, alternative generalizable and computationally efficient pipelines have been developed that incorporate unsupervised featurization and/or a classification or regression algorithm (Cheng, Han, and X. Lu 2017; Cheriadat 2014; Penatti et al. 2015; Inglada et al. 2017; Pérez-Suay et al. 2017). MOSAIKS is low-cost and generalizable like these latter models; however, unlike these other models, it offers accuracies for regression problems competitive with that of leading CNN architectures. Here we quantitatively assess the predictive performance of MOSAIKS relative to (a) a CNN trained end-to-end with the outcomes of interest, (b) a similarly cheap, unsupervised featurization used in place of random convolutional features in the MOSAIKS infrastructure and (c) a transfer learning approach. For (b), we use the features generated by the last hidden layer of a pre-trained variant of the CNN (trained on natural imagery). This common approach is unsupervised in that the weights of the CNN are not trained using the labels of the outcome of interest, and such an approach has been shown to have better predictive performance than many other unsupervised featurization

algorithms (e.g. GIST, SIFT, Bag of Visual Words) on satellite image tasks (Cheng, Han, and X. Lu 2017). Previous analyses show through direct comparison that our methodology significantly outperforms ridge regression models using GIST features (Bolliger et al. 2017). Fig S15 (reproduced from (Bolliger et al. 2017)) demonstrates this comparison, describing out-of-sample performance for the prediction of housing price class for homes in Arizona.

These exercises compare MOSAIKS performance to that of models suited to the data availability of different prediction domains (abundant within the U.S., and relatively scarce at dispersed locations globally). A fine tuned CNN is expected to perform well in the United States, where data are relatively high quality and the sample size is large (nearly a hundred thousand observations); whereas the transfer learning approach is designed to perform well in regions where the training data are more coarse and the sample sizes are smaller (hundreds of observations). The ability of MOSAIKS to perform on par with these approaches in each setting demonstrates its generalizability.

Comparison to a deep convolutional neural network and an alternative unsupervised featurization First, we compare the performance of MOSAIKS to that of a tuned Residual Network (ResNet)(He et al. 2016) – a common, versatile deep network architecture used in recent satellite-based learning tasks (Perez et al. 2017). We train this network *end to end* to predict outcomes in all seven tasks across the continental US, using as input the same imagery used by MOSAIKS.

Specifically, we train an 18-layer variant of the ResNet Architecture pre-trained on ImageNet using stochastic gradient descent to minimize the mean squared error (MSE) between the predictions and labels with an initial learning rate of 0.001 with a single decay milestone at 10 epochs, and momentum parameter of 0.9. We train the model for 50 epochs, at which point performance approaches an asymptote. The optimal values for learning rates were tuned on a validation set for the task of predicting population density. We employ a standard train/test split of 80%/20%, matching our approach when evaluating MOSAIKS.

Second, we compare MOSAIKS performance to a similarly cheap, unsupervised featurization generated by the last hidden layer of a pre-trained variant of the CNN used above, trained on natural imagery. To execute this comparison, we use the features from the last layer of a 152-layer variant of the ResNet Architecture, and then run ridge regression on these features for each task.

Fig. 2.3A in the main text compares the holdout accuracy of MOSAIKS to both alternative approaches, demonstrating that MOSAIKS (dark bars) achieves performance competitive with ResNet (middle bars) across all seven tasks, while providing substantially greater performance than ridge regression run on features from the pre-trained CNN (lightest bars).

Comparison to a transfer learning approach We also compare the performance of MOSAIKS to that of a transfer learning approach in which nighttime lights observations are used to tune a CNN that was pre-trained on ImageNet. The tuned CNN is then used to extract features from the satellite images and a linear model is trained to predict the outcome

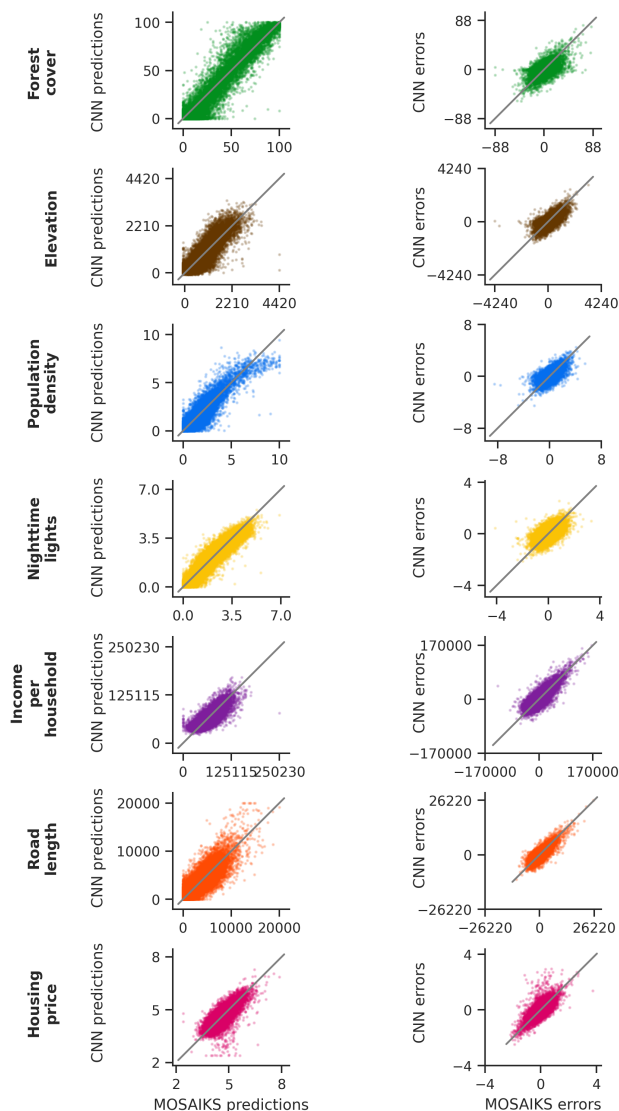


Figure S16: **Comparison of predictions and prediction errors between MOSAIKS and the ResNet-18 CNN.** The left column shows the relationship between predictions generated by MOSAIKS (x -axis) and predictions generated by the ResNet-18 CNN (y -axis). The right column shows the relationship between prediction errors from MOSAIKS (x -axis) and prediction errors from the CNN (y -axis). In both plots, each point indicates one grid cell ($\sim 1\text{km} \times 1\text{km}$) in the holdout test set; the test set sample size is approximately 20,000 for each task, although sample sizes vary somewhat due to data availability across tasks (Section S.3.5).

of interest. This approach leverages a large number of nighttime lights observations to better learn how to extract information from satellite imagery that is meaningful to tasks that may be reflected in nighttime lights (e.g. wealth). Comparing the performance of MOSAIKS to that of transfer learning tests the value of learning these features from nighttime lights,

relative to the unsupervised featurization of MOSAIKS.

We compare the performance of MOSAIKS to the transfer learning approach by replicating a subset of the analyses in (Jean, Burke, et al. 2016) and (Head et al. 2017). Using MOSAIKS, we predict wealth, electricity, mobile phone ownership, education, bed net count, female body mass index, water access, and hemoglobin level in Haiti, Nepal and Rwanda; we additionally predict child weight percentile, child height percentile and child weight for height percentile in Rwanda. These variables are recorded at geo-located “cluster” locations by the Demographic and Health Survey (DHS); the survey methodology is detailed in (Jean, Burke, et al. 2016). The variables and countries we provide performance metrics for were chosen based on the facility of obtaining and matching images and labels from the original authors and their replication code bases. We report performance for all tested variables and countries.

In this analysis, we use two MOSAIKS-based models to predict the DHS cluster labels. First, we use only the MOSAIKS random convolutional features (indicated as RCF), which we calculate for each image as detailed in Sec. S.3.3, and then average over the 100 images associated with each DHS cluster (see (Jean, Burke, et al. 2016) and (Head et al. 2017) for details on matching images to clusters; we use the same matching approach as the original authors). In a second model, denoted MOSAIKS-NL below, we use the MOSAIKS random convolutional features along with features based on nighttime lights. The nighttime light features for each cluster are counts of the number of nightlight values that fall within a set of 19 bins, as well as the minimum, mean and maximum of the values within the image. Bins were evenly spaced on a log scale from a luminosity of 0.1 to 500 (in units of nanoWatts/cm²/sr). We average nighttime light features for all 100 images associated with each DHS cluster, as we do for RCF.

We show results for the MOSAIKS-nighttime lights model for two reasons. First, it presents the most fair comparison to the transfer learning approach, which also leverages both nighttime lights and daytime imagery. Second, it demonstrates the ability of MOSAIKS to seamlessly combine information from different sensors – by appending their features in a linear model – to make predictions.

Training a model that uses features from multiple sensors A key benefit of the MOSAIKS approach is that it can easily combine information from multiple sensors. Recall that to train a model when using only the RCF from visual imagery we regress the outcome y_ℓ^s for each task s on features $\mathbf{x}_\ell$ as follows:

$$y_\ell^s = \mathbf{x}(\mathbf{I}_\ell)\boldsymbol{\beta}^s + \epsilon_\ell^s$$

And solve for $\boldsymbol{\beta}^s$ by minimizing the sum of squared errors plus an l_2 regularization term:

$$\min_{\boldsymbol{\beta}^s} \frac{1}{2} \|\mathbf{y}_\ell^s - \mathbf{x}(\mathbf{I}_\ell)\boldsymbol{\beta}^s\|_2^2 + \frac{\lambda^s}{2} \|\boldsymbol{\beta}^s\|_2^2$$

To include features from an additional sensor, S_ℓ , such as nighttime lights, one simply generates a new set of features, $\mathbf{z}(\mathbf{S}_\ell)$, – using the RCF algorithm or any other unsupervised

featurization approach – and includes the features in the regression model, giving:

$$y_\ell^s = \mathbf{x}(\mathbf{I}_\ell)\boldsymbol{\beta}^s + \mathbf{z}(\mathbf{S}_\ell)\boldsymbol{\gamma}^s + \epsilon_\ell^s$$

Then, one solves for $\boldsymbol{\beta}^s$ and $\boldsymbol{\gamma}^s$ by minimizing the sum of squared errors plus individual regularization terms for each sensor:

$$\min_{\boldsymbol{\beta}^s, \boldsymbol{\gamma}^s} \frac{1}{2} \|y_\ell^s - \mathbf{x}(\mathbf{I}_\ell)\boldsymbol{\beta}^s - \mathbf{z}(\mathbf{S}_\ell)\boldsymbol{\gamma}^s\|_2^2 + \frac{\lambda_1^s}{2} \|\boldsymbol{\beta}^s\|_2^2 + \frac{\lambda_2^s}{2} \|\boldsymbol{\gamma}^s\|_2^2$$

Regularizing the features from each sensor separately enables the model to treat features from individual sensors differently, which we found improves model performance. We implement a model that combines RCF and features from nighttime lights in Fig. S17. Features from additional sensors could be added to the model in a similar way.

Interpretation of test accuracy comparisons Note that the performance of these models represents a reasonable lower bound on potential performance; some task-specific enhancements could be used to improve predictive power for each of these methods. For example, more layers could be added to ResNet or alternative architectures could be tested for specific tasks. In the case of MOSAIKS and the pre-trained ResNet features, more flexible regression models could be used to estimate the relationship between features and labels, such as increasing K , using a nonlinear model, or leveraging a hurdle model in tasks with a large number of zero observations. While these task-specific changes may marginally improve performance of any of these approaches, prior research on similar image recognition tasks suggests further gains for the ResNet are likely to be minimal (Zoph and Le 2016). While the similarity of performance in Fig. 2.3A is perhaps surprising, it is also encouraging for further research. This comparison suggests that wide, shallow networks using local-level features (analogous to random convolutional features) are as descriptive as more complex, highly optimized CNN architectures for satellite remote sensing, across many tasks.

The similarity in performance of MOSAIKS and ResNet across tasks (Fig. 2.3A) is consistent with a hypothesis that both approaches are reaching the limit of information that is provided by satellite imagery for predicting the outcomes we test. Fig. S16 provides further evidence of such a hypothesis, as the test set predictions and prediction errors are highly correlated across ResNet and MOSAIKS in all tasks studied. A human prediction baseline has not been established, but could provide insight into whether there is substantial room for improvement in skill for each of these tasks, although we suspect many of these tasks will be difficult for non-expert humans (e.g. nightlights or house prices).

S.4.2 Comparing costs

In practice, high computational costs can limit the use of SIML methods – especially when resources are scarce, such as in government agencies of low-income countries (Haack and Ryerson 2016) or research teams and NGOs with limited budgets. Specifically designed to

<i>Component</i>	ResNet Time (GPU)	MOSAICS Time
Training set featurization ($N = 80k$)		~ 1.2 hours (GPU)
Model training	$\sim \mathbf{7.9}$ hours	~ 2.8 seconds (GPU) ~ 50 seconds (10 cores) $\sim \mathbf{1.8}$ minutes (laptop)
Holdout set featurization ($N = 20k$)		~ 18 minutes (GPU)
Holdout set prediction	$\sim \mathbf{40}$ seconds	< 0.01 seconds (GPU) ~ 0.1 seconds (10 cores) $\sim \mathbf{0.7}$ seconds (laptop)
Total cost to ecosystem	$\sim \mathbf{7.9}$ hours	~ 1.5 hours (GPU)
Total cost to user	$\sim \mathbf{7.9}$ hours	~ 2.8 seconds (GPU) ~ 50.1 seconds (10 cores) $\sim \mathbf{1.8}$ minutes (laptop)

Table S5: **Wall-clock times of components of MOSAICS compared with a fine-tuned CNN.** Bold times are those that a practitioner using each method would incur (assuming MOSAICS users have access to a standard laptop only). Model training time includes training *after* tuning for a single task for both ResNet and MOSAICS. MOSAICS was run using $K=8,192$ features. ResNet operations were run on an Amazon EC2 p3.2xlarge instance with a Tesla V100 GPU and 60GB of onboard RAM. Cost of computation on this machine is roughly \$3/hr. MOSAICS operations are shown for runs on this same GPU, a local workstation with ten cores (Intel Xeon CPU E5-263), and a standard laptop (MacBook Pro).

address this challenge, MOSAICS scales across many research tasks by decoupling featurization from task selection, model-fitting, and prediction. The computationally costly step of featurization is done centrally on a fast computer with a graphics processing unit (GPU); individual practitioners need only download the pre-computed features, merge on labels for the task they select, and run regressions. Because features are created and stored by a central entity, the research community makes use of a cached set of computations, reducing the overall computational burden of widespread SIML and any external social costs generated by these computations (Strubell, Ganesh, and McCallum 2019). Additionally, this decoupling of task-agnostic computations from task-specific computations allows practitioners to run more diagnostic analyses on their tasks, such as those presented in Fig. 2.3 of the main text.

From the perspective of a user who can access pre-computed MOSAICS features to train and validate a new task, we find that MOSAICS is $\sim 250\times$ to $10,000\times$ faster than a state-of-the-art neural net architecture (ResNet), depending on the computational resources available to a MOSAICS user (Table S5). Moreover, MOSAICS performance is competitive with the ResNet on all tasks we have studied (Fig 2.3A). From the perspective of the entire computational ecosystem, which bears the cost of image featurization in addition to model

training and testing, we find that MOSAIKS is $5.3\times$ faster than the ResNet when solving a single task. The relative efficiency of MOSAIKS grows with the number of tasks studied because MOSAIKS features can be reused across tasks. For the ResNet, the times in Table S5 reflect our wall-clock time on a single Amazon EC2 instance for a single task, so that the time costs are similar to that of introducing a single new domain *ex post*. For MOSAIKS, Table S5 includes wall-clock times on three different computational platforms, as users may have access to different resources. We show times using the same GPU as we use for the ResNet comparisons, times on a local workstation with ten cores (Intel Xeon CPU E5-263), and times on a standard laptop (MacBook Pro). For both ResNet and MOSAIKS, we report in Table S5 model training time *after* using cross-validation to select optimal hyperparameters. For MOSAIKS, model training time on the local workstation with 10 cores is ~ 6.8 minutes when including cross-validation to select penalization parameters in ridge regression. The ecosystem-wide costs of featurization per task shown in Table S5 decline as MOSAIKS becomes more widely adopted, because features can be cached centrally and distributed without modification to multiple users who are training and/or testing SIML in common locations.

We considered only one CNN architecture, which we chose because of its use in previous remote sensing applications (Jean, Burke, et al. 2016). We did not attempt to innovate in neural net architectural design or algorithms. While one could pursue targeted innovations in neural networks for remote sensing, such as in ref. (Zhong et al. 2017), we emphasize that our method is currently orders of magnitude faster for the user than off-the-shelf fine-tuned CNN methods (Table S5), does not require a GPU for prediction, and achieves competitive prediction performance (Fig. 2.3A). There is recent work that aims to train networks to learn a “common representation” that can generalize across tasks, but this is a subject of ongoing research (Ruder 2017), requires the tasks to be known in advance, and has yet to be demonstrated or evaluated at scale.

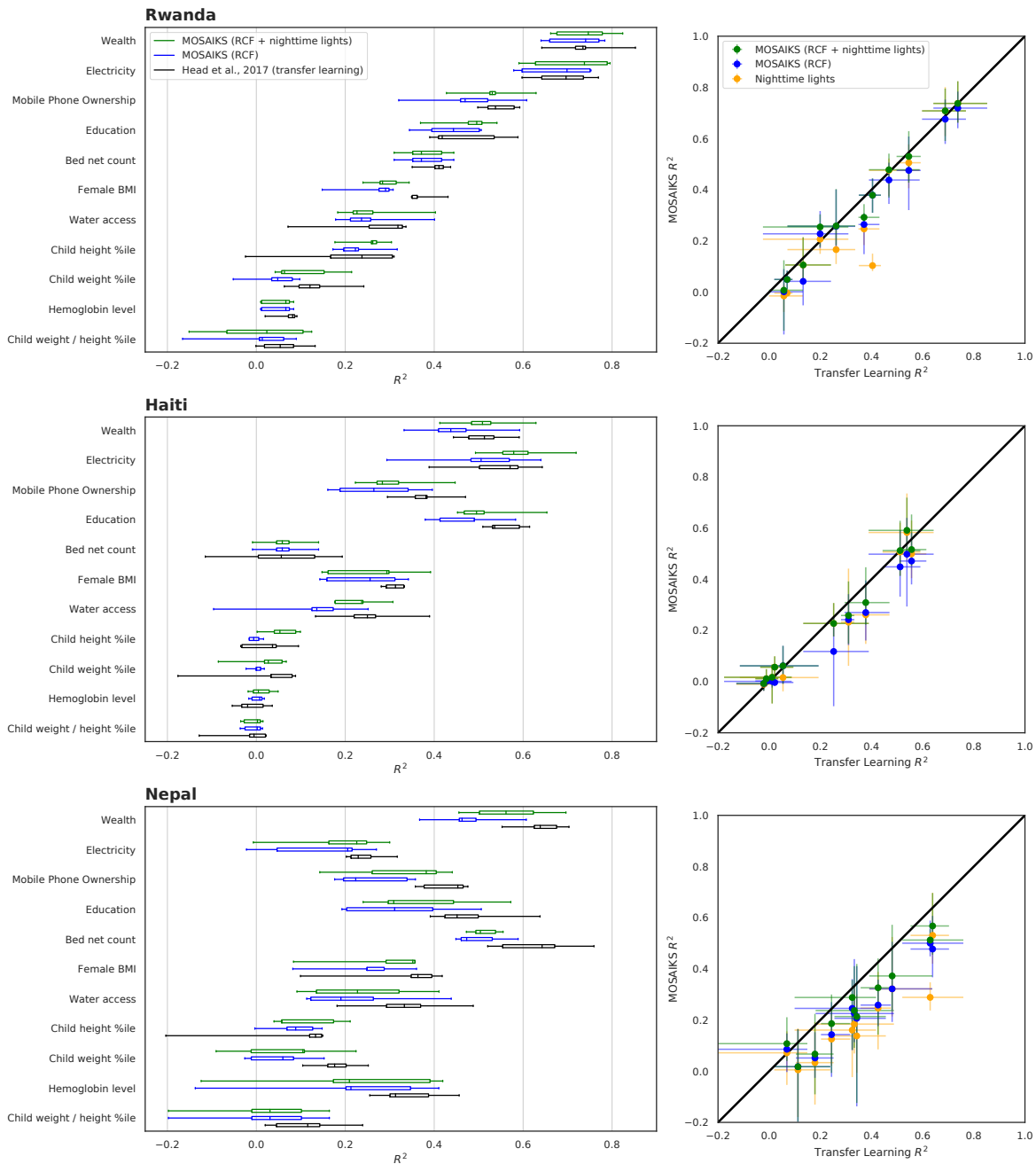


Figure S17: **Comparison of accuracy between MOSAIKS and a transfer learning model.** Box plots (left) show task-specific performance of MOSAIKS models (RCF in blue and RCF + nighttime lights in green) in contrast to a transfer learning model (black). Box and whiskers show the performance over the 5 cross-validation folds. Scatter plots (right) show the performance of MOSAIKS models, as well as a nighttime lights-only model (orange) versus the transfer learning model performance. Each point in the scatter is the average R^2 over the 5 cross-validation folds, while whiskers indicate the full range of performance across folds.

Chapter 3

The Effect of Large-Scale Anti-Contagion Policies on the COVID-19 Pandemic

Governments around the world are responding to the novel coronavirus (COVID-19) pandemic (F. Wu et al. 2020) with unprecedented policies designed to slow the growth rate of infections. Many actions, such as closing schools and restricting populations to their homes, impose large and visible costs on society, but their benefits cannot be directly observed and are currently understood only through process-based simulations.(Ferguson et al. 2020; Chinazzi et al. 2020; Kraemer et al. 2020) Here, we compile new data on 1,717 local, regional, and national non-pharmaceutical interventions deployed in the ongoing pandemic across localities in China, South Korea, Italy, Iran, France, and the United States (US). We then apply reduced-form econometric methods, commonly used to measure the effect of policies on economic growth,(Greene 2003; C. D. Romer and D. H. Romer 2010) to empirically evaluate the effect that these anti-contagion policies have had on the growth rate of infections. In the absence of policy actions, we estimate that early infections of COVID-19 exhibit exponential growth rates of roughly 38% per day. We find that anti-contagion policies have significantly and substantially slowed this growth. Some policies have different impacts on different populations, but we obtain consistent evidence that the policy packages now deployed are achieving large, beneficial, and measurable health outcomes. We estimate that across these six countries, interventions prevented or delayed on the order of 62 million confirmed cases, corresponding to averting roughly 530 million total infections. These findings may help inform whether or when these policies should be deployed, intensified, or lifted, and they can support decision-making in the other 180+ countries where COVID-19 has been reported.(World Health Organization 2020)

This chapter is adapted from (S. Hsiang, Allen, et al. 2020). The author of this thesis managed the overall project, policy data collection, and the projects source code (<https://github.com/bolliger32/gpl-covid>); collected geographic and population data for all countries; developed the theory linking the econometric analysis to epidemiological models; designed and implemented the epidemiological models and projections; designed and implemented model robustness checks; created Extended Data Figures 8 and 9; and adapted our analysis pipeline to run on the reproducibility platform Code Ocean. Solomon Hsiang conceived of and led the overall study, and led the writing of the published manuscript. A full list of author contributions is listed with the published manuscript.

3.1 Introduction

The COVID-19 pandemic is forcing societies worldwide to make consequential policy decisions with limited information. After containment of the initial outbreak failed, attention turned to implementing non-pharmaceutical interventions designed to slow contagion of the virus. In general, these policies aim to decrease virus transmission by reducing contact among individuals within or between populations, such as by closing restaurants or restricting travel, thereby slowing the spread of COVID-19 to a manageable rate. These large-scale anti-contagion policies are informed by epidemiological simulations (Kraemer et al. 2020; R. Li et al. 2020; Ferguson et al. 2020; Tang et al. 2020) and a small number of natural experiments in past epidemics (Hatchett, Mecher, and Lipsitch 2007). However, the actual effects of these policies on infection rates in the ongoing pandemic are unknown. Because the modern world has never confronted this pathogen, nor deployed anti-contagion policies of such scale and scope, it is crucial that direct measurements of policy impacts be used alongside numerical simulations in current decision-making.

Societies around the world are weighing whether the health benefits of anti-contagion policies are worth their social and economic costs. Many of these costs are plainly seen; for example, business restrictions increase unemployment and school closures impact educational outcomes. It is therefore not surprising that some populations have hesitated before implementing such dramatic policies, especially when their costs are visible while their health benefits – infections and deaths that would have occurred but instead were avoided or delayed – are unseen. Our objective is to measure the direct health benefits of these policies; specifically, how much these policies slowed the growth rate of infections. To do this, we compare the growth rate of infections within hundreds of sub-national regions before and after each of these policies is implemented locally. Intuitively, each administrative unit observed just prior to a policy deployment serves as the “control” for the same unit in the days after it receives a policy “treatment” (see Supplementary Information for accounts of these deployments). Our hope is to learn from the recent experience of six countries where early spread of the virus triggered large-scale policy actions, in part so that societies and decision-makers in the remaining 180+ countries can access this information.

Here we directly estimate the effects of 1,717 local, regional, and national policies on the

growth rate of infections across localities within China, France, Iran, Italy, South Korea, and the US (see Figure 3.1 and Supplementary Table 1). We compile subnational data on daily infection rates, changes in case definitions, and the timing of policy deployments, including (1) travel restrictions, (2) social distancing through cancellations of events and suspensions of educational/commercial/religious activities, (3) quarantines and lockdowns, and (4) additional policies such as emergency declarations and expansions of paid sick leave, from the earliest available dates to April 6, 2020 (see Supplementary Notes, also Extended Data Fig. 1). During this period, populations remained almost entirely susceptible to COVID-19, causing the natural spread of infections to exhibit almost perfect exponential growth.(Ma 2020; Muniz-Rodriguez et al. 2020) The rate of this exponential growth could change daily, determined by epidemiological factors, such as disease infectivity, as well as policies that alter behavior (Gerardo Chowell et al. 2016; Ma 2020; Tang et al. 2020). Because policies were deployed while the epidemic unfolded, we can estimate their effects empirically. To do this, we examine how the daily growth rate of infections in each locality changes in response to the collection of ongoing policies applied to that locality on that day.

3.2 Methods Summary

We employ well-established “reduced-form” econometric techniques(Angrist and Pischke 2008; Greene 2003) commonly used to measure the effects of events(C. D. Romer and D. H. Romer 2010; Burke, Solomon M Hsiang, and Miguel 2015) on economic growth rates. Similar to early COVID-19 infections, economic output generally increases exponentially with a variable rate that can be affected by policies and other conditions. Here, this technique aims to measure the total magnitude of the effect of *changes in policy*, without requiring explicit prior information about fundamental epidemiological parameters or mechanisms, many of which remain uncertain in the current pandemic. Rather, the collective influence of these factors is empirically recovered from the data without modeling their individual effects explicitly (see Methods). Prior work on influenza(Kandula et al. 2018), for example, has shown that such statistical approaches can provide important complementary information to process-based models.

To construct the dependent variable, we transform location-specific, subnational time-series data on infections into first-differences of their natural logarithm, which is the *per-day growth rate of infections* (see Methods). We use data from first- or second-level administrative units and data on active or cumulative cases, depending on availability (see Supplementary Information). We employ widely-used panel regression models(Angrist and Pischke 2008; Greene 2003) to estimate how the daily growth rate of infections changes over time within a location when different combinations of large-scale policies are enacted (see Methods). Our econometric approach accounts for differences in the baseline growth rate of infections across subnational locations, which may be affected by time-invariant characteristics, such as demographics, socio-economic status, culture, and health systems; it accounts for systematic patterns in growth rates within countries unrelated to policy, such as the effect

of the work-week; it is robust to systematic under-surveillance specific to each subnational unit; and it accounts for changes in procedures to diagnose positive cases (see Methods and Supplementary Information).

3.3 Results

We estimate that in the absence of policy, early infection rates of COVID-19 grow 43% per day on average across these six countries (Standard Error [SE]= 5%), implying a doubling time of approximately 2 days. Country-specific estimates range from 34% per day in the US (SE= 7%) to 68% per day in Iran (SE= 9%). We cannot determine if the high estimate for Iran results from true epidemiological differences, data quality issues (see Methods), the concurrence of the initial outbreak with a major religious holiday and pilgrimage (see Supplementary Notes), or sampling variability. Excluding Iran, the average growth rate is 38% per day (SE= 5%). Growth rates in all five other countries are independently estimated to be very near this value (Figure 3.2a). These estimated values differ from observed average growth rates because the latter are confounded by the effects of policy. These growth rates are not driven by the expansion of testing or increasing rates of case detection (see Methods and Extended Data Fig. 2) nor by data from individual regions (Extended Data Fig. 3).

Some prior analyses of pre-intervention infections in Wuhan suggest slower growth rates (doubling every 5–7 days)(J. T. Wu et al. 2020; Q. Li et al. 2020) using data collected before national standards for diagnosis and case definitions were first issued by the Chinese government on January 15, 2020.(Tsang et al. 2020) However, case data in Wuhan from before this date contain multiple irregularities: the cumulative case count decreased on January 9; no new cases were reported during January 9–15; and there were concerns that information about the outbreak was suppressed(*Wuhan pneumonia: 30 days from outbreak to out of control* 2020) (see Supplementary Table 2). When we remove these problematic data, utilizing a shorter but more reliable pre-intervention time series from Wuhan (January 16–21), we recover a growth rate of 43% per day (SE= 3%, doubling every 2 days) consistent with results from all other countries except Iran (Figure 3.2a, Supplementary Table 3).

During the early stages of an epidemic, a large proportion of the population remains susceptible to the virus, and if the spread of the virus is left uninhibited by policy or behavioral change, exponential growth continues until the fraction of the susceptible population declines meaningfully.(Fisman, Khoo, and Tuite 2014; Maier and Brockmann 2020; Gerardo Chowell et al. 2016; Ma 2020) After correcting for estimated rates of case-detection (Russell et al. 2020), we compute that the minimum susceptible fraction across administrative units in our sample is 72% of the total population (Cremona, Italy) and 87% of units would likely be in a regime of uninhibited exponential growth ($> 95\%$ susceptible) if policies were removed on the last date of our sample.

Consistent with predictions from epidemiological models(Bootsma and Ferguson 2007; Ferguson et al. 2020; Hatchett, Mecher, and Lipsitch 2007), we find that the combined effect of policies within each country reduces the growth rate of infections by a substantial

and statistically significant amount (Figure 3.2b, Supplementary Table 3). For example, a locality in France with a baseline growth rate of 0.33 (national average) that fully deployed all policy actions used in France would be expected to lower its daily growth rate by -0.17 to a growth rate of 0.16. In general, the estimated total effects of policy packages are large enough that they can in principle offset a large fraction of, or even eliminate, the baseline growth rate of infections—although in several countries, many localities have not deployed the full set of policies. Overall, the estimated effects of all policies combined are generally insensitive to withholding regional (i.e. state- or province-level) blocks of data from the sample (Extended Data Fig. 3).

In China, only three policies were enacted across 116 cities early in a seven week period, providing us with sufficient data to empirically estimate how the effects of these policies evolved over time without making assumptions about the timing of these effects (see Methods and Fig. 3.2b). We estimate that the combined effect of these policies reduced the growth rate of infections by -0.026 (SE= 0.046) in the first week following their deployment, increasing substantially in the second week to -0.20 (SE= 0.049), and essentially stabilizing in the third week near -0.28 (SE= 0.047). In other countries, we lack sufficient data to estimate these temporal dynamics explicitly and only report the average pooled effect of policies across all days following their deployment (see Methods). If other countries have transient responses similar to China, we would expect effects in the first week following deployment to be smaller in magnitude than the average effect we report. In Extended Data Fig. 5a and Supplementary Methods Section 3, we explore how our estimates would change if we impose the assumption that policies cannot affect infection growth rates until after a fixed number of days; however, we do not find evidence this improves model fit.

The estimates above (Figure 3.2b) capture the superposition of all policies deployed in each country, i.e., they represent the average effect of policies that we would expect to observe if all policies enacted anywhere in each country were implemented simultaneously in a single region of that country. We also estimate the effects of individual policies or clusters of policies (Figure 3.2c) that are grouped based on either their similarity in goal (e.g., library and museum closures) or timing (e.g., policies deployed simultaneously). Our estimates for these individual effects tend to be statistically noisier than the estimates for all policies combined. Some estimates for the same policy differ between countries, perhaps because policies are not implemented identically or because populations behave differently. Nonetheless, 22 out of 29 point estimates indicate that individual policies are likely contributing to reducing the growth rate of infections. Seven policies (one in South Korea, two in Italy, and four in the US) have point estimates that are positive, six of which are small in magnitude (< 0.1) and not statistically different from zero (5% level). Consistent with greater overall uncertainty in these dis-aggregated estimates, some in China, South Korea, Italy, and France are somewhat more sensitive to withholding regional blocks of data (Extended Data Fig. 4), but remain broadly robust to assuming a constant delayed effect of all policies (Extended Data Fig. 5b).

Based on these results, we find that the deployment of anti-contagion policies in all six countries significantly and substantially slowed the pandemic. We combine the estimates above with our data on the timing of the 1,717 policy deployments to estimate the total

effect of all policies across the dates in our sample. To do this, we use our estimates to predict the growth rate of infections in each locality on each day, given the actual policies in effect at that location on that date (Figure 3.3, blue markers). We then use the same model to predict what counterfactual growth rates would be on that date if the effects of all policies were removed (Figure 3.3, red markers), which we call the “no-policy scenario.” The difference between these two predictions is our estimated effect that all deployed policies had on the growth rate of infections. During our sample, we estimate that all policies combined slowed the average growth rate of infections by -0.252 per day ($SE = 0.045$, $p < 0.001$) in China, -0.248 ($SE = 0.089$, $p < 0.01$) in South Korea, -0.24 ($SE = 0.068$, $p < 0.001$) in Italy, -0.355 ($SE = 0.063$, $p < 0.001$) in Iran, -0.123 ($SE = 0.019$, $p < 0.001$) in France and -0.084 ($SE = 0.03$, $p < 0.01$) in the US. These results are robust to modeling the effects of policies without grouping them (Extended Data Fig. 6a and Supplementary Table 4) or assuming a delayed effect of policy on infection growth rates (Supplementary Table 5).

The number of COVID-19 infections on a date depends on the growth rate of infections on all prior days. Thus, persistent reductions in growth rates have a compounding effect on infections, until growth is slowed by a shrinking susceptible population. To provide a sense of scale for our results, we integrate the growth rate of infections in each locality from Figure 3.3 to estimate cumulative infections, both with actual anti-contagion policies and in the no-policy counterfactual scenario. To account for the declining susceptible population in each administrative unit, we couple our econometric estimates of the effects of policies with a Susceptible-Infected-Removed (SIR) model (Ma 2020; Gerardo Chowell et al. 2016) that adjusts the susceptible population in each administrative unit based on estimated case-detection rates (Russell et al. 2020; Meyerowitz-Katz and Merone 2020) (see Methods). This allows us to extend our projections beyond the initial exponential growth phase of infections, a threshold that many localities cross in our no-policy scenario.

Our results suggest that ongoing anti-contagion policies have already substantially reduced the number of COVID-19 infections observed in the world today (Figure 3.4). Our central estimates suggest that there would be roughly 37 million more cumulative confirmed cases (corresponding to 285 million more total infections, including the confirmed cases) in China, 11.5 million more confirmed cases in South Korea (38 million total infections), 2.1 million more confirmed cases in Italy (49 million total infections), 5 million more confirmed cases in Iran (54 million total infections), 1.4 million more confirmed cases in France (45 million total infections), and 4.8 million more confirmed cases (60 million total infections) in the US had these countries never enacted any anti-contagion policies since the start of the pandemic. The magnitudes of these impacts partially reflect the timing, intensity, and extent of policy deployment (e.g., how many localities deployed policies), and the duration for which they have been applied. Several of these estimates are subject to large statistical uncertainties (see intervals in Figure 3.4). Sensitivity tests (Extended Data Fig. 7) that assume a range of plausible alternative parameter values relating to disease dynamics, such as incorporating a Susceptible-Exposed-Infected-Removed (SEIR) model, suggest that interventions may have reduced the severity of the outbreak by a total of 55–66 million confirmed cases over the dates in our sample (central estimates). Sensitivity tests varying the

assumed infection-fatality ratio (Supplementary Table 6) suggest a corresponding range of 46–77 million confirmed cases (490–580 million total infections).

3.4 Discussion

Our empirical results indicate that large-scale anti-contagion policies are slowing the COVID-19 pandemic. Because infection rates in the countries we study would have initially followed rapid exponential growth had no policies been applied, our results suggest that these policies have provided large health benefits. For example, we estimate that there would be roughly $465\times$ the observed number of confirmed cases in China, $17\times$ in Italy, and $14\times$ in the US by the end of our sample if large-scale anti-contagion policies had not been deployed. Consistent with process-based simulations of COVID-19 infections (Kraemer et al. 2020; Tang et al. 2020; Ferguson et al. 2020; Maier and Brockmann 2020; R. Li et al. 2020; Kucharski et al. 2020), our analysis of existing policies indicates that seemingly small delays in policy deployment likely produced dramatically different health outcomes.

While the limitations of available data pose challenges to our analysis, our aim is to use what data exist to estimate the first-order impacts of unprecedented policy actions in an ongoing global crisis. As more data become available, related findings will become more precise and may capture more complex interactions. Furthermore, this analysis does not account for interactions between populations in nearby localities, (Gerardo Chowell et al. 2016) nor mobility networks (R. Li et al. 2020; Tang et al. 2020; Chinazzi et al. 2020; Kraemer et al. 2020). Nonetheless, we hope these results can support critical decision-making, both in the countries we study and in the other 180+ countries where COVID-19 infections have been reported. (World Health Organization 2020)

A key advantage of our reduced-form “top down” statistical approach is that it captures the real-world behavior of affected populations without requiring that we explicitly model underlying mechanisms and processes. This is useful in the current pandemic where many process-related parameters remain uncertain. However, our results cannot and should not be interpreted as a substitute for “bottom up” process-based epidemiological models specifically designed to provide guidance in public health crises. Rather, our results complement existing models, for example, by helping to calibrate key model parameters. We believe both forward-looking simulations and backward-looking empirical evaluations should be used to inform decision-making.

Our analysis measures changes in local infection growth rates associated with changes in anti-contagion policies. A necessary condition for this association to be interpreted as the plausibly causal effect of these policies is that the timing of policy deployment is independent of infection growth rates (Angrist and Pischke 2008). This assumption is supported by established epidemiological theory (Gerardo Chowell et al. 2016; Ma 2020; Anderson and May 1992) and evidence (Nishiura et al. 2010; WHO Ebola Response Team 2014), which indicate that infections in the absence of policy will grow exponentially early in the epidemic, implying that pre-policy infection growth rates should be constant over time and

therefore uncorrelated with the timing of policy deployment. Further, scientific guidance to decision-makers early in the current epidemic explicitly projected constant growth rates in the absence of anti-contagion measures, limiting the possibility that anticipated changes in natural growth rates affected decision-making.(Ferguson et al. 2020; Flaxman et al. 2020; Lourenço et al. 2020; Maier and Brockmann 2020) In practice, policies tended to be deployed in response to high total numbers of cases (e.g. in France)(Ministère des Solidarités et de la Santé 2020), in response to outbreaks in other regions (e.g. in China, South Korea, and Iran) (H. Tian et al. 2020), after delays due to political constraints (e.g. in the US and Italy), and often with timing that coincided with arbitrary events, like weekends or holidays (see Supplementary Notes for detailed chronologies).

Our analysis accounts for documented changes in COVID-19 testing procedures and availability, as well as differences in case-detection across locations; however, unobserved trends in case-detection could affect our results (see Methods). We analyze estimated case-detection trends(Russell et al. 2020) (Extended Data Fig. 2), finding that this potential bias is small, possibly elevating our estimated no-policy growth rates by 0.022 (7%) on average.

It is also possible that changing public knowledge during the period of our study affects our results. If individuals alter behavior in response to new information unrelated to anti-contagion policies, such as seeking out online resources, this could alter the growth rate of infections and thus affect our estimates. If increasing availability of information reduces infection growth rates, it would cause us to overstate the effectiveness of anti-contagion policies. We note, however, that if public knowledge is increasing in response to policy actions, such as through news reports, then it should be considered a pathway through which policies alter infection growth, not a form of bias. Investigating these potential effects is beyond the scope of this analysis, but it is an important topic for future investigations.

Finally, our analysis focuses on confirmed infections, but other outcomes, such as hospitalizations or deaths, are also of policy interest. Future work on these outcomes may require additional modeling approaches because they are relatively more context- and state-dependent. Nonetheless, we experimentally implement our approach on the daily growth rate of hospitalizations in France, where hospitalization data is available at the granularity of this study. We find that the total estimated effect of anti-contagion policies on the growth rate of hospitalizations is similar to our estimates for infection growth rates (Extended Data Fig. 6c).

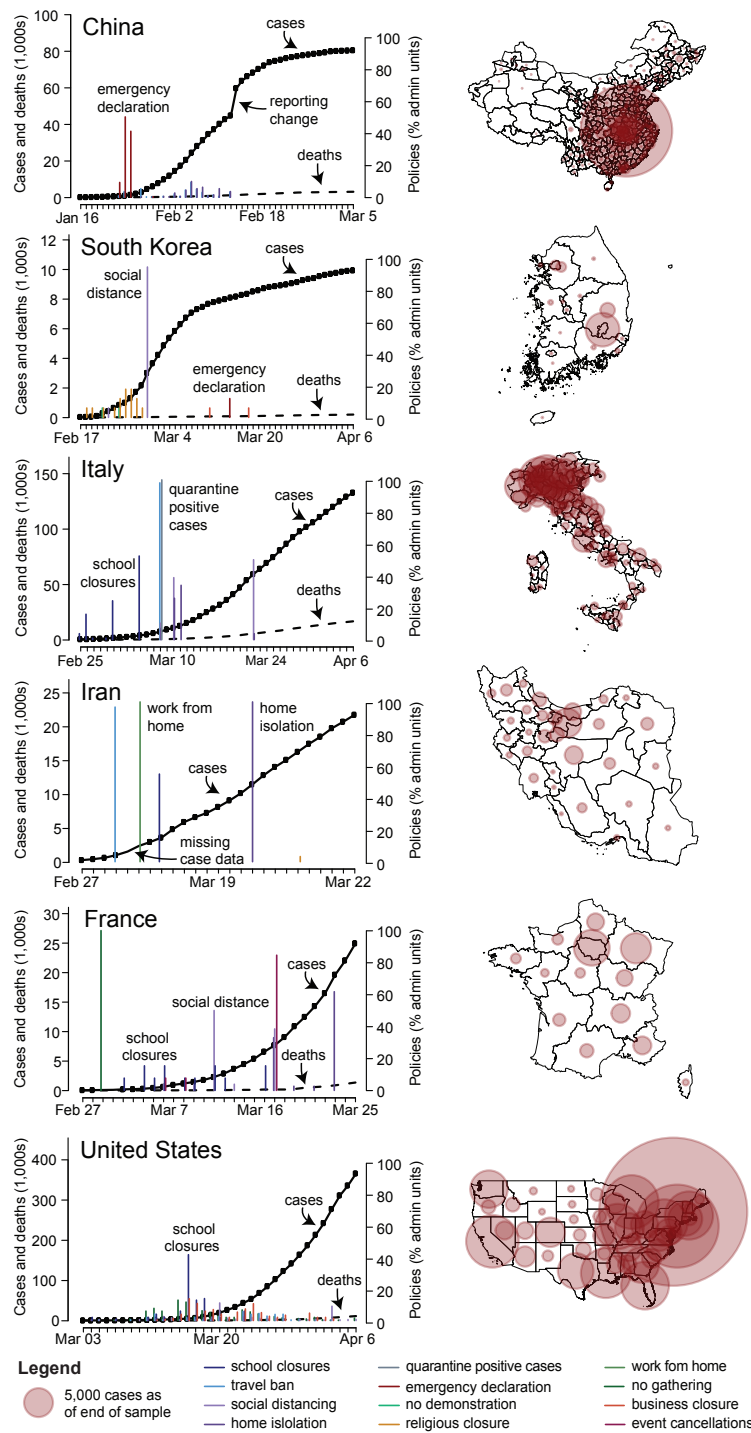
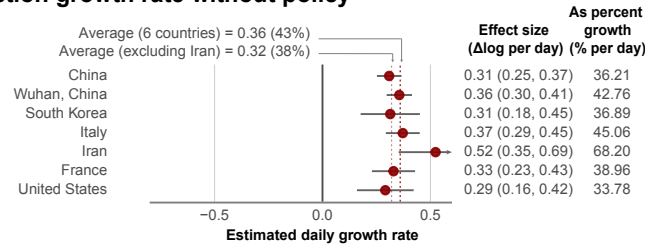
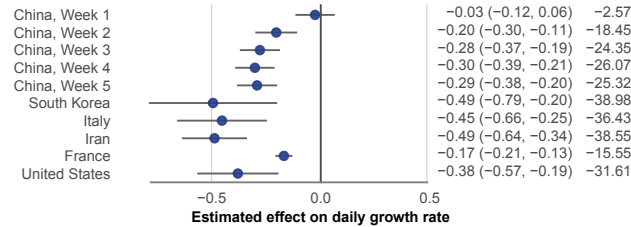


Figure 3.1: **Data on COVID-19 infections and large-scale anti-contagion policies.** Left: Daily cumulative confirmed cases of COVID-19 (solid black line, left axis) and deaths (dashed black line) over time. Vertical lines are deployments of anti-contagion policies, with height indicating the number of administrative units instituting a policy that day (right axis). For display purposes only, ≤ 5 policy types are shown per country and missing case data is imputed unless all sub-national units are missing. Right: Maps of cumulative confirmed cases by administrative unit on the last date of each sample.

a Infection growth rate without policy



b Effect of all policies combined



c Effect of individual policies

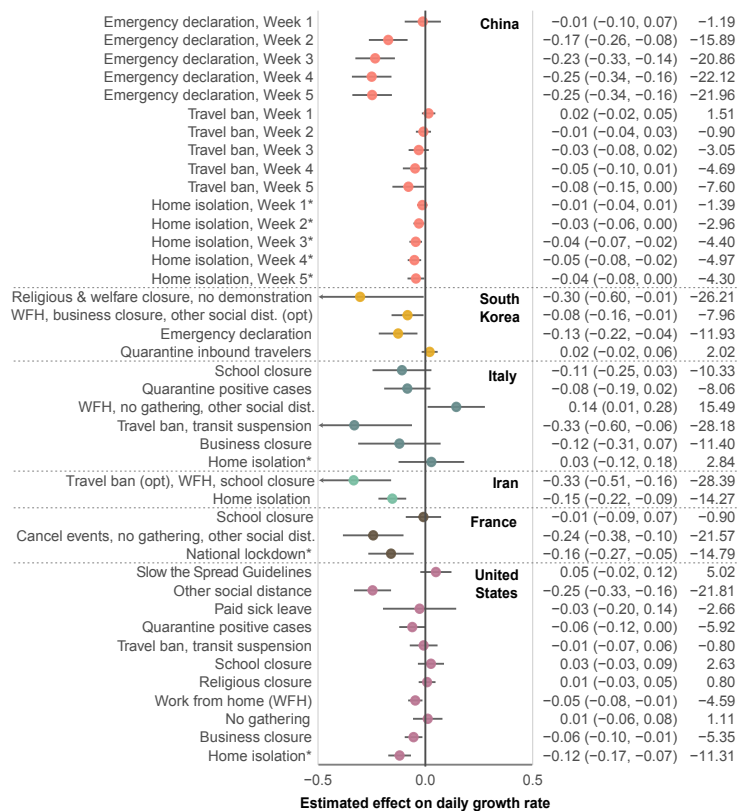


Figure 3.2: **Empirical estimates of unmitigated COVID-19 infection growth rates and the effect of anti-contagion policies.** Markers are country-specific estimates, whiskers are 95% CI. Columns report effect sizes as a change in the continuous-time growth rate (95% CI in parentheses) and the day-over-day percentage growth rate. (a) Estimates of daily COVID-19 infection growth rates in the absence of policy (dashed lines = averages with and without Iran, both excluding Wuhan-specific estimate). (b) Estimated combined effect of all policies on infection growth rates. (c) Estimated effects of individual policies or policy groups on the daily growth rate of infections, jointly estimated and ordered roughly chronologically within each country. *Reported effect of “home isolation” includes effects of other implied policies (see Methods). China: N=3669; South Korea: N=595; Italy: N=2898; Iran: N=548; France: N=270; US: N=1238.

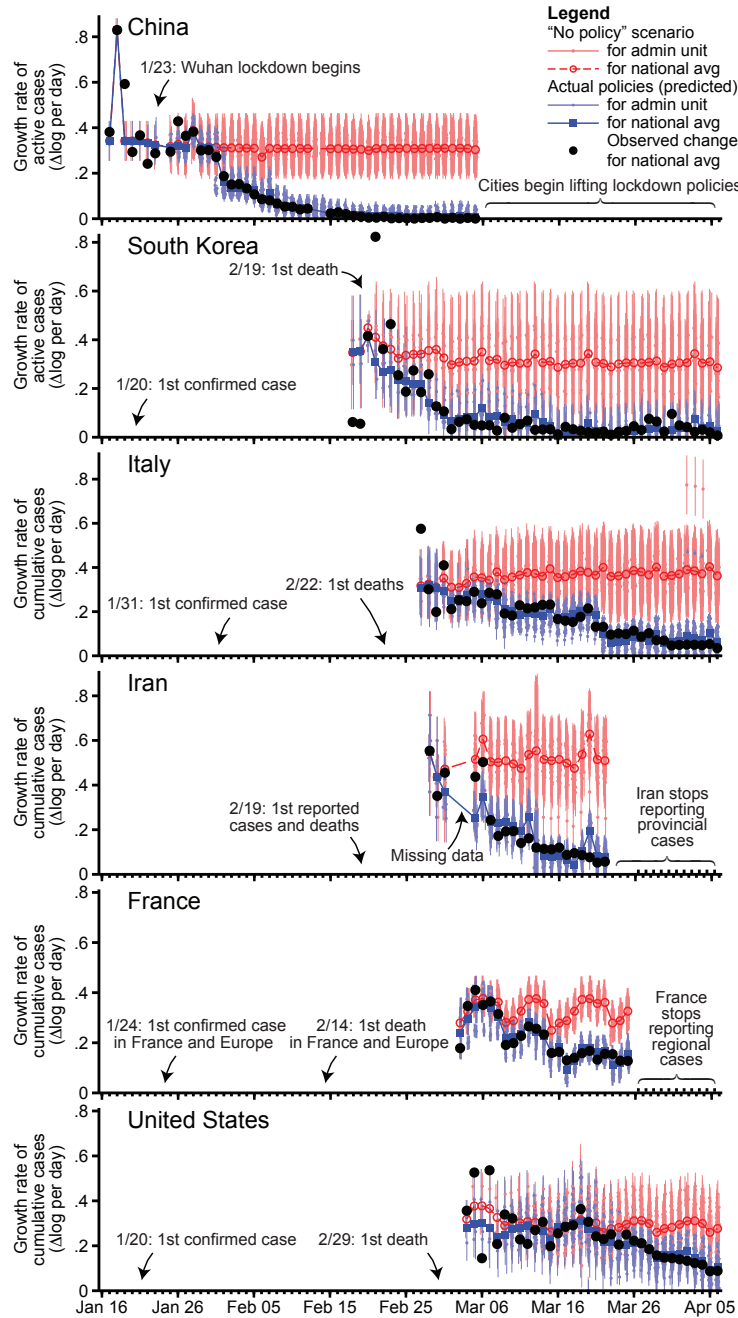


Figure 3.3: **Estimated infection growth rates based on actual anti-contagion policies and in a “no policy” counterfactual scenario.** Predicted daily growth rates of active (China, South Korea) or cumulative (all others) COVID-19 infections based on the observed timing of all policy deployments within each country (blue) and in a scenario where no policies were deployed (red). The difference between these two predictions is our estimated effect of actual anti-contagion policies on the growth rate of infections. Small markers are daily estimates for sub-national administrative units (vertical lines are 95% CI). Large markers are national averages. Black circles are observed daily changes in $\log(\text{infections})$, averaged across administrative units. Sample sizes are the same as Figure 3.2.

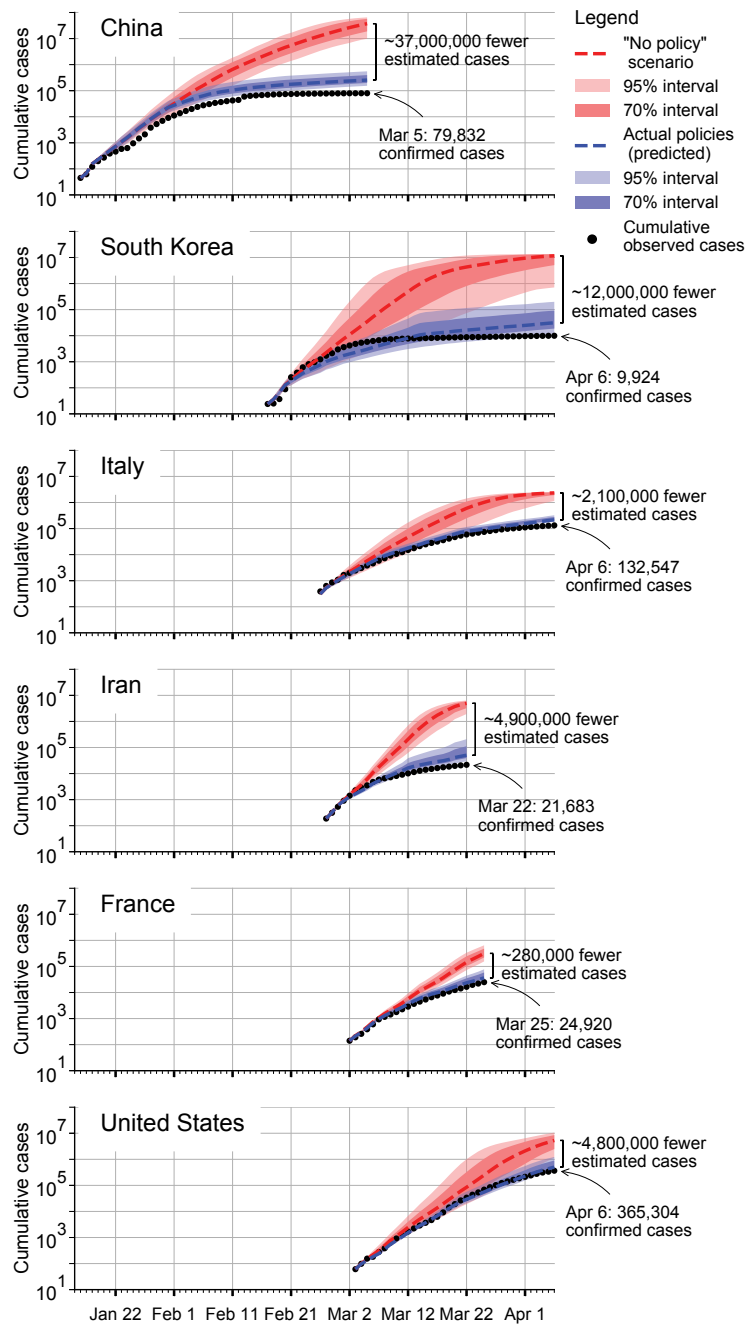


Figure 3.4: **Estimated cumulative confirmed COVID-19 infections with and without anti-contagion policies.** The predicted cumulative number of confirmed COVID-19 infections based on actual policy deployments (blue) and in the no-policy counterfactual scenario (red). Shaded areas show uncertainty based on 1,000 simulations where empirically estimated parameters are resampled from their joint distribution (dark = inner 70% of predictions; light = inner 95%). Black dotted line is observed cumulative infections. Infections are not projected for administrative units that never report infections in the sample, but which might have experienced infections in a no-policy scenario.

3.5 Methods

3.5.1 Data Collection and Processing

We provide a brief summary of our data collection processes here (see the Supplementary Notes for more details, including access dates). Epidemiological, case definition/testing regime, and policy data for each of the six countries in our sample were collected from a variety of in-country data sources, including government public health websites, regional newspaper articles, and crowd-sourced information on Wikipedia. The availability of epidemiological and policy data varied across the six countries, and preference was given to collecting data at the most granular administrative unit level. The country-specific panel datasets are at the region level in France, the state level in the US, the province level in South Korea, Italy and Iran, and the city level in China. Due to data availability, the sample dates differ across countries: in China we use data from January 16 - March 5, 2020; in South Korea from February 17 - April 6, 2020; in Italy from February 26 - April 6, 2020; in Iran from February 27 - March 22, 2020; in France from February 29 - March 25, 2020; and in the US from March 3 - April 6, 2020. Below, we describe our data sources.

China We acquired epidemiological data from an open source GitHub project (*COVID-19/2019-nCoV Time Series Infection Data Warehouse*) that scrapes time series data from Ding Xiang Yuan. We extended this dataset back in time to January 10, 2020 by manually collecting official daily statistics from the central and provincial (Hubei, Guangdong, and Zhejiang) Chinese government websites. We compiled policies by collecting data on the start dates of travel bans and lockdowns at the city-level from the “2020 Hubei lockdowns” Wikipedia page (*COVID-19 pandemic lockdown in Hubei* 2020) and various other news reports. We suspect that most Chinese cities have implemented at least one anti-contagion policy due to their reported trends in infections; as such, we dropped cities where we could not identify a policy deployment date to avoid miscategorizing the policy status of these cities. Thus our results are only representative for the sample of 116 cities for which we obtained policy data.

South Korea We manually collected and compiled the epidemiological dataset in South Korea, based on provincial government reports, policy briefings, and news articles. We compiled policy actions from news articles and press releases from the Korean Centers for Disease Control and Prevention (KCDC), the Ministry of Foreign Affairs, and local governments’ websites.

Iran We used epidemiological data from the table “New COVID-19 cases in Iran by province” (*COVID-19 pandemic in Iran* 2020) in the “2020 coronavirus pandemic in Iran” Wikipedia article, which were compiled from data provided on the Iranian Ministry of Health website (in Persian). We relied on news media reporting and two timelines of pandemic events

in Iran (Kantis, Kiernan, and Bardi 2020) (*COVID-19 pandemic in Iran* 2020) to collate policy data. From March 2-3, Iran did not report subnational cases. Around this period the country implemented three national policies: a recommendation against local travel (3/1), work from home for government employees (3/3), and school closure (3/5). As the effects of these policies cannot be distinguished from each other due to the data gap, we group them for the purpose of this analysis.

Italy We used epidemiological data from the GitHub repository (*COVID-19*) maintained by the Italian Department of Civil Protection (Dipartimento della Protezione Civile). For policies, we primarily relied on the English version of the COVID-19 dossier “Chronology of main steps and legal acts taken by the Italian Government for the containment of the COVID-19 epidemiological emergency” written by the Dipartimento della Protezione Civile, (*Coronavirus Emergency*) and Wikipedia. (*COVID-19 pandemic lockdown in Italy* 2020)

France We used the region-level epidemiological dataset provided by France’s government website (Roussel 2020) and supplemented it with numbers of confirmed cases by region on France’s public health website, which was previously updated daily through March 25. (*Santé Publique France*) We obtained data on France’s policy response to the COVID-19 pandemic from the French government website, press releases from each regional public health site, (*Agence Régionale de Santé: Agir pour la santé de tous*) and Wikipedia. (*COVID-19 pandemic in France* 2020)

United States We used state-level epidemiological data from usafacts.org, (*Coronavirus Locations: COVID-19 Map by County and State*) which they compile from multiple sources. For policy responses, we relied on a number of sources, including the U.S. Centers for Disease Control (CDC), the National Governors Association, as well as various executive orders from county- and city-level governments, and press releases from media outlets.

Policy Data Policies in administrative units were coded as binary variables, where the policy was coded as either 1 (after the date that the policy was implemented, and before it was removed) or 0 otherwise, for the affected administrative units. When a policy only affected a fraction of an administrative unit (e.g., half of the counties within a state), policy variables were weighted by the percentage of people within the administrative unit who were treated by the policy. We used the most recent population estimates we could find for countries’ administrative units (see the Population Data section in the Appendix). In order to standardize policy types across countries, we mapped each country-specific policy to one of the broader policy category variables in our analysis. In this exercise, we collected 168 policies for China, 59 for South Korea, 214 for Italy, 23 for Iran, 59 for France, and 1,194 for the United States (see Supplementary Table 1). There are some cases where we encode policies that are necessarily in effect whenever another policy is in place, due in particular to the far-reaching implications of home isolation policies. In China, wherever home isolation

is documented, we assume a local travel ban is enacted on the same day if we have not found an explicit local travel ban policy for a given locality. In France, we assume home isolation is accompanied by event cancellations, social distancing, and no-gathering policies; in Italy, we assume home isolation entails no-gathering, local travel ban, work from home, and social distancing policies; in the US, we assume shelter-in-place orders indicate that non-essential business closures, work from home policies, and no-gathering policies are in effect. For policy types that are enacted multiple times at increasing degrees of intensity within a locality, we add weights to the variable by escalating the intensity from 0 pre-policy in steps up to 1 for the final version of the policy (see the Policy Data section in the Appendix).

Epidemiological Data We collected information on cumulative confirmed cases, cumulative recoveries, cumulative deaths, active cases, and any changes to domestic COVID-19 testing regimes, such as case definitions or testing methodology. For our regression analysis (Figure 3.2), we use active cases when they are available (for China and South Korea) and cumulative confirmed cases otherwise. We document quality control steps in the Appendix. Notably, for China and South Korea we acquired more granular data than the data hosted on the Johns Hopkins University (JHU) interactive dashboard; (*COVID-19 Data Repository*) we confirm that the number of confirmed cases closely match between the two data sources (see Extended Data Fig. 1). To conduct the econometric analysis, we merge the epidemiological and policy data to form a single data set for each country.

3.5.2 Econometric analysis

Reduced-Form Approach The reduced-form econometric approach that we apply here is a “top down” approach that describes the behavior of aggregate outcomes y in data (here, infection rates). This approach can identify plausibly causal effects (Angrist and Pischke 2008; Greene 2003) induced by exogenous changes in independent policy variables z (e.g., school closure) without explicitly describing all underlying mechanisms that link z to y , without observing intermediary variables x (e.g., behavior) that might link z to y , or without other determinants of y unrelated to z (e.g., demographics), denoted w . Let $f(\cdot)$ describe a complex and unobserved process that generates infection rates y :

$$y = f(x_1(z_1, \dots, z_K), \dots, x_N(z_1, \dots, z_K), w_1, \dots, w_M) \quad (3.1)$$

Process-based epidemiological models aim to capture elements of $f(\cdot)$ explicitly, and then simulate how changes in z , x , or w affect y . This approach is particularly important and useful in forward-looking simulations where future conditions are likely to be different than historical conditions. However, a challenge faced by this approach is that we may not know the full structure of $f(\cdot)$, for example if a pathogen is new and many key biological and societal parameters remain uncertain. Crucially, we may not know the effect that large-scale policy (z) will have on behavior ($x(z)$) or how this behavior change will affect infection rates ($f(\cdot)$).

Alternatively, one can differentiate Equation 3.1 with respect to the k^{th} policy z_k :

$$\frac{\partial y}{\partial z_k} = \sum_{j=1}^N \frac{\partial y}{\partial x_j} \frac{\partial x_j}{\partial z_k} \quad (3.2)$$

which describes how changes in the policy affects infections through all N potential pathways mediated by $x_1, \dots, x_N$. Usefully, for a fixed population observed over time, empirically estimating an average value of the local derivative on the left-hand-side in Equation 3.2 does not depend on explicit knowledge of w . If we can observe y and z directly and estimate changes over time $\frac{\partial y}{\partial z_k}$ with data, then intermediate variables x also need not be observed nor modeled. The reduced-form econometric approach (Angrist and Pischke 2008; Greene 2003) thus attempts to measure $\frac{\partial y}{\partial z_k}$ directly, exploiting exogenous variation in policies z .

Model Active infections grow exponentially during the initial phase of an epidemic, when the proportion of immune individuals in a population is near zero. Assuming a simple Susceptible-Infected-Recovered (SIR) disease model (e.g., ref. [(Ma 2020)]), the growth in infections during the early period is

$$\frac{dI_t}{dt} = (S_t\beta - \gamma)I_t \underset{S_t \rightarrow 1}{=} (\beta - \gamma)I_t, \quad (3.3)$$

where I_t is the number of infected individuals at time t , β is the transmission rate (new infections per day per infected individual), γ is the removal rate (proportion of infected individuals recovering or dying each day) and S is the fraction of the population susceptible to the disease. The second equality holds in the limit $S \rightarrow 1$, which describes the current conditions during the beginning of the COVID-19 pandemic. The solution to this ordinary differential equation is the exponential function

$$\frac{I_{t_2}}{I_{t_1}} = e^{g \cdot (t_2 - t_1)}, \quad (3.4)$$

where I_{t_1} is the initial condition. Taking the natural logarithm and rearranging, we have

$$\log(I_{t_2}) - \log(I_{t_1}) = g \cdot (t_2 - t_1). \quad (3.5)$$

Anti-contagion policies are designed to alter g , through changes to β , by reducing contact between susceptible and infected individuals. Holding the time-step between observations fixed at one day ($t_2 - t_1 = 1$), we thus model g as a time-varying outcome that is a linear function of a time-varying policy

$$g_t = \log(I_t) - \log(I_{t-1}) = \theta_0 + \theta \cdot \text{policy}_t + \epsilon_t, \quad (3.6)$$

where θ_0 is the average growth rate absent policy, policy_t is a binary variable describing whether a policy is deployed at time t , and θ is the average effect of the policy on growth

rate g over all periods subsequent to the policy’s introduction, thereby encompassing any lagged effects of policies. ϵ_t is a mean-zero disturbance term that captures inter-period changes not described by $policy_t$. Using this approach, infections each day are treated as the initial conditions for integrating Equation 3.4 through to the following day.

We compute the first differences $\log(I_t) - \log(I_{t-1})$ using active infections where they are available, otherwise we use cumulative infections, noting that they are almost identical during this early period (except in China, where we use active infections). We then match these data to policy variables that we construct using the novel data sets we assemble and apply a reduced-form approach to estimate a version of Equation 3.6, although the actual expression has additional terms detailed below.

Estimation To estimate a multi-variable version of Equation 3.6, we estimate a separate regression for each country c . Observations are for subnational units indexed by i observed for each day t . Because not all localities began testing for COVID-19 on the same date, these samples are unbalanced panels. To ensure data quality, we restrict our analysis to localities after they have reported at least ten cumulative infections.

A necessary condition for unbiased estimates is that the timing of policy deployment is independent of natural infection growth rates (Angrist and Pischke 2008), a mathematical condition that should be true in the context of a new epidemic. In established epidemiological models, including the standard SIR model above, early rates of infection within a susceptible population are characterized by constant exponential growth. This phenomenon is well understood theoretically, (Gerardo Chowell et al. 2016), (Gerardo Chowell et al. 2016; Anderson and May 1992) (Kermack and McKendrick 1927) has been repeatedly documented in past epidemics (WHO Ebola Response Team 2014; Nishiura et al. 2010) (Mills, Robins, and Lipsitch 2004) as well as the current COVID-19 pandemic (Ma 2020), (Muniz-Rodriguez et al. 2020), and implies constant infection growth rates in the absence of policy intervention. Thus, we treat changes in infection growth rates as conditionally independent of policy deployments since the correlation between a constant variable and any other variable is zero in expectation.

We estimate a multiple regression version of Equation 3.6 using ordinary least squares. We include a vector of subnational unit-fixed effects θ_0 (i.e., varying intercepts captured as coefficients to dummy variables) to account for all time-invariant factors that affect the local growth rate of infections, such as differences in demographics, socio-economic status, culture, and health systems. (Greene 2003) We include a vector of day-of-week-fixed effects δ to account for weekly patterns in the growth rate of infections that are common across locations within a country, however, in China, we omit day-of-week effects because we find no evidence they are present in the data – perhaps due to the fact that the outbreak of COVID-19 began during a national holiday and workers never returned to work. We also include a separate single-day dummy variable each time there is an abrupt change in the availability of COVID-19 testing or a change in the procedure to diagnose positive cases. Such changes generally manifest as a discontinuous jump in infections and a re-scaling of

subsequent infection rates (e.g., See China in Figure 3.1), effects that are flexibly absorbed by a single-day dummy variable because the dependent variable is the first-difference of the logarithm of infections. We denote the vector of these testing dummies μ .

Lastly, we include a vector of P_c country-specific policy variables for each location and day. These policy variables take on values between zero and one (inclusive) where zero indicates no policy action and one indicates a policy is fully enacted. In cases where a policy variable captures the effects of collections of policies (e.g., museum closures and library closures), a policy variable is computed for each, then they are averaged, so the coefficient on this type of variable is interpreted as the effect if all policies in the collection are fully enacted. There are also instances where multiple policies are deployed on the same date in numerous locations, in which case we group policies that have similar objectives (e.g., suspension of transit and travel ban, or cancelling of events and no gathering) and keep other policies separate (i.e., business closure, school closure). The grouping of policies is useful for reducing the number of estimated parameters in our limited sample of data, allowing us to examine the impact of subsets of policies (e.g. Fig. 3.2c). However, policy grouping does not have a material impact on the estimated effect of all policies combined nor on the effect of actual policies, which we demonstrate by estimating a regression model where no policies are grouped and these values are recalculated (Supplementary Table 4, Extended Data Fig. 6).

In some cases (for Italy and the US), policy data is available at a more spatially granular level than infection data (e.g., city policies and state-level infections in the US). In these cases, we code binary policy variables at the more granular level and use population-weights to aggregate them to the level of the infection data. Thus, policy variables may take on continuous values between zero and one, with a value of one indicating that the policy is fully enacted for the entire population. Given the limited quantity of data currently available, we use a parsimonious model that assumes the effects of policies on infection growth rates are approximately linear and additively separable. However, future work that possesses more data may be able to identify important nonlinearities or interactions between policies.

For each country, our general multiple regression model is thus

$$g_{cit} = \log(I_{cit}) - \log(I_{ci,t-1}) = \theta_{0,ci} + \delta_{ct} + \mu_{cit} + \sum_{p=1}^{P_c} (\theta_{cp} \cdot policy_{pcit}) + \epsilon_{cit} \quad (3.7)$$

where observations are indexed by country c , subnational unit i , and day t . The parameters of interest are the country-by-policy specific coefficients θ_{cp} . We display the estimated residuals ϵ_{cit} in Extended Data Fig. 10, which are mean zero but not strictly normal (normality is not a requirement of our modeling and inference strategy), and we estimate uncertainty over all parameters by calculating our standard errors robust to error clustering at the day level. (Angrist and Pischke 2008) This approach allows the covariance in ϵ_{cit} across different locations within a country, observed on the same day, to be nonzero. Such clustering is important in this context because idiosyncratic events within a country, such as a holiday or a backlog in testing laboratories, could generate nonuniform country-wide changes in infection growth for individual days not explicitly captured in our model. Thus, this approach non-

parametrically accounts for both arbitrary forms of spatial auto-correlation or systematic misreporting in regions of a country on any given day (we note that it generates larger estimates for uncertainty than clustering by i). When we report the effect of all policies combined (e.g., Figure 3.2b) we are reporting the sum of coefficient estimates for all policies $\sum_{p=1}^{P_c} \theta_{cp}$, accounting for the covariance of errors in these estimates when computing the uncertainty of this sum.

Note that our estimates of θ and θ_0 in Equation 3.7 are robust to systematic under-reporting of infections, a major concern in the ongoing pandemic, due to the construction of our dependent variable. This remains true even if different localities have different rates of under-reporting, so long as the rate of under-reporting is relatively constant. To see this, note that if each locality i has a medical system that reports only a fraction ψ_i of infections such that we observe $\tilde{I}_{it} = \psi_i I_{it}$ rather than actual infections I_{it} , then the left-hand-side of Equation 3.7 will be

$$\begin{aligned} \log(\tilde{I}_{it}) - \log(\tilde{I}_{i,t-1}) &= \log(\psi_i I_{it}) - \log(\psi_i I_{i,t-1}) \\ &= \log(\psi_i) - \log(\psi_i) + \log(I_{it}) - \log(I_{i,t-1}) \\ &= \log(I_{it}) - \log(I_{i,t-1}) = g_t \end{aligned}$$

and is therefore unaffected by location-specific and time-invariant under-reporting. Thus systematic under-reporting does not affect our estimates for the effects of policy θ . As discussed above, potential biases associated with non-systematic under-reporting resulting from documented changes in testing regimes over space and time are absorbed by region-day specific dummies μ .

However, if the rate of under-reporting within a locality is changing day-to-day, this could bias infection growth rates. We estimate the magnitude of this bias (see Extended Data Fig. 2), and verify that it is quantitatively small. Specifically, if $\tilde{I}_{it} = \psi_{it} I_{it}$ where ψ_{it} changes day-to-day, then

$$\log(\tilde{I}_{it}) - \log(\tilde{I}_{i,t-1}) = \log(\psi_{it}) - \log(\psi_{i,t-1}) + g_t \quad (3.8)$$

where $\log(\psi_{it}) - \log(\psi_{i,t-1})$ is the day-over-day growth rate of the case-detection probability. Disease surveillance has evolved slowly in some locations as governments gradually expand testing, which would cause ψ_{it} to change over time, but these changes in testing capacity do not appear to significantly alter our estimates of infection growth rates. In Extended Data Fig. 2, we show one set of epidemiological estimates (Russell et al. 2020) for $\log(\psi_{it}) - \log(\psi_{i,t-1})$. Despite random day-to-day variations, which do not cause systematic biases in our point estimates, the mean of $\log(\psi_{it}) - \log(\psi_{i,t-1})$ is consistently small across the different countries: 0.05 in China, 0.064 in Iran, 0.019 in South Korea, -0.058 in France, 0.031 in Italy, and 0.049 in the US. The average of these estimates is 0.026, potentially accounting for 7.3% of our global average estimate for the no-policy infection growth rate (0.36). These estimates of $\log(\psi_{it}) - \log(\psi_{i,t-1})$ also do not display strong temporal trends, alleviating concerns that time-varying under-reporting generates sizable biases in our estimated effects of anti-contagion policies.

Transient dynamics In China, we are able to examine the transient response of infection growth rates following policy deployment because only three policies were deployed early in a seven-week sample period during which we observe many cities simultaneously. This provides us with sufficient data to estimate the temporal structure of policy effects without imposing assumptions regarding this structure. To do this, we estimate a distributed-lag model that encodes policy parameters using weekly lags based on the date that each policy is first implemented in locality i . This means the effect of a policy implemented one week ago is allowed to differ arbitrarily from the effect of that same policy in the following week, etc. These effects are then estimated simultaneously and are displayed in Fig. 3.2 (also Supplementary Table 3). Such a distributed lag approach did not provide statistically meaningful insight in other countries using currently available data because there were fewer administrative units and shorter periods of observation (i.e. smaller samples), and more policies (i.e. more parameters to estimate) in all other countries. Future work may be able to successfully explore these dynamics outside of China.

As a robustness check, we examine whether excluding the transient response from the estimated effects of policy substantially alters our results. We do this by estimating a “fixed lag” model, where we assume that policies cannot influence infection growth rates for L days, recoding a policy variable at time t as zero if a policy was implemented fewer than L days before t . We re-estimate Equation 3.7 for each value of L and present results in Extended Data Fig. 5 and Supplementary Table 5.

Alternative disease models Our main empirical specification is motivated with an SIR model of disease contagion, which assumes zero latent period between exposure to COVID-19 and infectiousness. If we relax this assumption to allow for a latent period of infection, as in a Susceptible-Exposed-Infected-Recovered (SEIR) model, the growth of the outbreak is only asymptotically exponential. (Ma 2020) Nonetheless, we demonstrate that SEIR dynamics have only a minor potential impact on the coefficients recovered by using our empirical approach in this context. In Extended Data Figs. 8 and 9 we present results from a simulation exercise which uses Equations 3.9-3.11, along with a generalization to the SEIR model (Ma 2020) to generate synthetic outbreaks (see Supplementary Methods Section 2). We use these simulated data to test the ability of our statistical model (Equation 3.7) to recover both the unimpeded growth rate (Extended Data Fig. 8) as well as the impact of simulated policies on growth rates (Extended Data Fig. 9) when applied to data generated by SIR or SEIR dynamics over a wide range of epidemiological conditions.

3.5.3 Projections

Daily growth rates of infections To estimate the instantaneous daily growth rate of infections if policies were removed, we obtain fitted values from Equation 3.7 and compute a predicted value for the dependent variable when all P_c policy variables are set to zero. Thus, these estimated growth rates $\hat{g}_{cit}^{no\ policy}$ capture the effect of all locality-specific factors on the growth rate of infections (e.g., demographics), day-of-week-effects, and adjustments

based on the way in which infection cases are reported. This counterfactual does not account for changes in information that are triggered by policy deployment, since those should be considered a pathway through which policies affect outcomes, as discussed in the main text. Additionally, the “no-policy” counterfactual does not model previously unobserved changes in behavior that might occur if fundamentally new behaviors emerge even in the absence of government intervention. When we report an average no-policy growth rate of infections (Figure 3.2a), it is the average value of these predictions for all observations in the original sample. Location-and-day specific counterfactual predictions ($\hat{g}_{cit}^{no\ policy}$), accounting for the covariance of errors in estimated parameters, are shown as red markers in Figure 3.3.

Cumulative infections To provide a sense of scale for the estimated cumulative benefits of effects shown in Figure 3.3, we link our reduced-form empirical estimates to the key structures in a simple SIR system and simulate this dynamical system over the course of our sample. The system is defined as the following:

$$\frac{dS_t}{dt} = -\beta_t S_t I_t \tag{3.9}$$

$$\frac{dI_t}{dt} = (\beta_t S_t - \gamma) I_t \tag{3.10}$$

$$\frac{dR_t}{dt} = \gamma I_t \tag{3.11}$$

where S_t is the susceptible population and R_t is the removed population. Here β_t is a time-evolving parameter, determined via our empirical estimates as described below. Accounting for changes in S becomes increasingly important as the size of cumulative infections ($I_t + R_t$) becomes a substantial fraction of the local subnational population, which occurs in some no-policy scenarios. Our reduced-form analysis provides estimates for the growth rate of active infections ($\hat{g}$) for each locality and day, in a regime where $S_t \approx 1$. Thus we know

$$\left. \frac{dI_t}{dt} / I_t \right|_{S \approx 1} = \hat{g}_t = \beta_t - \gamma \tag{3.12}$$

but we do not know the values of either of the two right-hand-side terms, which are required to simulate Equations 3.9-3.11. To estimate γ , we note that the left-hand-side term of Equation 3.11 is

$$\frac{dR_t}{dt} \approx \frac{d}{dt}(\text{cumulative_recoveries} + \text{cumulative_deaths})$$

which we can observe in our data for China and South Korea. Computing first differences in these two variables (to differentiate with respect to time), summing them, and then dividing by active cases gives us estimates of γ (medians: China=0.11, Korea=0.05). These values differ slightly from the classical SIR interpretation of γ because in the public data we are able to obtain, individuals are coded as “recovered” when they no longer test positive for

COVID-19, whereas in the classical SIR model this occurs when they are no longer infectious. We adopt the average of these two medians, setting $\gamma = .08$. We use medians rather than simple averages because low values for I induce a long right-tail in daily estimates of γ and medians are less vulnerable to this distortion. We then use our empirically-based reduced-form estimates of $\hat{g}$ (both with and without policy) combined with Equations 3.9-3.11 to project total cumulative cases in all countries, shown in Figure 3.4. We simulate infections and cases for each administrative unit in our sample beginning on the first day for which we observe 10 or more cases (for that unit) using a time-step of 4 hours. Because we observe confirmed cases rather than total infections, we seed each simulation by adjusting observed I_t on the first day using country-specific estimates of case detection rates. We adjust existing estimates of case under-reporting (Russell et al. 2020) to further account for asymptomatic infections assuming an infection-fatality ratio of 0.75% (Meyerowitz-Katz and Merone 2020). We assume $R_t = 0$ on the first day. To maintain consistency with the reported data, we report our output in confirmed cases by multiplying our simulated $I_t + R_t$ values by the aforementioned proportion of infections confirmed. We estimate uncertainty by resampling from the estimated variance-covariance matrix of all regression parameters. In Extended Data Fig. 7, we show sensitivity of this simulation to the estimated value of γ as well as to the use of a Susceptible-Exposed-Infected-Recovered (SEIR) framework. In Supplementary Table 6, we show sensitivity of this simulation to the assumed infection-fatality ratio (see Supplementary Methods Section 1).

3.6 Data and Code Availability

The datasets generated during and/or analysed during the current study are available at <https://github.com/bolliger32/gpl-covid> (DOI: 10.5281/zenodo.3832367). For easier replication, we have also created a CodeOcean “capsule” – which contains a pre-built computing environment in addition to the source code and data. This is available at <https://codeocean.com/capsule/1887579/tree/v1> (DOI: 10.24433/CO.6625287.v2). Future updates and/or extensions to data or code will be listed at <http://www.globalpolicy.science/covid19>.

3.7 End Notes

3.7.1 Acknowledgements

We thank Brenda Chen for her role initiating this work and Avi Feller for his feedback. Funding: SAP, EK, PL, JT are supported by a gift from the Tuaropaki Trust. TC is supported by an AI for Earth grant from National Geographic and Microsoft. DA, AH, IB are supported through joint collaborations with the Climate Impact Lab. KB is supported by the Royal Society Te Apārangi Rutherford Postdoctoral Fellowship. HD and ER are supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE

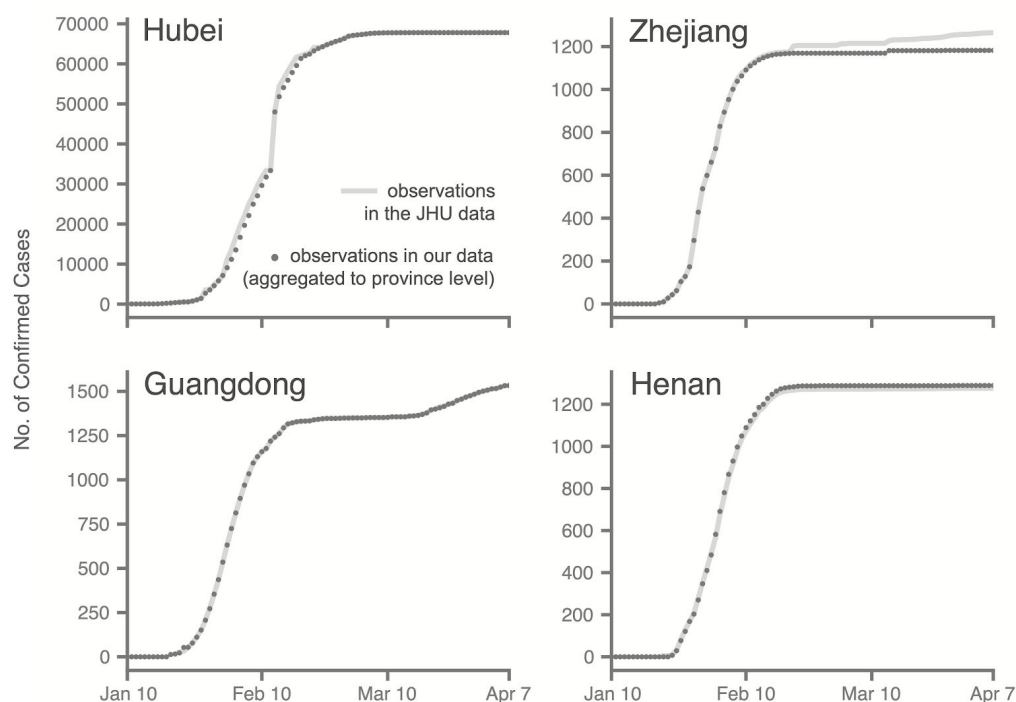
1106400 and 1752814, respectively. Opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of supporting organizations.

3.7.2 Ethics Declaration

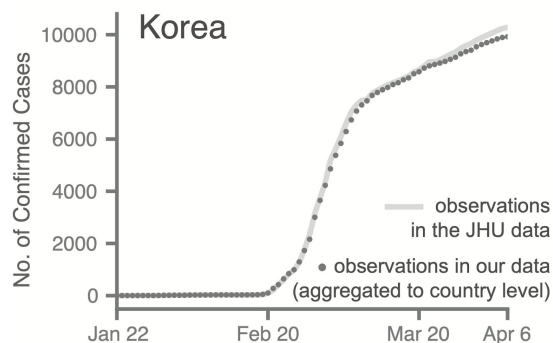
The authors declare no conflicts of interest.

3.8 Extended Data

a



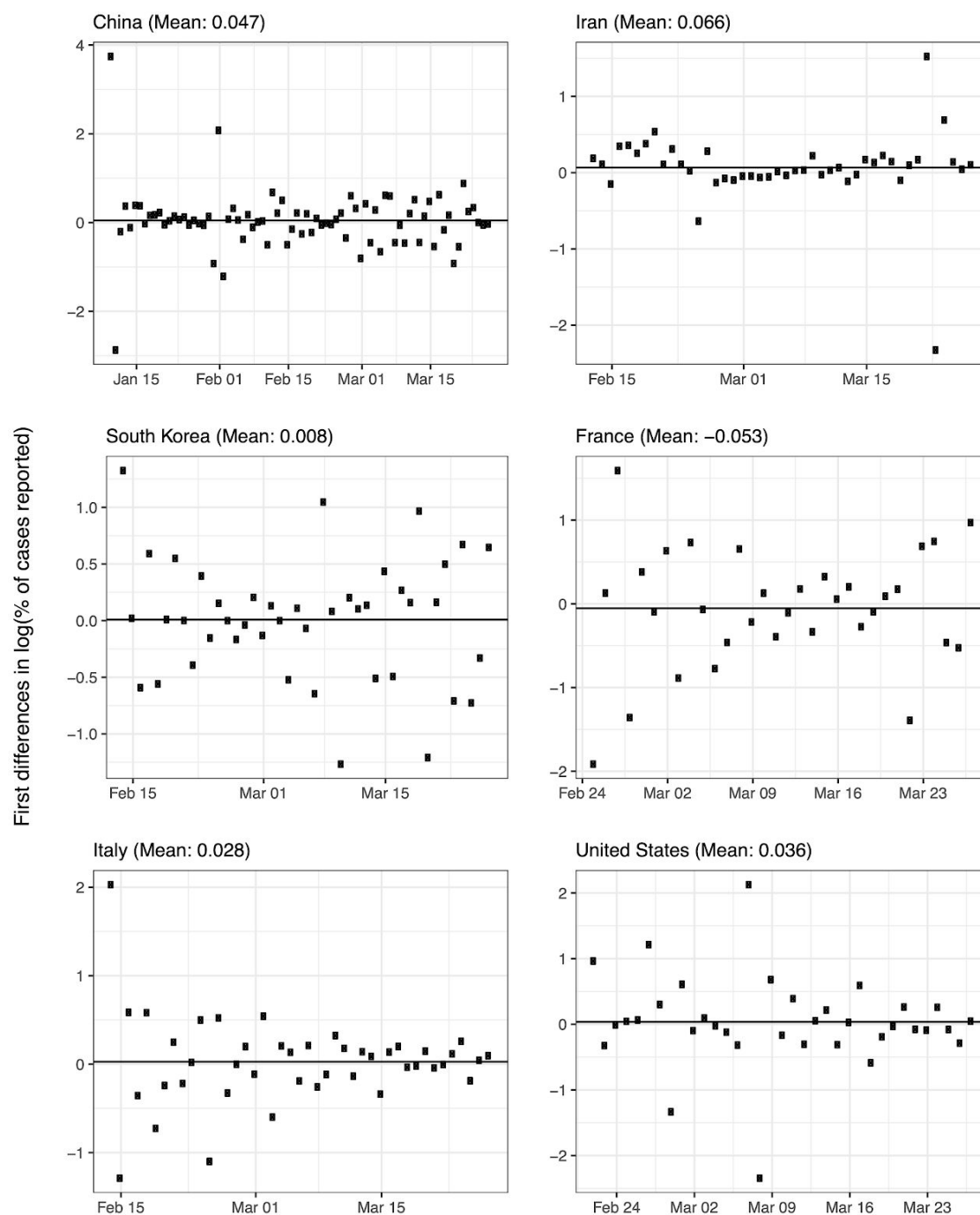
b



Extended Data Figure 1 | Validating disaggregated epidemiological data against aggregated data from the Johns Hopkins Center for Systems Science and Engineering. Comparison of cumulative confirmed cases from a subset of regions in our collated epidemiological dataset to the same statistics from the 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository by the Johns Hopkins Center for Systems Science and Engineering (JHU CSSE).¹ We conduct this comparison for Chinese provinces and South Korea, where the data we collect are from

local administrative units that are more spatially granular than the data in the JHU CSSE database. **a**, In China, we aggregate our city-level data to the province level, and **b**, in Korea we aggregate province-level data up to the country level. Small discrepancies, especially in later periods of the outbreak, are generally due to imported cases (international or domestic) that are present in national statistics but which we do not assign to particular cities (in China) or provinces (in Korea).

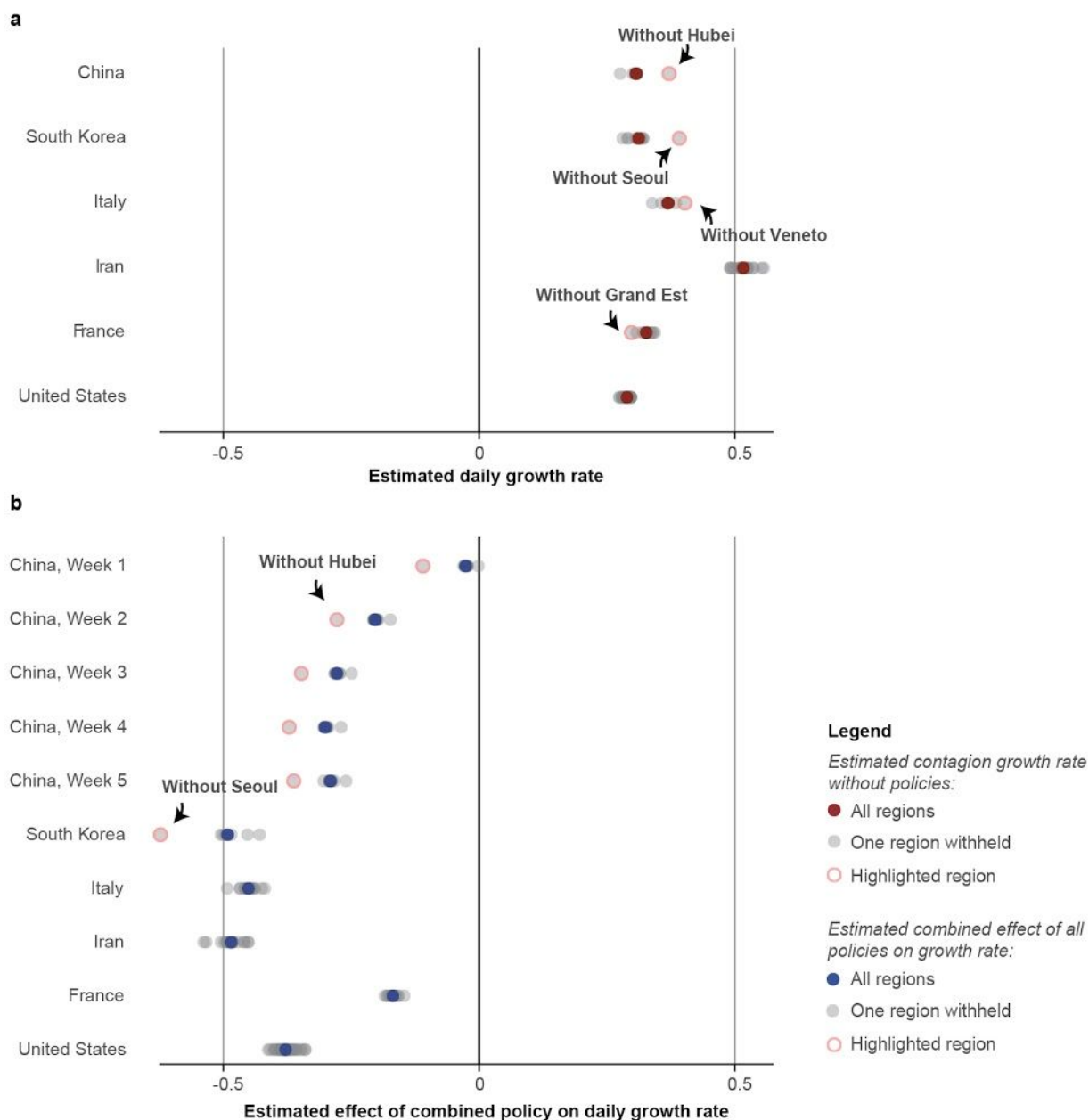
¹ <https://github.com/CSSEGISandData/COVID-19> (access date: April 7, 2020)



Extended Data Figure 2 | Estimated trends in case detection over time within each country. Systematic trends in case detection may potentially bias estimates of no-policy infection growth rates (see Equation 8). We estimate the potential magnitude of this bias using data from the Centre for Mathematical Modelling of Infectious Diseases.² Markers indicate daily first-differences in the logarithm of the fraction of estimated symptomatic

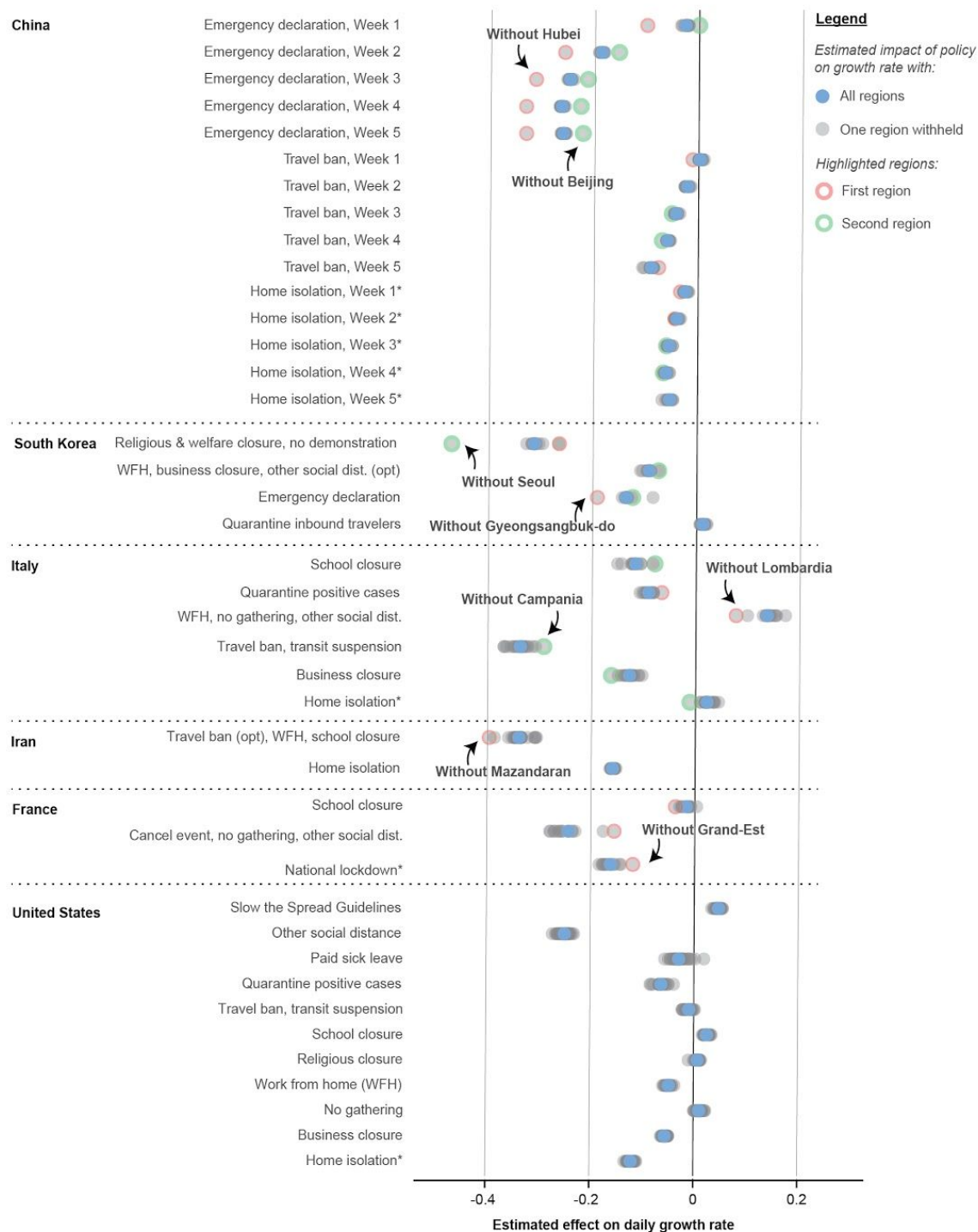
cases reported for each country over time. The average value over time (solid line and value denoted in panel title) is the average growth rate of case detection, equal to the magnitude of the potential bias. For example, in the main text we estimate that the infection growth rate in the United States is 0.29 (Figure 2A), of which growth in case detection might contribute 0.049 (this figure). Sample sizes are 75 in China, 41 in Iran, 40 in South Korea, 29 in France, 40 in Italy, and 32 in the US.

² Russell, T., Joel Hellewell, and S. Abbot. "Using a delay-adjusted case fatality ratio to estimate under-reporting." *Centre for Mathematical Modelling of Infectious Diseases Repository* (2020). URL: https://cmmid.github.io/topics/covid19/severity/global_cfr_estimates.html (access date: April 18, 2020)



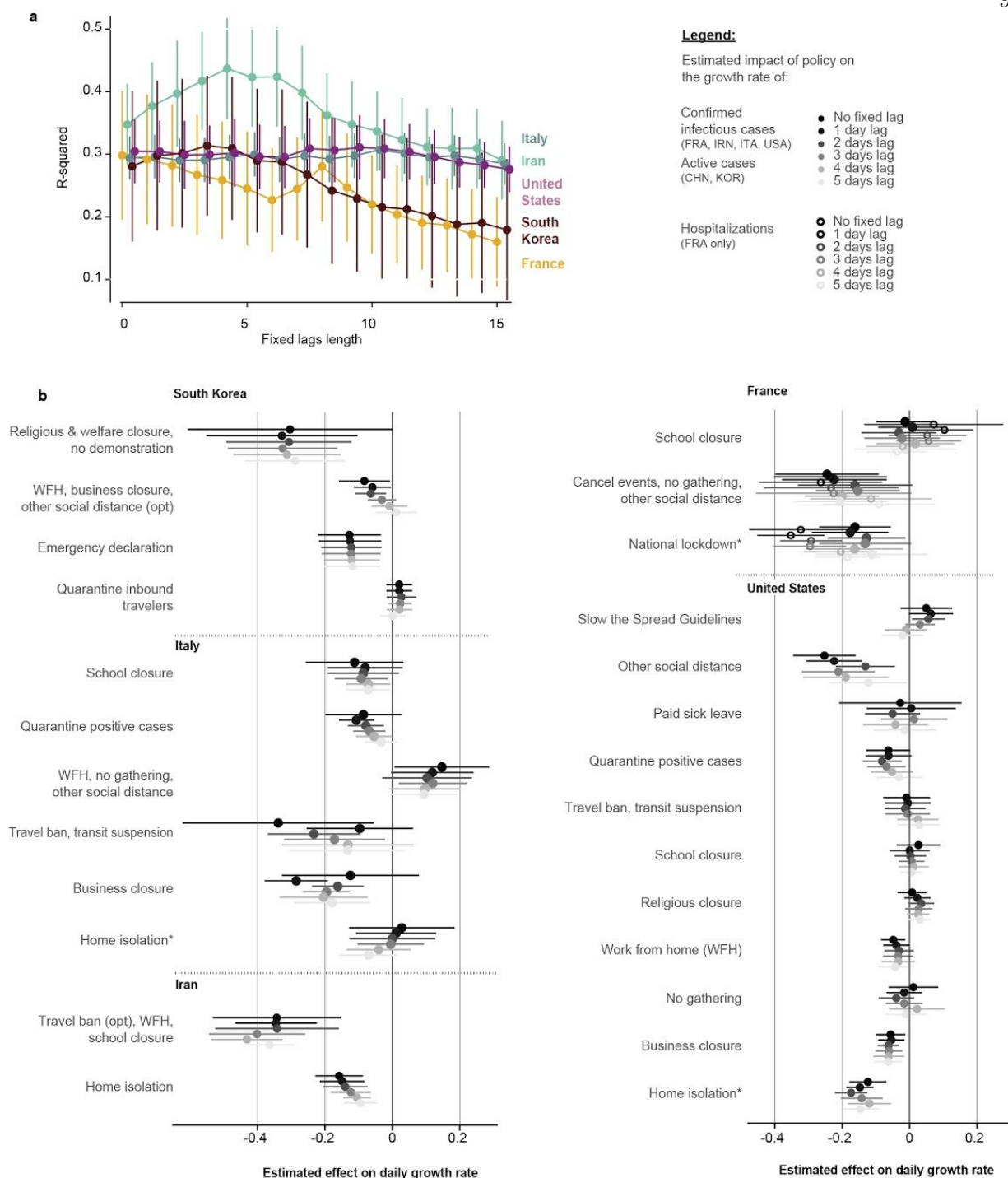
Extended Data Figure 3 | Robustness of the estimated no-policy growth rate of infections and the combined effect of policies to withholding blocks of data from entire regions. For each country, we re-estimated Eq. 7 using real data k times, each time withholding one of the k first-level administrative regions (“Adm1,” i.e. state or province) in that country. Each gray circle is either (a) the estimated no-policy growth rate or (b) the total effect of all policies combined, from one of these k regressions. Red and blue circles show estimates from the full sample, identical to results presented in panels A and B of Figure 2, respectively.

For each country panel, if a single region is influential, the estimated value when it is withheld from the sample will appear as an outlier. Some regions that appear influential are highlighted with an open pink circle. As in Figure 2B of the main text, we estimate a distributed lag model for China and display each of the estimated weekly lag effects (red circle is the same “without Hubei” sample for lags). The full sample includes 3,684 observations in China, 595 in South Korea, 2,898 in Italy, 548 in Iran, 270 in France, and 1,238 in the US.



Extended Data Figure 4 | Robustness of the estimated effects of individual policies to withholding blocks of data from entire regions. Same as Extended Data Figure 3, but for individual policies (analogous to Figure 2C in the main

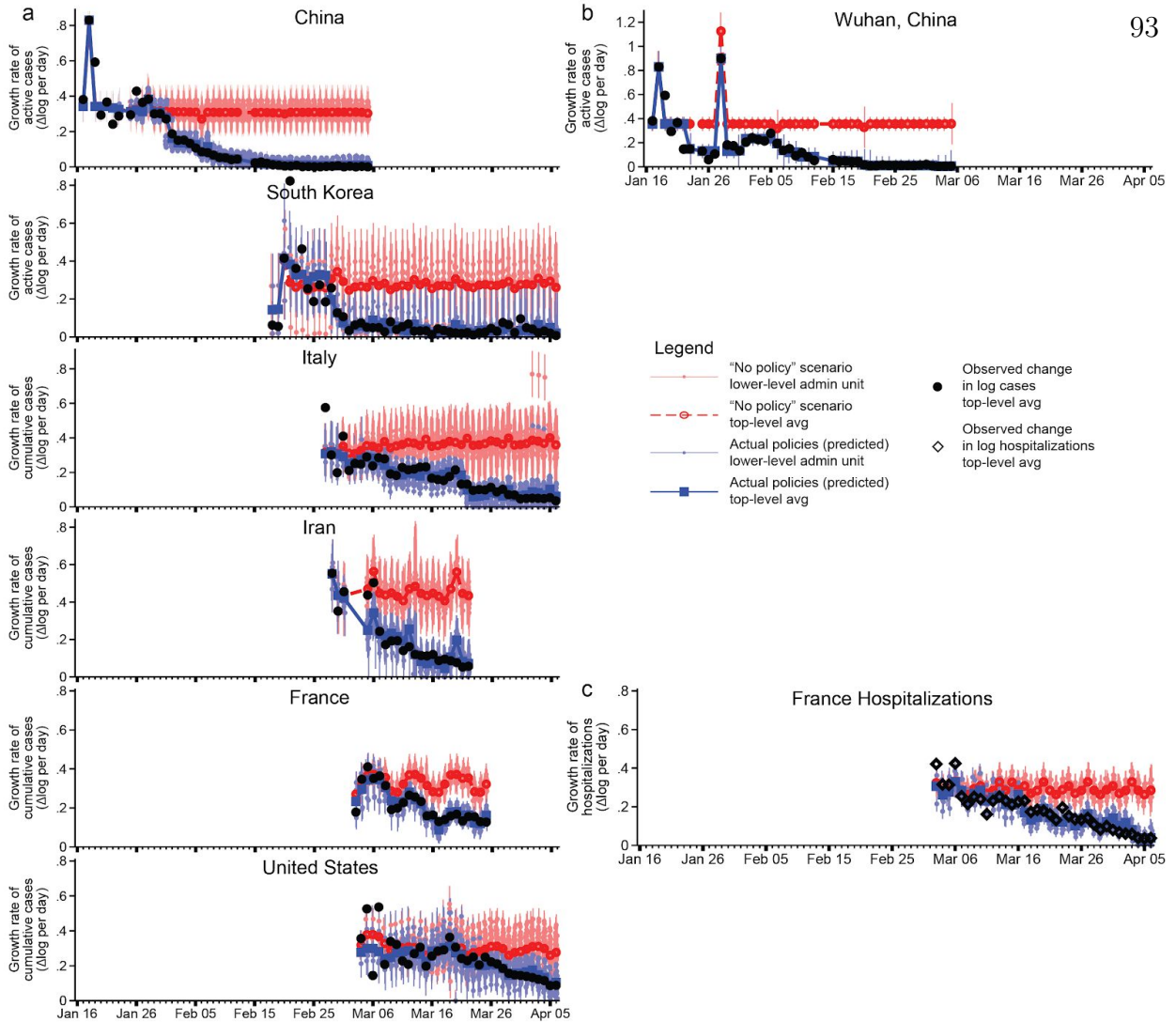
text). In cases where two regions are influential, a second region is highlighted with an open green circle. The full sample includes 3,669 observations in China, 595 in South Korea, 2,898 in Italy, 548 in Iran, 270 in France, and 1,238 in the US.



Extended Data Figure 5 | Evidence supporting models where policies affect infection growth rates in the days following deployment.

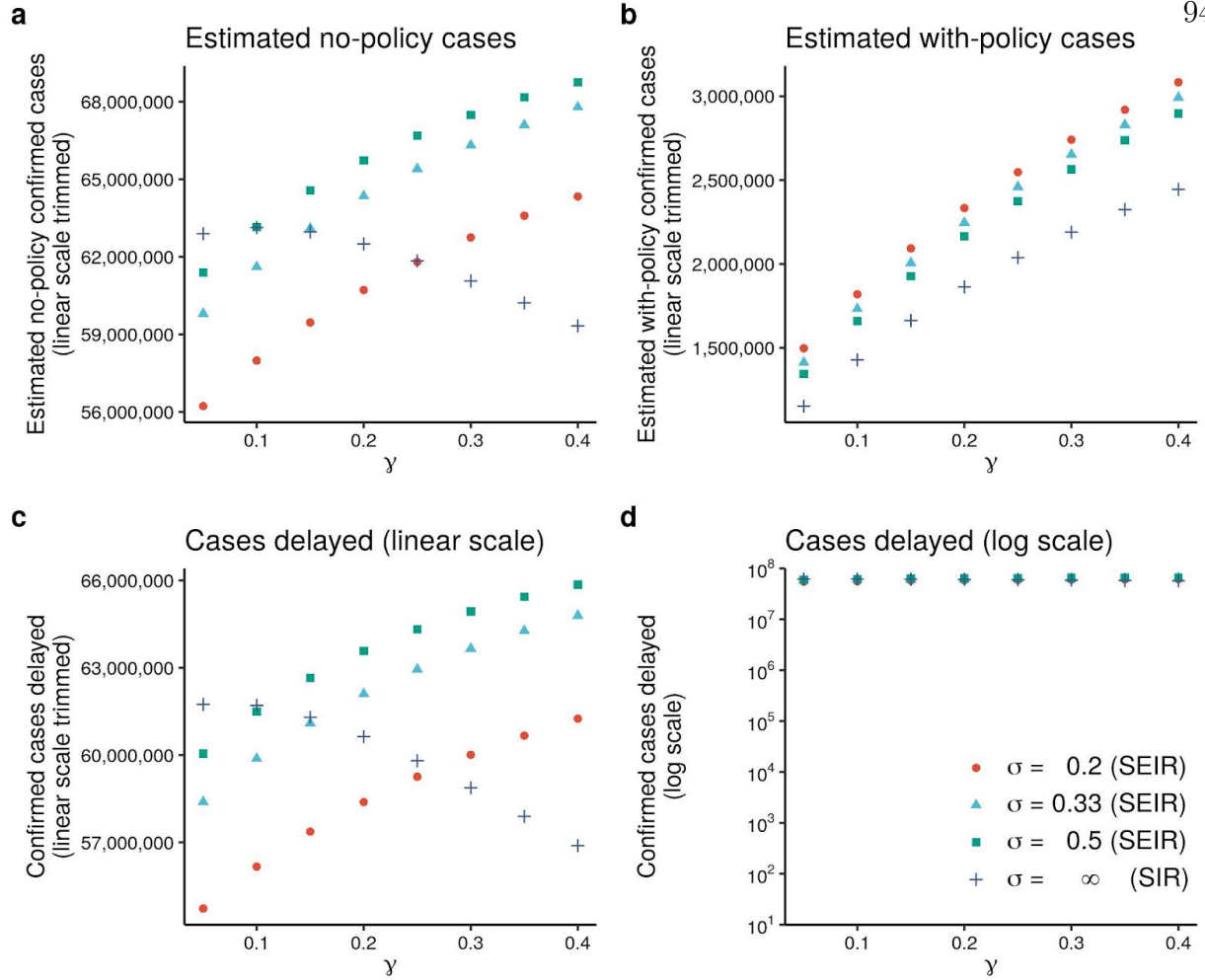
Existing evidence has not demonstrated whether policies should affect infection growth rates in the days immediately following deployment. It is therefore not clear *ex ante* whether the policy variables in Eq. 7 should be encoded as “on” immediately following a policy deployment. We estimate “fixed-lag” models in which a fixed delay between a policy’s deployment and its effect is assumed (see Supplementary Methods Section 3). If a delay model is more consistent with real world infection dynamics, these fixed lag models should recover larger estimates for the impact of policies and exhibit better model fit. **a**, R-squared values

associated with fixed-lag lengths varying from zero to fifteen days. Center values represent the R squared value in our sample, whiskers are 95% CI computed through resampling with replacement. In-sample fit generally declines or remains unchanged if policies are assumed to have a delay longer than four days. **b**, Estimated effects for no lag (the model reported in the main text) and for fixed-lags between one and five days. Center values represent the point estimate, error bars are 95% CI. Estimates generally are unchanged or shrink towards zero (e.g. *Home isolation* in Iran), consistent with mis-coding of post-policy days as no-policy days. The sample size is 595 in South Korea, 2,898 in Italy, 548 in Iran, 270 in France, and 1,238 in the US.



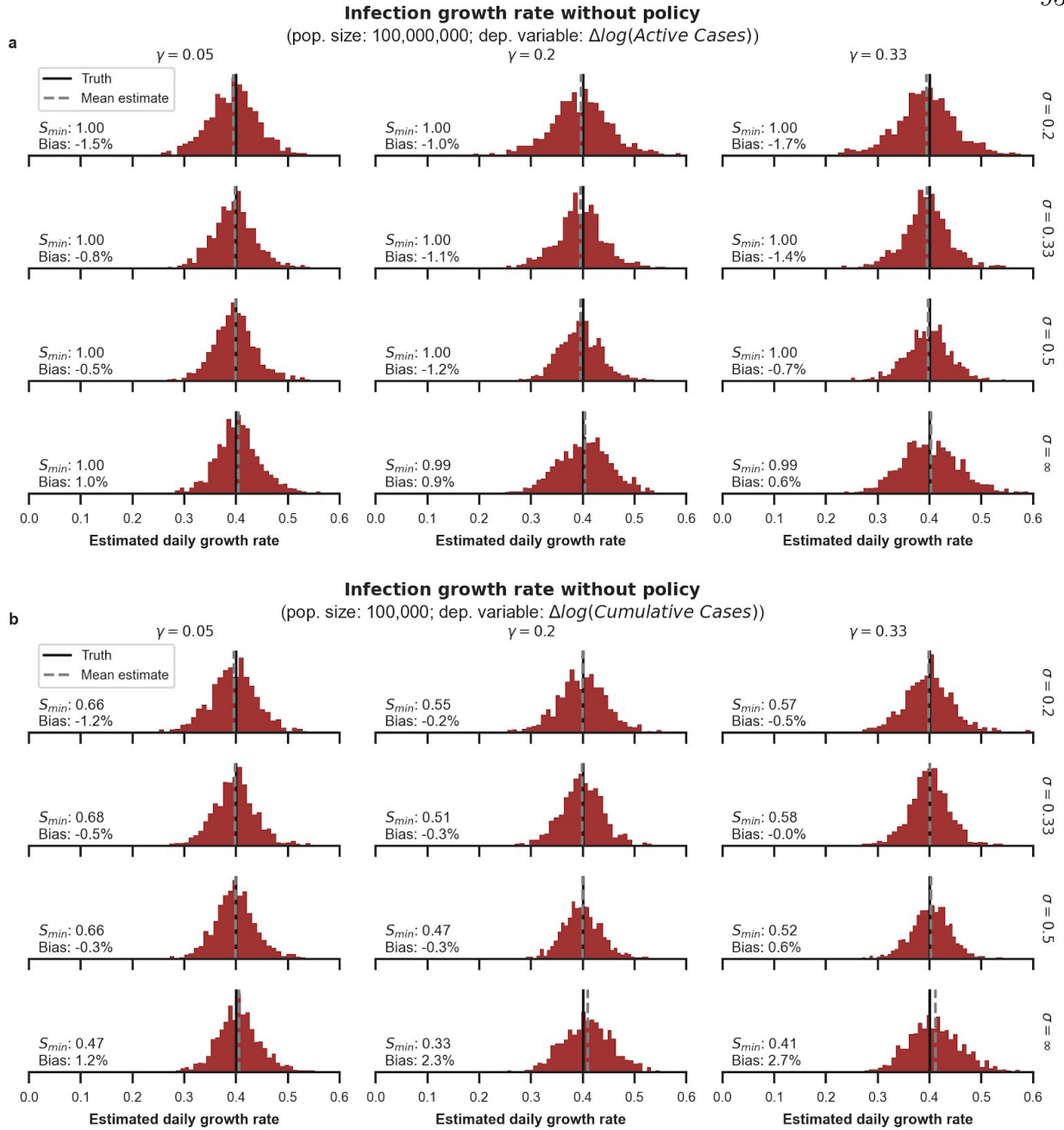
Extended Data Figure 6 | Estimated infection or hospitalization growth rates with actual anti-contagion policies and in a "no policy" counterfactual scenario. **a**, The estimated daily growth rates of active (China, South Korea) or cumulative (all others) infections based on the observed timing of all policy deployments within each subnational unit (blue) and in a scenario where no policies were deployed (red). Identical to Figure 3 in the main text, but using an alternative disaggregated encoding of policies that does not group any policies into policy packages. The sample size is 3,669 in China, 595 in South Korea, 2,898 in Italy, 548 in Iran, 270 in France, and 1,238 in the US. **b**, Same as Figure 3 in the main text, but Eq. 7 is implemented for a single example

administrative unit, Wuhan, China. The sample size is 46 observations. **c**, Same as Figure 3 in the main text, but using hospitalization data from France rather than cumulative cases (the French government stopped reporting cumulative cases after March 25, 2020). The sample size is 424 observations. For all panels, the difference between the with- and no-policy predictions is our estimated effect of actual anti-contagion policies on the growth rate of infections (or hospitalizations). The markers are daily estimates for each subnational administrative unit (vertical lines are 95% confidence intervals). Black circles are observed changes in $\log(\text{infections})$ (or diamonds for $\log(\text{hospitalizations})$), averaged across observed administrative units.



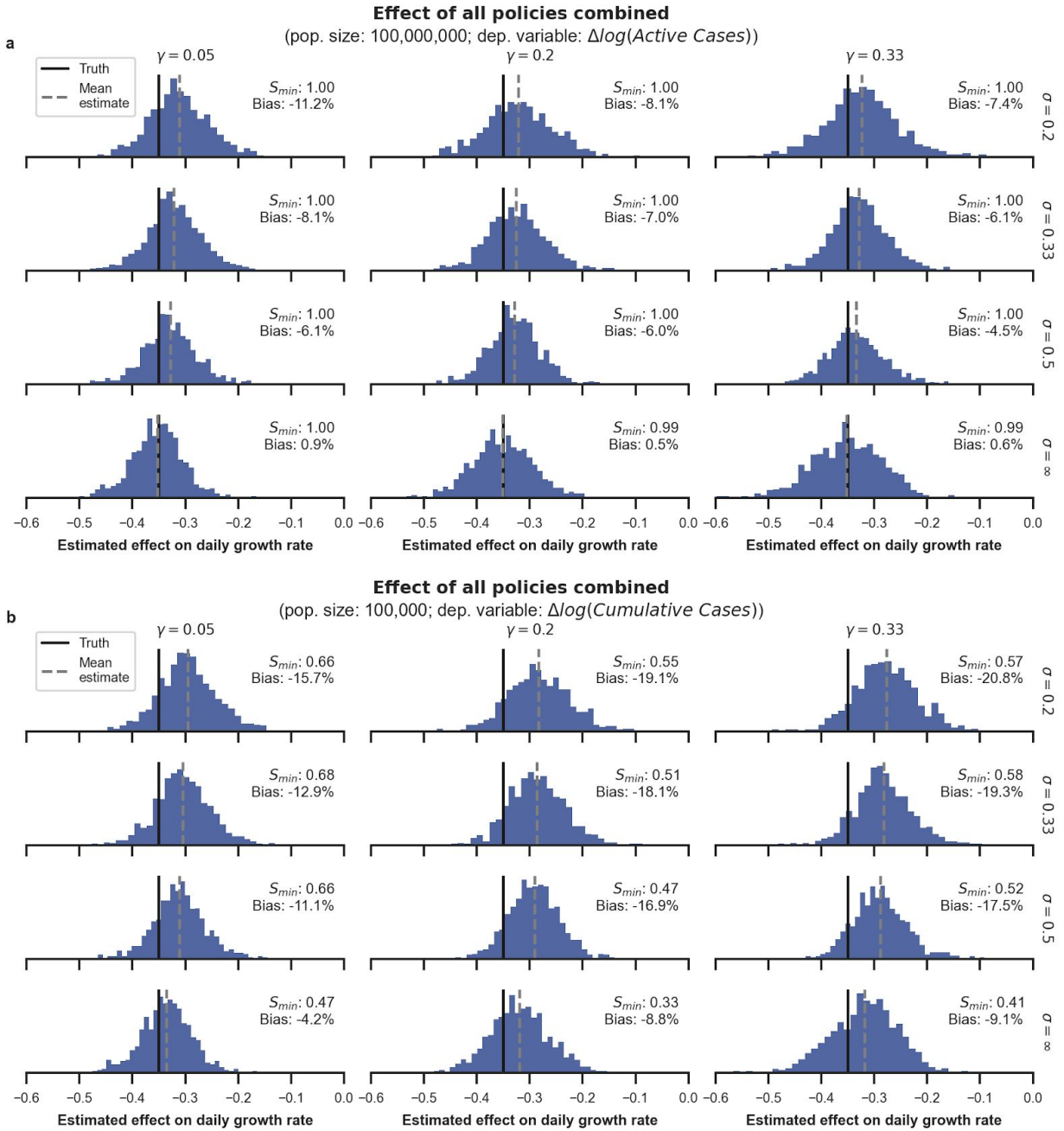
Extended Data Figure 7 | Sensitivity of estimated averted/delayed infections to the choice of γ and σ in an SIR/SEIR framework. This figure displays the sensitivity of total averted/delayed cases presented in Figure 4 of the main text to alternative modeling assumptions. We compute total cases across the respective final days in our samples for the six countries presented in our analysis. The figure displays how these totals vary with eight values of γ (0.05-0.4) and four values of σ (0.2, 0.33, 0.5, ∞), where the final value of σ (∞) corresponds to the SIR model. **a**, The simulated total number of infections under no policy. **b**, Same, but using

actual policies. **c**, The difference between **(a)** and **(b)**, which is the total number of averted/delayed infections. **d**, Same as **(c)**, but on a logarithmic scale similar to Figure 4 in the main text (**a-c** are on a linear scale, trimmed to show details). Figure 4 in the main text uses $\gamma = 0.079$, which we calculate using empirical recovery/death rates in countries where we observe them (China and South Korea, see Methods). If we assume a 14-day delay between infected individuals becoming non-infectious and being reported as “recovered” in the data, we would calculate $\gamma = 0.18$. Figure 4 in the main text assumes $\sigma = \infty$.



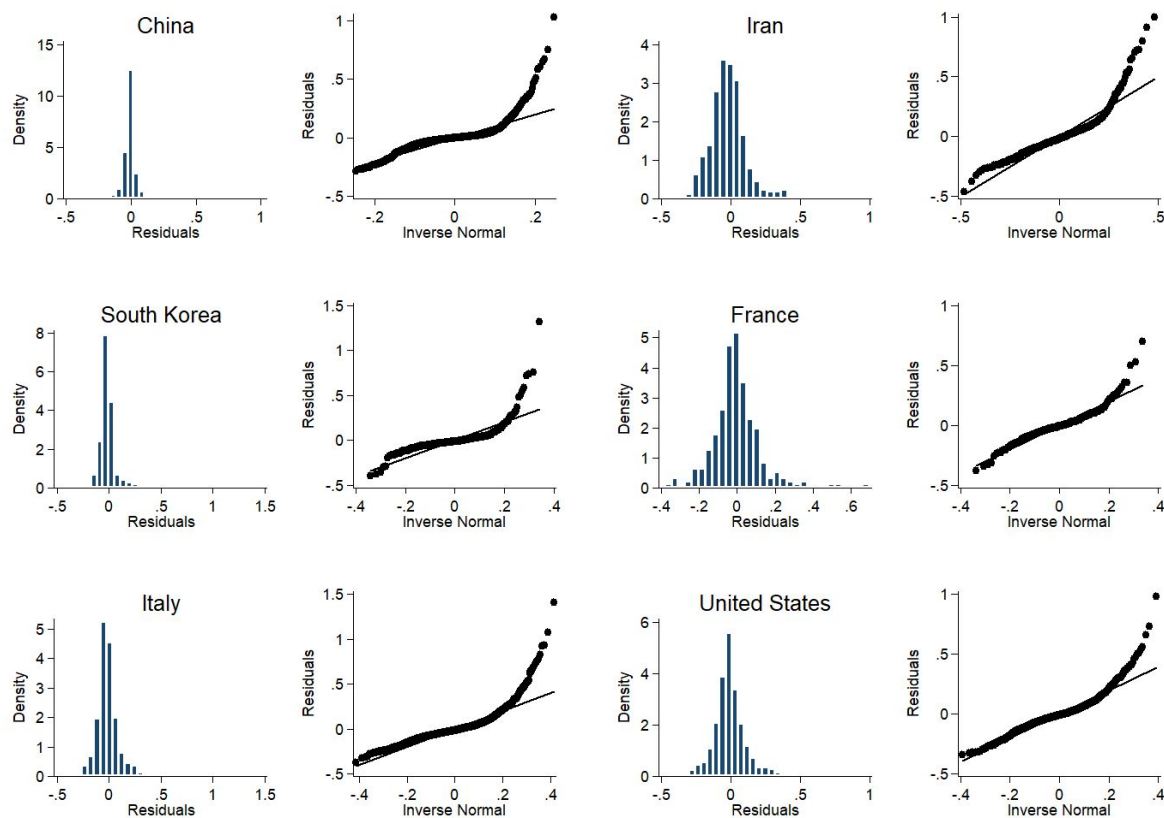
Extended Data Figure 8 | Simulating reduced-form estimates for the no-policy growth rate of infections for different population regimes and disease dynamics. We examine the performance of reduced-form econometric estimators through simulations in which different underlying disease dynamics are assumed (see Supplementary Information Section 3). Each histogram shows the distribution of econometrically estimated values across 1,000 simulated outbreaks. Estimates are for the no-policy infection growth rate (analogous to Figure 2a) when three different policies are deployed at random moments in time. The black line shows the correct value imposed on the simulation and the red histogram shows the distribution of estimates using the regression in Eq. 7, applied to data output from the simulation. The grey dashed line shows the mean of this distribution. The 12 subpanels describe the results when various values are assigned to the mean infectious period (γ^{-1}) and mean latency period

(σ^{-1}) of the disease. “ $\sigma = \infty$ ” is equivalent to SIR disease dynamics. In each panel, S_{min} is the minimum susceptible fraction observed across all 1,000 45-day simulations shown in each panel. In the real datasets used in the main text, after correcting for country-specific under-reporting, S_{min} across all units analyzed is 0.72 and 95% of the analyzed units finish with $S_{min} > 0.91$. “Bias” refers to the distance between the dashed grey and black line as a percentage of the true value. **a**, Simulations in near-ideal data conditions in which we observe active infections within a large population (such that the susceptible fraction of the population remains high during the sample period, similar to those in our data for Chongqing, China). **b**, Simulations in a non-ideal data scenario where we are only able to observe cumulative infections in a small population (similar to those in our sample for Cremona, Italy).



Extended Data Figure 9 | Simulating reduced form estimates for anti-contagion policy effects for different population regimes and assumed disease dynamics. Same as Extended Data Figure 8, but

estimates are for the combined effect of three different policies (analogous to Figure 2b) that are deployed at random moments in time.



Extended Data Figure 10 | Regression residuals for the growth rates of COVID-19 by country. These plots show the estimated residuals from Equation 7 for each country-specific econometric model. Histograms (left) show the estimated unconditional probability density function. Quantile

plots (right) show quantiles of the cumulative density function (y-axis) plotted against the same quantiles for a Normal Distribution. For additional details, see Fig. 3 and the Econometric Analysis section of Methods.

3.9 Supplementary Information

The Supplementary Information contains three sections: Supplementary Notes, Supplementary Methods, and Supplementary Tables.

The Supplementary Notes section details the basic accounting of policy deployment decisions in each of the six countries we analyze, and describes the data acquisition and processing procedure for the epidemiological and policy data used in this paper. The sources for both types of data come from a variety of in-country data sources, which include government public health websites, regional newspaper articles, and Wikipedia crowd-sourced information. We have supplemented this data with international data compilations. A list of the epidemiological and policy data compiled for this analysis can be found [here](#).

The Supplementary Methods section describes sensitivity analyses and simulations performed to verify the robustness of our model, including: the sensitivity of our counterfactual projections to varying epidemiological parameters; and the sensitivity of our regression estimates to alternative lag structures, withholding of data, and differing policy groupings.

The Supplementary Tables section contains tables detailing:

1. Number of unique anti-contagion policies tabulated by administrative division of each country.
2. Epidemiological data in Wuhan prior to policy intervention, and estimates of the initial infection growth rate and case doubling times.
3. Full regression results from our main regression estimating the effect of policy on growth rates.
4. Estimates of the effect of policies on infection growth rate using a more disaggregated grouping of policies.
5. Estimates of the effect of policies on infection growth rate assuming a variety of lag structures.
6. Estimates of the cumulative number of confirmed cases averted assuming a variety of country-specific underreporting ratios.

Supplementary Notes

Chronology of Policy Implementation

In the main article, we analyze how infection growth rates change when anti-contagion policies are deployed. In order for these estimated associations to be interpreted as plausibly causal, it must be the case that changes in policy are not correlated with changes in infection growth rates that would have occurred in the absence of policy actions.¹ For locations and periods where policy was actually deployed, it is not possible to directly observe how growth rates would have changed in the absence of policy actions.² In general, it is well-established that in the absence of policy, early rates of infection within a susceptible population follow almost-perfect exponential growth.^{3, 4, 5, 6, 7, 8, 9, 10} This fact would suggest that in the absence of policy, infection growth rates would be constant^{11, 12} and, therefore, changes in the infection growth rate could not be correlated with the timing of policy deployments (since correlation with a constant variable is zero). However, it is nonetheless also worthwhile to consider how the timing of policy actions were determined in practice. In this section, we provide a basic accounting of policy deployment decisions in each of the six countries we analyze, focusing on the timing of major events and reported motivations for notable policy deployments. These accounts are intended to familiarize readers with the general decision-making context of each country, although they are not exhaustive and are not intended to capture the full complexity of decision-making in the ongoing pandemic.

As stated in the main text, policies were generally not initially deployed with any reference to observed or anticipated natural changes in growth rates. In general, policies tended to be deployed in response to high total numbers of cases or to outbreaks in other regions, after long delays due to political constraints, and often with timing that coincided with arbitrary events, like the start and end of the week, or holidays. For example, epidemic planning in France explicitly tied policy actions to case-counts in specific regions, and epidemiological guidance in the US explicitly recommended policy actions depending on when thresholds in total case counts were passed. In South Korea, policy actions across the country were initially triggered by an idiosyncratic outbreak in a Shincheonji Church in Daegu, and in Iran they were triggered by an outbreak in Qom coinciding with a religious pilgrimage. In China and Italy, nation-wide policy actions were deployed relatively swiftly by the central government in response to a regional outbreak, with limited variation across locations that

¹ [Angrist & Pischke, 2008](#)

² [Holland, 1986](#)

³ [Mena-Lorcat, 1992](#)

⁴ [Brauer & Chowell, 2012](#)

⁵ [WHO Ebola Response Team](#)

⁶ [Mills et al, 2004](#)

⁷ [Andreason et al, 2008](#)

⁸ [Nishiura et al, 2010](#)

⁹ [Towers et al, 2014](#)

¹⁰ [WHO Ebola Response Team](#)

¹¹ [Ma et al, 2020](#)

¹² [Muniz-Rodriguez et al, 2020](#)

could be correlated with changes in local growth rates. In China, Italy, and the US, policies tended to be deployed during the weekend or on a Monday or Friday, with the preferred day of the week varying across these countries.

China

In China, most anti-contagion policies were deployed in a centralized manner. On January 27, 2020, towards the beginning of the outbreak in Wuhan, the central government extended the Chinese New Year national holiday that originally ended on January 30 to February 2, 2020, preventing all schools and businesses across the country from resuming operations. The holiday was later further extended within various provinces to end in mid February.

The main policy instruments that cities deployed were: level 1 emergency declaration, travel bans, and home isolation. The timing of these policies varied across cities, but was still strongly influenced by provincial government mandates. Indeed, cities routinely cite documents from upper-level governments when issuing statements or policies.¹³ From January 23 to January 29, 2020, all provincial governments declared Level 1 emergency responses.¹⁴ The declaration often included the closing of entertainment venues, banning of public gatherings, and extensive temperature monitoring at airports, railway stations and highway checkpoints.¹⁵ In addition to these measures, cities started to implement travel bans, limiting travelling in or out of the cities, as well as home isolation policies, which limited residents from leaving their home for non-essential activities. The first travel ban was enacted in Wuhan, on January 23, 2020, with all the cities in Hubei province adopting the same policy shortly afterwards.

Many Chinese cities appeared to enact policies in response to outbreaks in other localities or other political considerations that did not simply reflect the spread of infections across regions. For example, Tian et al. (2020)¹⁶ examined the policy responses in China during the first 50 days of the pandemic and found that about half of the cities that implemented transmission control measures did so before the first local case was reported (see Figure S2 in Tian et al. (2020)). In another example, many cities in Hubei did not implement a home isolation policy until February 10, 2020, lagging behind many other cities in other provinces, even though the count of cases was highest in Hubei.

A disproportionate fraction of policies were deployed on Fridays in China (36.2%).

¹³ Here we list some examples from Xiaogan, Hubei: [孝感市新型冠状病毒感染的肺炎防控指挥部（3号令）](#); [孝感市新型冠状病毒感染的肺炎防控指挥部（4号令）](#); [孝感市新型冠状病毒感染的肺炎防控指挥部（8号令）](#); [孝感市新型冠状病毒感染的肺炎防控指挥部（9号令）](#) (Policy document No. 3, 4, 8, and 9 from the COVID-19 Prevention and Control Center in Xiaogan, Hubei)

¹⁴ Tian, Huaiyu, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz UG Kraemer, Bingying Li et al. "An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China." *Science* (2020).

¹⁵ http://www.gov.cn/xinwen/2020-01/29/content_5472881.htm

¹⁶ Tian, Huaiyu, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz UG Kraemer, Bingying Li et al. "An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China." *Science* (2020).

South Korea

South Korea confirmed its first imported coronavirus case from Wuhan, China on January 20, 2020.¹⁷ By January 29, 2020, there were four confirmed cases, all originating from Wuhan. In response, the Korean Center for Disease Control began systematically screening all travelers arriving from China.¹⁸

On February 18, 2020, South Korea confirmed its 31st case in Daegu Metropolitan City, an individual later identified as a member of the Shincheonji Church.¹⁹ Over the next two days, the number of cases more than doubled, almost half of them linked to the Shincheonji Church.²⁰ While the outbreak was localized in Daegu, both Daegu and Gyeongsangbuk-do (the province that surrounds Daegu) immediately shut down all Shincheonji-affiliated places of worship. The outbreak in Daegu triggered most other provinces to follow suit by closing all religious facilities associated with Shincheonji. This included provinces with significantly lower infection rates at that point in time, such as Jeollabuk-do, Jeollanam-do, Gwangju and even Jeju, an island-province far removed from the epicenter of the outbreak.²¹

The Shincheonji outbreak prompted the central government to activate the highest health alert level on February 23, 2020.²² This was a turning point in the coronavirus response across the country,^{23, 24, 25, 26, 27} with South Korean President Moon Jae-in declaring,

“The [coronavirus] situation [in the country] is entirely different after the Shincheonji outbreak.... The central government, local governments, health officials and medical personnel and the entire people must wage an all-out, concerted response to the [COVID-19] problem.”

By February 26, 2020, there were over 1,100 confirmed cases, approximately half of which were attributable to the Shincheonji Church.²⁸ The identification of the superspreader linked to the Shincheonji Church and the consequent virus outbreak resulted in a large number of governmental countermeasures deployed at the end of February. As the number of confirmed cases surpassed 7,000 in Daegu and some cities in Gyeongsangbuk-do in mid-March, the South Korean government declared

¹⁷ [List | Press Release | News Room : KCDC](#)

¹⁸ [The case definition of 2019 novel coronavirus will be expanded | Press Release | News Room : KCDC](#)

¹⁹ [S. Korea's 31st virus patient still undergoing treatment: KCDC](#)

²⁰ [List | Press Release | News Room : KCDC](#)

²¹ [제주도, 신천지 시설 7곳 폐쇄조치... 신도 현황은 '깜깜' \(Jeju shut down 7 Shincheonji places... information on believers still unknown\)](#)

²² [Coronavirus latest: South Korea declares red alert after rapid spike in cases](#)

²³ “대규모로 일어나고 있는 신천지 집단 감염 사태 이전과 이후는 전혀 다른 상황입니다... 정부와 지방자치단체, 방역당국과 의료진, 나아가 지역주민과 전국민이 혼연일체가 되어 총력 대응해야 하는 중차대한 시점입니다.”

²⁴ [\[전문\]文대통령 “신천지 집단감염 사태 전·후는 전혀 다른 상황” \(President Moon “The situation is entirely different after the Shincheonji outbreak”\)](#)

²⁵ [South Korea Raises Threat Alert Level](#)

²⁶ [\[전문\]文대통령 “신천지 집단감염 사태 전·후는 전혀 다른 상황” \(President Moon “The situation is entirely different after the Shincheonji outbreak”\)](#)

²⁷ [South Korea Raises Threat Alert Level](#)

²⁸ [List | Press Release | News Room : KCDC](#)

a state of emergency for those regions hardest hit by the virus.²⁹ Throughout the country, additional policies were adopted, including strengthened travel restrictions, wider testing regimes, recommendations for social distancing, and guidelines for certain business operations.

On March 19, 2020, as the number of confirmed cases in other countries and the risk of importing the pathogen increased, the central government extended screening measures to all inbound travelers arriving from every country,^{30, 31} requiring them to report their symptoms daily for 14 days via a self-diagnosis mobile application.³² As the situation in other countries worsened, the government intensified its efforts in limiting the spread of the virus from abroad. On March 22, 2020, inbound travelers from Europe were subjected to a home-quarantine for 14 days.³³ This policy was revised on March 27, 2020 to also include all symptomatic travelers arriving from the US.³⁴ Shortly after, on April 1, 2020, this policy was expanded to include inbound travelers arriving from all countries.³⁵

Italy

The first confirmed cases in Italy were imported cases detected on January 31, 2020.³⁶ By mid February, the virus had spread to Lombardy, and later Veneto, in Northern Italy.³⁷ The hardest hit municipalities within these regions (Bertonico, Casalpusterlengo, Castelgerundo, Castiglione d'Adda, Codogno, Fombio, Maleo, San Fiorano, Somaglia, and Terranova dei Passerini in Lombardy, and Vo'Eugane in Veneto)³⁸ were the first areas in Italy to be placed under lockdown on February 22, 2020.

The Italian government made the decision at the national level to close schools across the country on March 5, 2020.⁴⁰ On March 8, 2020, Prime Minister Giuseppe Conte placed Lombardy and 14 other northern provinces under lockdown, and placed restrictions on the rest of the country (such as suspending any public or private gatherings and other social distancing measures).⁴¹ He then placed the entire country under lockdown on March 10, 2020,⁴² with increasing restrictions on commercial

²⁹ [Moon declares virus-hit Daegu, part of North Gyeongsang Province as special disaster zones](#)

³⁰ “정부는 3월 19일(목) 0시부터 특별입국절차 적용대상을 국내의 모든 내·외국인 입국자로 확대하기로 하였다... 이는 최근 입국자 검역 과정에서 발생한 다수의 확진 사례와 전 세계적인 코로나19 전파 속도 등을 고려하여 해외 위험요인이 국내로 재유입되는 것을 강력하게 차단하기 위한 조치이다.” (Translation: “The government decided to expand Special Entry Procedure to all Koreans and foreign nationals entering Korea from March 19... In consideration of the number of cases confirmed from quarantining inbound travelers, and the speed of global transmission, this measure is designed to strongly block the inflow of overseas risk factors into Korea.”)

³¹ [코로나바이러스감염증-19 중앙재난안전대책본부 정례브리핑 \(3월 17일\) \(Coronavirus Infectious Diseases-19 Regular Disaster Prevention Headquarters Regular Briefing \(March 17\)\)](#)

³² [Korea Implements Special Entry Procedure | Official Korea Tourism Organization](#)

³³ [The updates on COVID-19 in Korea as of 23 March | Press Release | News Room : KCDC](#)

³⁴ [The updates on COVID-19 in Korea as of 27 March | Press Release | News Room : KCDC](#)

³⁵ [The updates on COVID-19 in Korea as of 31 March | Press Release | News Room : KCDC](#)

³⁶ [Coronavirus, primi due casi in Italia: sono due turisti cinesi \(Coronavirus, first two cases in Italy: from two Chinese tourists\)](#)

³⁷ [Factbox: Latest on coronavirus spreading in China and beyond](#)

³⁸ <https://metro.co.uk/2020/02/25/towns-ITA-lockdown-coronavirus-12298246/>

³⁹ [Italian towns on lockdown after 2 virus deaths, clusters](#)

⁴⁰ [Italy orders closure of all schools and universities due to coronavirus](#)

⁴¹ [Italy Locks Down Much of the Country's North Over the Coronavirus](#)

⁴² [Coronavirus: Italy extends emergency measures nationwide](#)

activities passed on March 11, 2020.⁴³ Lastly on March 22, 2020, additional commercial activities and industries were closed, and people's ability to travel across the country were further curtailed.^{44, 45}

Throughout this period, Italian policymakers weighed implementing restrictive anti-contagion policies against maintaining individual freedoms, and were delayed in reaching consensus on when to implement policies due to coordinating across the different branches of government.⁴⁶

A majority (62.7%) of all policies in Italy were deployed on a Sunday.

Iran

The first major outbreak in Iran was connected to a major Shia pilgrimage in the city of Qom that brought travelers from Iran and throughout the Middle East to visit the Fatima Masumeh shrine, which often involves kissing or touching the shrine.⁴⁷ The virus, initially centered in Qom and neighboring Tehran, spread rapidly throughout the country. Early in the pandemic, government communication and policy responses appeared mixed.⁴⁸ It has been reported that, as the virus spread, factions within the government competed for control over the nation's coronavirus response, leading to implementation of policy that was not strongly coordinated.⁴⁹

⁴³ [Coronavirus, negozi e locali chiusi in tutta Italia fino al 25 marzo. Garantiti servizi essenziali, alimentari e farmacie. Possibili riduzioni trasporti. Conte: "Torneremo ad abbracciarci"](#) (Coronavirus, Conte: "In Italy until March 25, shops except food and pharmacies are closed. Possible reductions on transport. Effects for 14 days)

⁴⁴ [Virus News: Italian Industry Shuts Down](#)

⁴⁵ [Coronavirus: Lombardy region announces stricter measures](#)

⁴⁶ [Italy, Pandemic's New Epicenter, Has Lessons for the World](#)

⁴⁷ [How Iran Became a New Epicenter of the Coronavirus Outbreak](#)

⁴⁸ For example, early in the pandemic, officials appeared confident in the country's ability to manage the crisis. But, this quickly gave way to a reportedly uncoordinated policy response: *"Iranian health officials initially boasted of their public health prowess. They ridiculed quarantines as 'archaic' and portrayed Iran as a global role model. President Hassan Rouhani suggested a week ago that by this past Saturday life would have returned to normal..."*

Embarrassed anew by the spread of the disease, the Iranian authorities have responded with a hodgepodge of contradictory measures mixing elements of a crackdown with attempts to save face." [Iran's Coronavirus Response: Pride, Paranoia, Secrecy, Chaos](#) (Accessed 5/5/20)

⁴⁹ For example, the New York Times reported on infighting between the civilian government (led by President Rouhani) and the military after the Supreme Leader Khamenei authorized the military to take over control of the country's coronavirus response: *"While [Khamenei] told the military to work with the civilian government, Mr. Khamenei effectively authorized it to sideline Mr. Rouhani's government if needed. Almost instantly the infighting and conflicting public messaging began..."*

Mr. Rouhani demanded that the armed forces adhere to his command. The generals refused and said Mr. Khamenei had authorized them to act independently.

The military men proposed closing off Tehran and vast swaths of the country... Mr. Rouhani refused...

Apparently ignoring Mr. Rouhani, General Bagheri said on Friday [March 13, 2020] that the military would severely restrict traffic, close businesses and clear people off the streets in Tehran and at least 11 provinces... Barely two days later, Mr. Rouhani said there would be no shutting of cities, no lockdowns, no forced closure of businesses...

By Monday, Mr. Rouhani appeared to have dissuaded the military from imposing strict rules. But local authorities defied him, independently closing provinces and several cities." [Power Struggle Hampers Iran's Coronavirus Response](#) (Accessed 5/5/20)

Throughout this period, the Iranian government increased the stringency of its response as the count of confirmed cases increased. The first two cases and deaths were reported simultaneously on February 19, 2020. The next day, school closures were announced in the province of Qom and travel in the region was discouraged. On February 22, 2020, the government closed schools in 14 provinces and closed down major gathering sites such as football matches and theaters. By March 5, 2020 schools were closed nationwide and government employees were required to work from home. Home isolation was overseen by the military on March 13, 2020, described by The Guardian as a “near-curfew [following] growing exasperation among MPs that calls for Iranian citizens to stay at home had been widely ignored, as people continued to travel before the Nowruz New Year holidays.”⁵⁰ Military implementation of home isolation was reportedly hindered due in part to pushback from the civilian government.⁴⁹

France

The ORSAN (*organisation de la réponse du système de santé en situations sanitaires exceptionnelles*) plan,⁵¹ originally designed to deal with the 2009 H1N1 epidemic, is a predefined, structured emergency plan that guides epidemic responses in France. The ORSAN plan, which is implemented by President Emmanuel Macron and Prime Minister Édouard Philippe based on the advice of the *Conseil scientifique Covid-19*, divides the epidemic outbreak into three distinct phases (“*stades*”) accompanied by different anti-contagion policies:⁵²

Stade 1: The phase when the virus has not prevalently spread within the French population. Policies at this point aim to track infected individuals and identify the source of contagion in order to prevent future clusters. *Stade 1* was declared on February 25, 2020.

Stade 2: The phase when the virus has spread to a limited number of regions within the country. During this phase, policies deployed are designed to slow the propagation of the virus within these regions. For instance, under *stade 2*, localities may close schools or limit access to elderly care facilities on a case-by-case basis. *Stade 2* was declared on February 29, 2020 when the cumulative number of cases within the country reached 100 cases, and the national government prohibited large gatherings of over 5,000 people. On March 9, 2020, this national policy was subsequently changed to limiting gatherings to no more than 1,000 people.

Stade 3: The phase when every region of the country is affected by the virus. Policies are intended to mitigate the impact of the epidemic. *Stade 3* officially started on March 14, 2020 with a speech from President Emmanuel Macron. He announced a nationwide school closure starting the following Monday. The next day the government decided to close all non-essential business and the national lockdown began on March 17, 2020.

⁵⁰ [Revolutionary Guards to enforce coronavirus controls in Iran](#)

⁵¹ [Orsan plan](#)

⁵² [PRÉPARATION AU RISQUE ÉPIDÉMIQUE Covid-19](#)

The central government relied heavily on the *Conseil scientifique Covid-19* for the management of the crisis. The *Conseil scientifique Covid-19*'s analysis was largely based on epidemiological models inspired by the Imperial College model (Ferguson et al, 2006 ; Luca et al, 2018 ; Ferguson et al, 2005).⁵³ According to the *Conseil scientifique Covid-19*'s first report, the goal of anti-contagion measures was to limit the number of individuals who would require hospitalization.⁵⁴

United States

Although the first confirmed COVID-19 case in the US was documented on January 20, 2020,⁵⁵ many jurisdictions across the country did not implement mandatory policies at the county or state level until early- to mid-March, with some states delaying more restrictive policy action until even later. At the national level, both policy action and public engagement surrounding COVID-19 moved slowly. On January 24, 2020, the head of the National Institute of Allergy and Infectious Diseases, Anthony Fauci, commented that:

*"We don't want the American public to be worried about this because their risk is low."*⁵⁶

On January 28, 2020, Alex Azar, the secretary of Health and Human Services gave a Press Briefing in which the tone remained largely the same:

*"As of today, the CDC has reported 5 cases of the novel coronavirus infection here in the United States. China has now reported more than 4,500 cases. Americans should know that this is a potentially very serious public health threat, but, at this point, Americans should not worry for their own safety."*⁵⁷

However, on January 31, 2020, Azar declared a public health emergency for the US.⁵⁸ The first national policy action, implemented on February 2, 2020, suspended entry into the United States of foreign nationals who had visited China in the previous 14 days and mandated quarantine for Americans who had recently visited hard hit regions of China.⁵⁹ The federal government followed this Presidential

⁵³ [Ferguson, N., Cummings, D., Cauchemez, S. et al. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* 437, 209–214 \(2005\).](#)

[Ferguson, N., Cummings, D., Fraser, C. et al. Strategies for mitigating an influenza pandemic. *Nature* 442, 448–452 \(2006\).](#)
[Luca, G.D., Kerckhove, K.V., Coletti, P. et al. The impact of regular school closure on seasonal influenza epidemics: a data-driven spatial transmission model for Belgium. *BMC Infect Dis* 18, 29 \(2018\).](#)

⁵⁴ *"Il est dès lors apparu indispensable au Conseil scientifique de prendre en compte comme objectif collectif principal et immédiat la réduction maximale de l'afflux prévisible de cas graves en réanimation."*
 (Translation: "It therefore appeared essential to the Scientific Council to set the main goal as the maximum reduction of the foreseeable influx of serious cases in intensive care.")

[Avis du Conseil scientifique COVID-19 16 mars 2020 Avis du Conseil scientifique COVID-19 12 mars 2020](#)

⁵⁵ [First Case of 2019 Novel Coronavirus in the United States](#)

⁵⁶ [Three months in: A timeline of how COVID-19 has unfolded in the US](#)

⁵⁷ [Remarks at Coronavirus Press Briefing | Alex M. Azar II | Press | January 28, 2020 Washington, D.C.](#)

⁵⁸ [Secretary Azar Declares Public Health Emergency for United States for 2019 Novel Coronavirus](#)

⁵⁹ In the press briefing detailing these actions, Robert Redfield, director of the CDC, focused on the number of cases internationally, saying: "As of today, there are nearly 9,700 cases in China, with more than 200 deaths. Additionally, currently there are another 23 countries that have confirmed, totally, 132 cases. This also includes 12 individuals who have been

Proclamation with a similar ban on February 29, 2020, prohibiting entry for non-U.S. citizens who had been in Iran.⁶⁰ On March 11, 2020, the federal administration further suspended arrivals by halting travel from the Schengen Area in Europe.⁶¹

On March 16, 2020 the White House announced “15 Days to Slow the Spread,” a set of social distancing guidelines that urged Americans to work from home, engage in remote schooling, avoid unnecessary shopping trips, refrain from visiting restaurants and bars, and not gather in groups of more than 10 people.⁶² It was widely reported that these federal guidelines drew on a report by the Imperial College COVID-19 Response Team (Ferguson et al, 2020) that recommended non-pharmaceutical interventions to slow the spread of the virus.⁶³ This report assumed that infection was “seeded in each country at an exponentially growing rate” (page 4) and provided guidance to policymakers that “suppression policies are best triggered early in the epidemic, with a cumulative total of 200 ICU cases per week being the latest point at which policies can be triggered” and still keep peak ICU demand below surge limits.⁶⁴

Besides the federal travel bans and social distancing guidelines, the majority of US anti-contagion policies were enacted in a decentralized manner by state governors. Washington State, the first state to suffer an outbreak, declared a state of emergency on February 29, 2020 in response to the first in-state COVID-19 death.⁶⁵ A few weeks later, Washington Governor Jay Inslee passed “the immediate two-week closure of all restaurants, bars, and entertainment and recreational facilities, as well as additional limits on large gatherings” starting on March 16, 2020, where his “announcement comes after the recent spike in numbers of COVID-19 cases in the state and across the country.”⁶⁶ A week later, on March 24, 2020, Governor Inslee announced a state-wide shelter-in-place order. In his announcement, he wrote: “We have now confirmed that more than 2,000 Washingtonians have contracted the virus. There are likely thousands more that have not yet been diagnosed. COVID-19 has taken more than 100 lives in our state, a number that will also continue to rise...So, tonight, I am issuing a ‘Stay at Home’ order to fight this virus...This includes a ban on all gatherings, and closures of many businesses.”⁶⁷

In California, another state with early confirmed cases of COVID-19, the Health and Human Services Agency in the California Department of Public Health issued on March 7, 2020 “statewide guidance to help both school and public health officials inform their decision making” on whether to close schools

confirmed in six countries who did not travel to China.” [Press Briefing by Members of the President’s Coronavirus Task Force | January 31, 2020](#)

⁶⁰ [Suspension of Entry as Immigrants and Nonimmigrants of Certain Additional Persons Who Pose a Risk of Transmitting 2019 Novel Coronavirus](#)

⁶¹ [Suspension of Entry as Immigrants and Nonimmigrants of Certain Additional Persons Who Pose a Risk of Transmitting 2019 Novel Coronavirus](#)

⁶² [15 Days to Slow the Spread](#)

⁶³ [White House Takes New Line After Dire Report on Death Toll](#)

⁶⁴ [Impact of non-pharmaceutical interventions \(NPIs\) to reduce COVID-19 mortality and healthcare demand](#)

⁶⁵ [Inslee signs bill package to support state effort combating the COVID-19 outbreak](#)

⁶⁶ [Inslee announces statewide shutdown of restaurants, bars and expanded social gathering limits](#)

⁶⁷ ["Stay Home, Stay Healthy" address transcript](#)

and school districts.⁶⁸ They included different scenarios each with a separate course of action, where scenarios changed based on whether there were two or more confirmed cases in the community; one or more confirmed case(s) by a student or teacher at the school; or if multiple schools within the school district had a student or teacher test positive. The recommendations for some of these scenarios included closing the school or the entire school district. On March 19, 2020, Newsom announced the state-wide shelter in place order. The governor said his decision to deploy the policy came after epidemiological models that projected the spread of COVID-19 in California.⁶⁹

New York State, which would become the center of the US outbreak, confirmed its first case on March 1, 2020. Governor Andrew Cuomo responded by saying that the general risk to New Yorkers “remains low” and that “we are fully coordinated, and we are fully mobilized, and we are fully prepared to deal with the situation as it develops.”⁷⁰ On March 9, 2020, during an interview with Katy Tur of MSNBC, when asked “What about schools? Are you considering closing them? What will it take to close them?” Cuomo responded:

“We close them on a case-by-case basis...You have higher numbers in certain areas [of the state], lower in others. If the numbers are low, God bless. If the numbers are high, take action. As I mentioned before where we have a cluster in Westchester that has more cases than New York City. So how do you handle that hotbed, that hot spot, as they call it? Closing schools, closing gatherings, etcetera, extraordinary efforts where you have higher density of cases.”

Only a few days later, on March 16, 2020, New York and neighboring states Connecticut and New Jersey, announced that they would limit social and recreational gathering to 50 people⁷¹ per national CDC recommendations issued the day before.⁷² The three states also coordinated the implementation of business closures (including restaurants, bars, movie theaters and casinos) effective on March 16, 2020.⁷³ The decision to implement these uniform restrictions all at once across the three neighboring states was to “prevent 'state shopping' where residents of one state travel to another and vice versa.”

⁷⁴

In other states that have subsequently been affected by the virus, policies to close businesses and schools, reduce gatherings and cancel events, and reduce travel have been implemented. In some cases, the timing of these policies was determined by a holiday or in response to a high number of

⁶⁸ [CDE School Guidance for COVID-19](#)

⁶⁹ [Gov. Gavin Newsom orders all of California to shelter in place](#)

⁷⁰ [A timeline of Cuomo's and Trump's responses to coronavirus outbreak | April 3, 2020 | ABC News](#)

⁷¹ [Amid Lack of Federal Direction, Governor Cuomo, Governor Murphy and Governor Lamont Announce Regional Approach to Combatting COVID-19](#)

⁷² [Get Your Mass Gatherings or Large Community Events Ready](#)

⁷³ [Audio & Rush Transcript: Amid Lack of Federal Direction, Governor Cuomo, Governor Murphy and Governor Lamont Announce Regional Approach to Combatting Covid-19](#)

⁷⁴ [Amid Lack of Federal Direction, Governor Cuomo, Governor Murphy and Governor Lamont Announce Regional Approach to Combatting COVID-19](#)

confirmed cases in-state. For example Ohio (on March 15, 2020),⁷⁵ Michigan (on March 16, 2020),⁷⁶ Illinois (on March 16, 2020)⁷⁷ and Massachusetts (on March 17, 2020)⁷⁸ responded preemptively to the St. Patrick's Day holiday (on March 17, 2020) by ordering the closures of bars and restaurants. In Georgia, the closure of public K-12 schools starting on March 18, 2020, the limiting of large gatherings and closing of bars on March 23, 2020, implementing a statewide shelter in place on April 2, 2020, were deployed in response to "the number of COVID-19 cases in Georgia [which has] continue[d] to rise," as stated in the Executive Orders issued by Governor Bill Kemp.⁷⁹ On March 18, 2020, Kentucky closed all businesses that interfaced with the public, such as recreational facilities, hair salons, and concert venues, and recommended against visiting senior, psychiatric, or acute care centers.⁸⁰

Policies across the US were deployed on all days of the week, but a disproportionate number were deployed on Mondays (25.0%) and Fridays (23.8%) relative to other days of the week (10.2% on average).

Epidemiological Data

The epidemiological datasets and sources used in this paper are described below. The main health variables of interest are:

1. *cum_confirmed_cases*: The total number of confirmed positive cases in the administrative area since the first confirmed case.
2. *cum_deaths*: The total number of individuals that have died from COVID-19.
3. *cum_recoveries*: The total number of individuals that have recovered from COVID-19.
4. *cum_hospitalized*: The total number of hospitalized individuals.
5. *cum_hospitalized_symptom*: The total number of symptomatic hospitalized individuals.
6. *cum_intensive_care*: The total number of individuals that have received intensive care.
7. *cum_home_confinement*: The total number of individuals that have been self-quarantined in their homes as a result of a positive test.
8. *active_cases*: The number of individuals who currently still test positive on the date of the observation.
9. *active_cases_new*: The number of new active cases since the previous date.
10. *cum_tests*: The total number of tests (includes both positive and negative results) conducted in an administrative unit.

Additional metadata accompanying the health outcome variables:

1. *date*: The date of observation.
2. *adm0_name*: The ISO3 (country) code to which this observation belongs.

⁷⁵ [All Ohio Bars and Restaurants Ordered Closed at 9 p.m. and Until Further Notice](#)

⁷⁶ [Coronavirus - Governor Whitmer Signs Executive Order Temporarily Closing Bars, Theaters, Casinos, and Other Public Spaces: Limiting Restaurants to Delivery and Carry-Out Orders](#)

⁷⁷ [Pritzker closes bars and restaurants as federal officials to double staff to handle crush at O'Hare](#)

⁷⁸ [Gov. Baker Closes Schools, Bans Mid-Size Gatherings, Eating at Restaurants](#)

⁷⁹ [2020 Executive Orders](#)

⁸⁰ [Kentucky's Response to COVID-19](#)

3. *adm1_name*: The name of the Adm1 region (typically state or province) to which this observation belongs.
4. *adm2_name*: If the dataset contains observations at the Adm2 level, then this is the name of the Adm2 region to which this observation belongs (e.g. counties in the United States).
5. *adm[1,2]_id*: Any alphanumeric ID scheme to identify different administrative units (e.g. FIPS code in the United States).
6. *lat*: The latitude of the centroid of the administrative unit.
7. *lon*: The longitude of the centroid of the administrative unit.
8. *policies_enacted*: The number of active policies that are in place for the administrative unit as of that date. This variable is not population weighted.
9. *testing_regime*: A categorical variable used to identify when an administrative region changed their COVID-19 testing regime. This is zero-indexed, with the ordering only indicating chronological progression (there is no external meaning to Regime 2 vs. Regime 1 vs. Regime 0, and there is no consistency enforced for coding across countries). For example, if China changes their testing regime twice, all observations prior to the first regime change would be coded *testing_regime*=0, all observations in between the two changes would be coded *testing_regime*=1, and all observations after the second change would be coded *testing_regime*=2.
10. *population*: The population of the administrative unit.
11. *pop_is_imputed*: A binary variable equal to 1 if the population is imputed, and 0 otherwise. Used for imputing the population of some cities in China.

Data Imputation

In certain instances where health outcome observations are missing or suffer from data quality issues, we have imputed to fill in the missing values. Imputed health outcome variables are denoted by *[health_outcome]_imputed*. We do not use any imputed health data in our analyses; it is only used for display purposes.

We impute by:

1. Taking the natural log of the non-missing observations pertaining to that health outcome variable.
2. Linearly interpolating over the missing dates for that health outcome variable.
3. Exponentiating the interpolated values back into levels and rounding to the nearest integer.

China

We have collated a city-level time series health outcome dataset in China for 339 cities from January 10, 2020 to April 7, 2020.

For data from January 24, 2020 onwards, we relied on the public dataset Ding Xiang Yuan⁸¹ (DXY) that reports daily statistics across Chinese cities. Since DXY only publishes the most recent

⁸¹ [全球新冠病毒最新实时疫情地图 \(The latest real-time global COVID-19 map\)](#)

(cross-sectional) statistics (and not the historical data), we used the time series dataset scraped from DXY in an open source GitHub project.⁸² The web scraper program checks for updates at least once a day for the statistics published on DXY and records any changes in the number of cumulative confirmed cases, cumulative recoveries or cumulative deaths.

We assumed that no updates to the statistics meant there had been no new cases. We dropped a small number of cases that had been recorded but not assigned to a specific city (many of these cases are imported ones from other cities). We also dropped confirmed cases in prison populations (we assumed the spread of COVID-19 in prisons was not affected by the implementation of city-level lockdowns or travel ban policies).

For city level health outcomes prior to January 24, 2020, we manually collected official daily statistics from the central⁸³ and provincial (Hubei,⁸⁴ Guangdong,⁸⁵ and Zhejiang⁸⁶) Chinese government websites. We did not collect city level health outcomes recorded prior to January 24, 2020 in provinces that had fewer than ten confirmed cases at that date. We made this decision since our analysis dropped observations with fewer than ten cumulative confirmed cases to prevent noisy data during the early transmission phase from disproportionately biasing the estimated results.

After merging the two datasets, we conducted a few quality checks:

(1) We checked that cumulative confirmed cases, cumulative recoveries, and cumulative deaths were increasing over time. In instances when cumulative outcomes decreased over time, we assumed that the recent numbers were more reliable, and treated the earlier number of cumulative cases as missing (this was often due to data entry errors or cases where patients that were reported to have been diagnosed with COVID-19, but were later found out to actually have tested negative). The magnitude of these errors was relatively small. We filled in any missing data with the imputation methodology described in the health data overview section.

(2) We validated our city-level dataset by aggregating observations up to the provincial level and comparing the time trends from the aggregated dataset to that of the provincial dataset collated by Johns Hopkins University.⁸⁷ We confirmed that the two datasets matched very closely (see Figure A2 Panel A).

Testing Regime Changes:

During our sample period starting January 16, 2020, the criteria for being diagnosed with COVID-19 changed five times in China.⁸⁸ On January 18, 2020, China began using the reverse transcription

⁸² [BlankerL/DXY-COVID-19-Data: 2019新型冠状病毒疫情时间序列数据仓库 \(COVID-19/2019-nCoV Infection Time Series Data Warehouse\)](#)

⁸³ [疫情通报 \(National Health Commission of PRC: COVID-19 Report\)](#)

⁸⁴ [信息发布--湖北省卫生健康委员会 \(Information release - Hubei Provincial Health Commission\)](#)

⁸⁵ [广东省卫生健康委员会网站 \(Information Release - Guangdong Provincial Health Commission\)](#)

⁸⁶ [浙江省人民政府门户网站疫情通告 \(Information Release - Zhejiang Provincial Health Commission\)](#)

⁸⁷ [CSSEGISandData/COVID-19: Novel Coronavirus \(COVID-19\) Cases, provided by JHU CSSE](#)

⁸⁸ ['Confusion breeds distrust:' China keeps changing how it counts coronavirus cases](#)

polymerase chain reaction (RT-PCR) test in addition to genome sequencing to confirm the SARS-CoV-2 infection in suspected cases.⁸⁹ China also no longer required failure in antibiotic treatment and began considering patients who were not exposed to markets in Wuhan but had contact with symptomatic persons from Wuhan.⁹⁰ On January 28, 2020, China began considering patients not necessarily linked to Wuhan with at least two out of the previous three required clinical manifestations.⁹¹ On February 13, 2020, China created a separate “clinically confirmed” case definition for the Hubei province, which counted patients who met clinical criteria through chest imaging and may not have had epidemiological links or a positive PCR test.⁹² On February 20, 2020, China reversed this decision and removed the separate “clinically confirmed” case definition for Hubei.⁹³ On March 4, 2020, China expanded the possible laboratory confirmation tests for SARS-CoV-2 to include serology.⁹⁴ We included this information in the dataset because it could have potentially changed the levels and short-term growth rates of the number of confirmed cases.

The testing regime date changes are encoded within the data cleaning script.

South Korea

We have collated a provincial-level time series health outcome dataset in South Korea from January 20, 2020 to April 6, 2020.

Most provinces in South Korea have been publishing data on their number of confirmed coronavirus cases. Seoul,⁹⁵ Daegu,⁹⁶ Gyeongsangbuk-do,⁹⁷ Jeollabuk-do,⁹⁸ and Sejong⁹⁹ provinces have been reporting the number of confirmed cases on a daily basis. For these provinces, we recorded this published health data.

Given that the province of Gangwon-do¹⁰⁰ does not report provincial-level health data, we refer to the daily number of new cases reported by each of its counties (Chuncheon-si,¹⁰¹ Wonju-si,¹⁰² Gangneung-si,¹⁰³ Taebaek-si,¹⁰⁴ Sokcho-si,¹⁰⁵ and Samcheok-si¹⁰⁶). As a result, we manually collected

⁸⁹ [Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China](#)

⁹⁰ [Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China](#)

⁹¹ [Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China](#)

⁹² [Why China's Huge Increase in New COVID-19 Cases Is Actually a Step in the Right Direction](#). We have found another data source indicating that this case definition change happened on February 5 ([Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China](#)); we control for both dates.

⁹³ [Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China](#)
[Affiliation; China records 2 straight days of fewer than 1,000 new COVID-19 cases](#)

⁹⁴ [Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China](#)

⁹⁵ [서울특별시 코로나19 발생현황 \(Seoul COVID-19 Status\)](#)

⁹⁶ [대구광역시 코로나19 확진자 추이 \(Daegu COVID-19 The Confirmed Cases\)](#)

⁹⁷ [경상북도 코로나19 발생동향 \(Gyeongsangbuk-do COVID-19 Status\)](#)

⁹⁸ [전라북도 코로나19 일일상황보고 \(Jeollabuk-do COVID-19 Daily Reports\)](#)

⁹⁹ [세종특별자치시 코로나19 현황판 \(Sejong COVID-19 Status\)](#)

¹⁰⁰ [강원도청 \(Gangwon-do Provincial Government\)](#)

¹⁰¹ [춘천시 코로나19 현황 \(Chuncheon COVID-19 Status\)](#)

¹⁰² [원주시 코로나19 현황 \(Wonju COVID-19 Status\)](#)

¹⁰³ [강릉시 코로나바이러스감염증-19 비상대책 \(Gangneung COVID-19 Emergency Plan\)](#)

¹⁰⁴ [태백시청 \(Taebaek City Government\)](#)

¹⁰⁵ [속초시청 \(Sokcho City Government\)](#)

¹⁰⁶ [삼척시청 \(Samcheok City Government\)](#)

the number of new confirmed cases from each county's webpage and aggregated the numbers to the provincial level.

The remaining provinces (Gyeonggi-do,¹⁰⁷ Incheon,¹⁰⁸ Busan,¹⁰⁹ Ulsan,¹¹⁰ Gwangju,¹¹¹ Chungcheongnam-do,¹¹² Chungcheongbuk-do,¹¹³ Gyeongsangnam-do,¹¹⁴ Jeju,¹¹⁵ and Jeollanam-do¹¹⁶) did not explicitly publish the number of cumulative confirmed cases. However, they did publish patient-level data, including the date when patients had tested positive. For these provinces, we constructed the measure of cumulative confirmed cases by counting the number of daily confirmed cases and adding it to the previous date's total.

Most provinces did not publish the number of deaths. Instead, we checked the daily policy briefings posted on the government homepages mentioned in the footnotes and manually collected mortality data. In instances when mortality data were not found in the briefings, we obtained the mortality data from other sources, such as through social media sources (e.g. Facebook) and blogs maintained by local governments. Lastly, we supplemented these sources with mortality data reported in news articles.

Testing regime changes:

We collected information on testing regime changes using press releases from the Korean Center for Disease Control and Prevention (KCDC). In the press release menu, the KCDC uploaded daily briefing announcements which contained information on testing criteria and changes to its testing regime.¹¹⁷ Initially, the South Korean government only tested people who: 1) demonstrated respiratory symptoms within 14 days after visiting Wuhan South China Seafood Wholesale Market and 2) those who had pneumonia symptoms within 14 days after returning from Wuhan.¹¹⁸

As the outbreak spread, the KCDC broadened the criteria for testing. Starting January 28, 2020, the agency isolated 1) those who had fever or respiratory symptoms upon returning from Hubei province and 2) those who had symptoms of pneumonia upon returning from mainland China.^{119,120} We coded this as the first change in the testing regime.

¹⁰⁷ [경기도 코로나19 발생동향 \(Gyeonggi-do COVID-19 Status\)](#)

¹⁰⁸ [인천광역시 코로나19 상황판 \(Incheon COVID-19 Status\)](#)

¹⁰⁹ [부산광역시 코로나19 발생현황 \(Busan COVID-19 Status\)](#)

¹¹⁰ [울산광역시 코로나19 발생현황 \(Ulsan COVID-19 Status\)](#)

¹¹¹ [광주광역시청 \(Gwangju COVID-19 Status\)](#)

¹¹² [충청남도 코로나19 발생현황 \(Chungcheongnam-do COVID-19 Status\)](#)

¹¹³ [충청북도 코로나19 발생현황 \(Chungcheongbuk-do COVID-19 Status\)](#)

¹¹⁴ [코로나19 경상남도 현황 \(COVID-19 Gyeongsangnam-do Status\)](#)

¹¹⁵ [제주특별자치도 코로나19 현황 \(Jeju COVID-19 Status\)](#)

¹¹⁶ [전라남도 코로나19 현황 \(Jeollanam-do COVID-19 Status\)](#)

¹¹⁷ [List | Press Release | News Room : KCDC](#)

¹¹⁸ [중국 후베이성 우한시 폐렴환자 집단발생 | 보도자료 | 알림·자료 \(Pneumonia Outbreak in Wuhan City, Hubei, China\)](#)

¹¹⁹ [신종코로나바이러스감염증 국내 발생 현황\(1월 26일, 사례정의 확대\) | 보도자료 | 알림·자료 \(COVID-19 Domestic Status \(Jan 26, Case Definition Broadened\)\)](#)

¹²⁰ NB: The KCDC English website explains the testing regime change in a more condensed format: "Any citizens identified with a fever or respiratory symptoms and have visited Wuhan will be isolated and tested at a nationally designated isolation hospital, and any foreigners staying in Korea will be conducted in cooperation with police." [Urges cooperation in preventing the spread of 2019-nCoV in community | Press Release | News Room : KCDC](#)

The second testing regime change occurred on February 4, 2020, when the KCDC announced that people who had had any “routine contacts” with confirmed cases were required to self quarantine for a 14-day period. The agency defines two categories of contacts: close contacts and routine contacts. The former is defined as a person who has been within two meters of, in the same room as, or exposed to any respiratory secretions of an infected individual. The latter refers to whether the individual conducted any activity in the same place and at the same time as the infected person. Prior to this regime change, the KCDC separated those two cases and applied different quarantine policies; starting February 4, 2020, any routine contacts were also required to be self-quarantined.¹²¹

Shortly thereafter, South Korea aggressively expanded the scope of their testing. Starting February 7, 2020, the KCDC broadened the definition of suspected cases to 1) anyone who developed a fever or respiratory symptoms within 14 days after returning from China, 2) anyone who developed a fever or respiratory symptoms within 14 days after being in close contact with a confirmed case, and 3) anyone suspected of contracting COVID-19 based on their travel history to affected countries and their clinical symptoms.¹²² Moreover, the KCDC announced that the test would be free for all suspected cases and confirmed cases.¹²³ As a result of these efforts, KCDC announced that they would begin to test 3,000 people daily, a marked increase from only 200 people a day previously.¹²⁴

The KCDC revised their guidelines on February 20, 2020 in order to test more people. Their press release stated: “Suspected cases with a medical professional’s recommendation, regardless of travel history, will get tested. Additionally, those who are hospitalized with unknown pneumonia will also be tested. Lastly, anybody in contact with a diagnosed individual will need to self-isolate, and will only be released when they test negative on the thirteenth day of isolation.”¹²⁵

As the number of patients grew rapidly, the KCDC decided to focus on more vulnerable groups. In their February 29, 2020 press release, the agency stated: “The KCDC has asked local government and health facilities to focus on tests and treatment, especially targeting those aged 65+ and those with underlying conditions who need early detection and treatment.” This change was coded as our next testing regime change in the dataset.¹²⁶

On March 22, 2020, the KCDC began conducting COVID-19 diagnostic testing for every inbound traveler entering from Europe. This was coded as another testing regime change. Of the 1,442 inbound travelers from Europe arriving March 22, 2020, 152 were symptomatic and were quarantined and

¹²¹ [알림 > 보도자료 내용보기 "신종 코로나바이러스 감염증 대응지침 일부 변경" \(Revision in the Guidance Documents for COVID-19\)](#)

¹²² [The updates on novel Coronavirus in Korea \(since 3 January\) | Press Release | News Room : KCDC](#)

NB: The date of this press release is February 8, 2020, but the definition of “suspected cases” was effective starting from February 7, 2020.

¹²³ NB: The testing fee was already somewhat affordable; a person needed to pay 160,000 KRW (about \$130 USD). A related article can be found here: [5 신종코로나 진단검사 비용은 얼마? \(How much is the COVID-19 testing fee?\)](#)

¹²⁴ [신종 코로나바이러스 감염증 중앙사고수습본부 정례 브리핑 \(2월 7일\) \(Daily briefing on COVID-19, February 7\)](#)

¹²⁵ [The updates of COVID-19\(as of Feb.19\) in Korea | Press Release | News Room : KCDC](#)

¹²⁶ [The updates of COVID-19 in Korea \(As of 29 Feb. 2020\)](#)

tested at an airport quarantine facility. The remaining 1,290 travelers were asymptomatic and were moved to a temporary living facility to be tested.¹²⁷

On March 27, 2020, this policy was expanded, where all inbound travellers from the US with symptoms (regardless of nationality) were required to be tested at the airport.¹²⁸ We code this as our final testing regime change.

The data on the testing regime date changes are in the “*KOR_policy_data_sources.csv*.”

Italy

We have collated a regional and provincial level time series health outcome dataset in Italy from February 24, 2020 to April 7, 2020.

This data came from the [GitHub repository](#) maintained by the Italian Department of Civil Protection (*Dipartimento della Protezione Civile*). Health outcomes included the number of confirmed cases, the number of deaths, the number of recoveries, and the number of active cases. These figures have been updating daily at 5 or 6 pm (Central European Time). The regional-level dataset was pulled directly from “*dati-regioni/dpc-covid19-ita-regioni.csv*,” and the provincial-level dataset was pulled from “*dati-province/dpc-covid19-ita-province.csv*.”

Testing regime changes:

The testing regime change in Italy occurred when the Director of Higher Health Council announced on February 26, 2020 that COVID-19 testing would only be performed on symptomatic patients, as the majority of the previous tests performed were negative.

The data on the testing regime date changes are in the “*ITA_policy_data_sources.csv*.”

Iran

We have collated a provincial-level time series health outcome dataset in Iran from February 19, 2020 to March 22, 2020 (however health data are not reported by the Iranian Ministry of Health for March 2-3, 2020).

The Iranian government had been announcing its new daily number of COVID-19 confirmed cases at the provincial level on the Ministry of Health’s website. This data has been compiled daily in the table “New COVID-19 cases in Iran by province”¹²⁹ located in the “2020 coronavirus pandemic in Iran” article on Wikipedia.

¹²⁷ [List | Press Release | News Room : KCDC](#)

¹²⁸ [The updates on COVID-19 in Korea as of 27 March | Press Release | News Room : KCDC](#)

¹²⁹ [2020 coronavirus pandemic in Iran](#)

We spot-checked the data in the Wikipedia table against the Iranian Ministry of Health announcements¹³⁰ using a combination of Google Translate and a comparison¹³¹ of the numbers in the announcements (which were written in Persian script) to the Persian numbers.

Testing regime changes:

On March 6, 2020, the Ministry of Health announced¹³² a national coronavirus plan, which included contacting families by phone to identify potential cases, along with the disinfecting of public places. The plan was to begin in the provinces of Qom, Gilan, and Isfahan, and then would be rolled out nationwide. On March 13, 2020, the government announced a military-enforced home isolation policy throughout the nation.¹³³ This announcement included nationwide disinfecting of public places. While a follow-up announcement of the March 6 high testing regime stating its complete rollout was not found, the March 13 announcement did reference the implementation of the public spaces component of the earlier plan across the country. We thus assumed that the high testing regime had also been fully rolled out on March 13, 2020.

Testing Regime Changes:

The data on the testing regime date changes are in the “*IRN_policy_data_sources.csv*.”

France

We have collated a regional-level time series confirmed cases dataset in France from February 15, 2020 to March 25, 2020, and regional-level time series hospitalization data from March 3, 2020 to April 6, 2020.

We used the number of confirmed COVID-19 cases by *région* from France’s government website.¹³⁴ The sources listed for this dataset were the French public health website,¹³⁵ the Ministry of Solidarity and Health,¹³⁶ French newspapers that reported government information,¹³⁷ and regional public health websites.¹³⁸ Given that these data were not published on a daily basis, we supplemented the dataset by scraping the number of confirmed cases by *région* on the French public health website through March 25, 2020, which is the last date the subnational case data are made publicly available.¹³⁹

¹³⁰ Example of Ministry of Health data:

[شناسایی 1209 بیمار جدید مبتلا به کوید 19 در کشور](#) (Identification of 1209 new patients with COVID-19 in the country)

¹³¹ Google Translate sometimes translates various Persian numbers as "1". Persian numbers compared here: [Persian numbers](#)

¹³² [ANOTHER senior Iranian official dies from coronavirus](#)

¹³³ [Revolutionary Guards to enforce coronavirus controls in Iran](#)

¹³⁴ [Fr-SARS-CoV-2](#)

¹³⁵ [Infection à coronavirus](#)

¹³⁶ [Points de situation coronavirus COVID-19](#)

¹³⁷ [France 3 Régions: Actualités](#)

¹³⁸ [Agence régionale de santé | Agir pour la santé de tous](#)

¹³⁹ [Infection au nouveau Coronavirus \(SARS-CoV-2\), COVID-19, France et Monde](#)

Hospitalization data come from the same source¹⁴⁰ (*Santé Publique France*) as the case data. *Santé Publique France* announced they would stop posting regional-level case data because they were not reliable, and only provide hospitalization data instead.

Testing Regime Changes:

The one testing regime change in France occurred on March 13, 2020 with the beginning of the epidemic “stade 3”, when the government started to give severe cases in hospitals priority for testing.¹⁴¹ The testing regime date changes are encoded within the data cleaning script.

United States

We have collated a state-level time series health outcome dataset in the United States from January 22, 2020 to April 7, 2020.

The data come from the Github repository associated with the usafacts.org interactive dashboard. As of the time of writing, the data are available [here](#). The repository and dashboard are updated essentially in real-time, at least daily.

Testing regime changes:

To determine the testing regime, we used estimated daily counts of the cumulative number of tests conducted in every state, as aggregated by the largely crowdsourced effort named "The Covid Tracking Project" (covidtracking.com). We estimated the total number of tests as the sum of confirmed positive and negative cases. For some states and some days, there have been no negative case counts, in which case we utilize just the confirmed positive cases. We also ensured that the confirmed number of positive cases agreed with the counts in the John Hopkins University COVID-19 dataset.¹⁴²

We programmatically determined possible testing regime changes by filtering for any consecutive days during which the testing rate increased at least 250% from one day to the next, and where this jump was an increase of at least 150 total tests over one day. After visually inspecting the candidates, we confirmed that the automatically detecting testing regime changes represent visually distinguishable changes in testing rates. The testing regime date changes are encoded within the data cleaning script.

Policy Data

The policy events, datasets, and sources used in this paper are described below. For each country, the relevant country-specific policies identified were then mapped to a harmonized policy categorization used across all countries.

¹⁴⁰ <https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/>

¹⁴¹ [Coronavirus : en quoi consiste le « stade 3 » de l'épidémie ?](#)

¹⁴² [CSSEGISandData/COVID-19: Novel Coronavirus \(COVID-19\) Cases, provided by JHU CSSE](#)

The policy categories are by default coded as binary variables, where $[policy_variable] = 0$ before the policy is implemented in that area, and $[policy_variable] = 1$ on the date the policy is implemented (and for all subsequent dates until the policy is lifted). There are instances when the value of the policy variable is between 0 and 1; for further details, refer to the Policy Intensity subsection.

The main policy categories identified across the six different countries fall into four broad classes:

1. Restricting travel:
 - a. *travel_ban_local*: A policy that restricts people from entering or exiting the administrative area (e.g. county or province) treated by the policy.
 - b. *travel_ban_intl_in*: A policy that either bans foreigners from specific countries from entering the country, or requires travelers coming from abroad to self-isolate upon entering the country.
 - c. *travel_ban_intl_out*: A policy that suspends international travel to specific foreign countries that have high levels of COVID-19 outbreak.
 - d. *travel_ban_country_list*: A list of countries for which the national government has issued a travel ban or advisory. This information supplements the policy variable *travel_ban_intl_out*.
 - e. *transit_suspension*: A policy that suspends any non-essential land-, rail-, or water-based passenger or freight transit.
2. Distancing through cancellation of events and suspension of educational/commercial/religious activities:
 - a. *school_closure*: A policy that closes school and other educational services in that area.
 - b. *business_closure*: A policy that closes offices, non-essential businesses, and non-essential commercial activities in that area. Non-essential services are defined by area. This policy also includes the limiting of business hours and reducing restaurant and bar operations.
 - c. *religious_closure*: A policy that prohibits gatherings at a place of worship, specifically targeting locations that are epicenters of the COVID-19 outbreak. See the section on Korean policy for more information on this policy variable.
 - d. *work_from_home*: A policy that requires people to work remotely. This policy may also include encouraging workers to take holiday/paid time off.
 - e. *event_cancel*: A policy that cancels a specific pre-scheduled large event (e.g. parade, sporting event, etc). This is different from prohibiting all events over a certain size.
 - f. *no_gathering*: A policy that prohibits any type of public or private gathering. (whether cultural, sporting, recreational, or religious). Depending on the country, the policy can prohibit a gathering above a certain size, in which case the number of people is specified by the “*no_gathering_size*” variable.
 - g. *no_gathering_inside*: A policy that specifically prohibits indoor gatherings. See the section on French policy for more information on this policy variable.
 - h. *no_demonstration*: A policy that prohibits protest-specific gatherings. See the section on Korean policy for more information on this policy variable.

- i. *social_distance*: A policy that encourages people to maintain a safety distance (often between one to two meters) from others. This policy differs by country, but includes other policies that close cultural institutions (e.g. museums or libraries), or encourage establishments to reduce density.
 - j. *welfare_services_closure*: A policy that mandates the closure of social welfare facilities, specifically mental rehabilitation facilities, social welfare centers, and homeless use facilities. See the section on Korean policy for more information on this policy variable.
3. Quarantine and lockdown:
- a. *pos_cases_quarantine*: A policy that mandates that people who have tested positive for COVID-19, or subject to quarantine measures, have to confine themselves at home. The policy can also include encouraging people who have fevers or respiratory symptoms to stay at home, regardless of whether they tested positive or not.
 - b. *home_isolation*: A policy that prohibits people from leaving their home regardless of their testing status. For some countries, the policy can also include the case when people have to stay at home, but are allowed to leave for work- or health-related purposes. This policy is also a superset of other, less stringent policies. As such, the estimates of the full home isolation effect encapsulates the effect one would observe if none of the less stringent policies were yet in place when home isolation was implemented.
4. Additional policies
- a. *emergency_declaration*: A decision made at the city/municipality, county, state/provincial, or federal level to declare a state of emergency. This allows the affected area to marshal emergency funds and resources as well as activate emergency legislation.
 - b. *paid_sick_leave*: A policy where employees receive pay while they are not working due to the illness.

Optional policies:

In the cases when the aforementioned policies are optional, we denote this as *[policy_variable]_opt*.

Population weighting of policy variables:

In cases where only a portion of the administrative unit (e.g. half of the counties within the state) are affected by the implementation of the policy, we weight the policy variable by the percentage of population within the administrative unit that is treated by the policy. This is denoted as *[policy_variable]_popwt*, and the value that this variable can take on is a continuous number between 0 and 1. Sources for the population data are detailed in a later section.

Policy intensity

policy_intensity is a continuous value between 0 and 1 that modulates the intensity/restrictiveness of a policy. By default this value is 0 when the policy has not been implemented and 1 when the policy is implemented (i.e. the policy variables are treated as indicator variables). However, in instances when a policy has evolved over time, then less restrictive implementations of the policy are weighted by a *policy_intensity* value that is between 0 and 1, and the most restrictive version of the policy has a value of 1.

For simplicity, if a given policy has undergone one version change, then the *policy_intensity* of the less restrictive edition is equal to 0.5, and the value of the more restrictive edition is equal to 1. If there have been two version changes, then the *policy_intensity* of the less restrictive edition is equal to 0.33, the value of the second most restrictive edition is equal to 0.67, and the value of the most restrictive edition is equal to 1, etc.

Additionally, in certain countries, a type of broad policy category (e.g. *business_closure*) could have multiple sub-policies that fall within this category (e.g. closing of restaurants/bars, closing of gyms/casinos/recreational facilities, and closing of non-food/non-recreational facilities). In this case, the policies are treated additively, where each of the three sub-policies have a *policy_intensity* weight of 0.33, and if an administrative unit has implemented only two of three of the policies on a given date, then the *business_closure* variable would equal 0.66.

Lastly, for some countries, the deployment of a certain policy (e.g. *home_isolation*) implies that other policies are also turned on (e.g. *travel_ban_local* or *work_from_home*), even if there is no separate policy that explicitly states this. The relationships between policies are country-specific, and these assumptions are encoded in our country-by-country policy data dictionary, which can be found in the GitHub repository at [data/raw/multi_country/policy_implication_rules.json](#).

In the instances when the implementation of a certain policy implies that other policies are also in place, these secondary policies are only drawn from the set of policies that have already been deployed within that country. For example, in Italy when *home_isolation* is enacted for an administrative unit on a specific date, this implies that *business_closure* and *work_from_home* should also be turned on for this administrative unit on this date, since these two policies have been implemented within the country. However, this would not imply that the policy variable *no_demonstrations* (a policy that has been implemented in other countries within our sample, but not in Italy) is turned on, since this policy had not been deployed in any Italian administrative unit prior to this specific date.

More specifically, we compute *policy_intensity* using this approach:

1. For non-population-weighted policy variables: For a given policy category on a specific date (e.g. *business_closure* on March 15, 2020), take the maximum of the mandatory policy intensities for all units lower (e.g. *Adm0*) than, equal to, and higher (e.g. *Adm2*) than the

analysis unit (e.g. *Adm1*). Assign this maximum *policy_intensity* value to the unit of analysis. If there is no mandatory version of the policy that applies to the unit of analysis, then take the maximum of the optional policy intensities and assign it to the optional policy variable for the analysis unit.

2. For population-weighted policy variables: Take the maximum of the mandatory policy intensities for all units lower (e.g. *Adm0*) than and equal to the analysis unit (e.g. *Adm1*), and assign that as the default mandatory intensity for all units higher (e.g. *Adm2*). If the policy is not mandatory at the analysis or lower unit, then assign the maximum of the optional *policy_intensity* value as the default optional intensity for all higher units. For any higher unit that has a specific policy, assign the appropriate version (mandatory or optional) of the policy variable at that higher unit the maximum of that intensity and the default intensity, with mandatory always taking priority over optional. For all that don't have a specific policy, assign them the default intensity (again, assigning this to the optional or mandatory version as appropriate). Then calculate the population-weighting at the analysis unit level (e.g. *Adm1*), separately for both optional and mandatory variables. Each higher unit should only have a non-zero intensity for optional or for mandatory (or neither), but not both.
3. For broadly defined policy variables like *social_distance* that could encompass a variety of country-specific policies: The *policy_intensity* assignment differs by country. If the specific policies employed at the various administrative levels are the same policy, then the approach in (1) is used. If they are different policies within the same broad category, then we add instead of taking the maximum, allow for both optional and mandatory policies, and otherwise follow the approach of (1). This addition is appropriate across different administrative divisions because of (1).

China

We obtain data on China's policy response to the COVID-19 pandemic by culling data on the start dates of travel bans and lockdowns at the city-level from the "2020 Hubei lockdowns" Wikipedia page,¹⁴³ and various news reports.

To combat the spread of COVID-19, the Chinese government imposed Level 1 emergency declarations, travel restrictions and quarantine measures. The lockdown of the city of Wuhan, the origin of the pandemic, occurred on January 23, 2020. Immediately following the Wuhan lockdown, neighboring cities followed suit, banning travel into and out of their borders, shutting down businesses, and placing residents under household quarantine. The same policy measures were implemented in cities across China for the next three weeks.

Starting around the same time as the initial lockdown, from January 23 to January 29, 2020, all provincial governments declared Level 1 emergency responses.¹⁴⁴ The declaration often included the

¹⁴³ [2020 Hubei lockdowns](#)

¹⁴⁴ Tian, Huaiyu, Yonghong Liu, Yidan Li, Chieh-Hsi Wu, Bin Chen, Moritz UG Kraemer, Bingying Li et al. "An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China." *Science* (2020).

closing of entertainment venues, banning of public gatherings, and extensive temperature monitoring at airports, railway stations and highway checkpoints.¹⁴⁵

Some lockdowns occurred during the national Chinese New Year holiday (January 24–30, 2020) when schools and most workers were on break. On January 27, 2020, China extended the official holiday to February 2, 2020, while many additional provinces delayed resuming work and opening schools for even longer.¹⁴⁶ The Chinese New Year holiday is analogous to containment policies such as school closures and restrictions on non-essential work. We do not specifically estimate the effect of this holiday extension, as most cities were in lockdown during the extended holiday, and a lockdown is a more restrictive containment measure. A lockdown requires all residents to stay home, except for medical reasons or essential work, and only allows one person from each household to go outside once every one to five days (exact policy varied by city).

South Korea

We obtained data on South Korea's policy response to the COVID-19 pandemic from various news sources, as well as press releases from the Korean Centers for Disease Control and Prevention (KCDC), the Ministry of Foreign Affairs, and local governments' websites. The policy variables coded in the dataset are: *welfare_services_closure*, *business_closure_opt*, *emergency_declaration*, *no_demonstration*, *religious_closure*, *event_cancel*, *school_closure*, *social_distance_opt*, *travel_ban_intl_in_opt*, *travel_ban_intl_out_opt*, *work_from_home_opt*, and *pos_cases_quarantine*.

On February 28 2020, the KCDC recommended the closure of 14 types of social welfare facilities to reduce the spread of infection among vulnerable groups in the population.¹⁴⁷ These include childcare centers, vocational rehabilitation centers for the disabled, senior citizen centers, mental rehabilitation facilities, and homeless use facilities. We code this in the variable *welfare_services_closure*. Even though it was technically a recommendation, we did not code this policy as optional because a majority of facility types listed in the press release (senior citizen centers, job centers, childcare centers, etc.) are under public administration, so these facilities likely would have followed recommendations. Indeed, some news articles have reported that all children's centers in Busan are closed¹⁴⁸ as well as over 3,600 facilities in Seoul.¹⁴⁹

We created another variable, *business_closure_opt*, which applies to two provinces: Seoul and Gyeonggi-do. On March 11, 2020, the mayor of Seoul advised that popular commercial establishments such as karaoke places, clubs, and cyber cafes be closed.¹⁵⁰ Seven days later, the governor of Gyeonggi-do issued an executive order limiting the usage of commonly frequented commercial

¹⁴⁵ http://www.gov.cn/xinwen/2020-01/29/content_5472881.htm

¹⁴⁶ [China Extends Lunar New Year Holiday to Feb 2, Shanghai to Feb 9](#)

¹⁴⁷ [코로나19 여파 “사회복지 이용시설 휴관 권고” \(Social welfare facilities recommended to shut down\)](#)

¹⁴⁸ [부산 지역아동센터 모두 휴관...더 외로운 저소득층 아이들 \(Busan child-care facilities shut down - worse for the lower-income children\)](#)

¹⁴⁹ [서울시, 노인복지관 등 사회복지시설 3601곳 휴관 \(Seoul, 3601 social welfare facilities shut down\)](#)

¹⁵⁰ [코로나19 확산을 막기 위한 서울시 일일보고 \(Seoul daily report on limiting the spread of COVID-19\)](#)

establishments and requiring a higher standard of cleanliness.¹⁵¹ We coded this as an optional business closure given that the policy discourages usage of these facilities but did not explicitly order them to shut down.

Daegu and Gyeongsangbuk-do have been two of the regions hardest hit by COVID-19. The government of South Korea declared an emergency for those two areas on March 15, 2020.¹⁵² We incorporated this information into the variable *emergency_declaration*.

The variable *no_demonstration* reflects the efforts of some regions limiting any protests calling for slowing the spread of the outbreak. On February 24, 2020, Incheon stopped a protest in front of the Incheon Metropolitan City Hall.¹⁵³ Two days later, Seoul prohibited protests in downtown areas where massive demonstrations used to take place.¹⁵⁴

Many province level COVID-19 policies have targeted religious gatherings at Shincheonji Church of Jesus, since its religious gatherings have been linked to the explosion in the number of cumulative confirmed cases. Provincial governments tried to shut down Shincheonji-related places of worship, and the related policy implementation is encoded in the variable *religious_closure*. The regions which utilized this policy option are: Daegu,¹⁵⁵ Gyeongsangbuk-do,¹⁵⁶ Seoul,¹⁵⁷ Jeju,¹⁵⁸ Gyeonggi-do,¹⁵⁹ Jeollanam-do,¹⁶⁰ Gyeongsangnam-do,¹⁶¹ Incheon,¹⁶² Ulsan,¹⁶³ Busan,¹⁶⁴ Jeollabuk-do,¹⁶⁵ Chungcheongbuk-do,¹⁶⁶ Gwangju,¹⁶⁷ Chungcheongnam-do,¹⁶⁸ and Daejeon.¹⁶⁹

¹⁵¹ (브리핑) 이재명, "PC방·노래연습장·클럽형태업소에 밀접이용제한 행정명령" (Cyber cafes, karaokes, and clubs under the administrative order limiting close-distance usage)

¹⁵² 19> 뉴스 & 이슈 > 보도자료 내용보기 " [카드뉴스] 중앙재난안전대책본부 정례브리핑(3.14.), 특별재난지역선포(대구, 경북 경산·청도·봉화) " (Daily briefing on announcing the emergency declaration for the regions: Daegu; Gyeongsangbuk-do Gyeongsan, Cheongdo, Bonghwa)

¹⁵³ 보도자료 조회 "인천시, 인천애들 잠정 사용중단(금지) 조치" (Incheon prohibits usage of Incheon City Government Square)

¹⁵⁴ 코로나19 확산 방지를 위해 도심 집회 제한 강화 (Stronger limits on demonstrations in downtown)

¹⁵⁵ 신천지 관련시설 폐쇄조치, 확산 방지에 행정력 집중...대구시 경찰청과 긴밀히 협조 (Shincheonji-related facilities shut down, Daegu struggling to limit the spread of the virus with the police power)

¹⁵⁶ 경북, 신천지 1612명 중 221명 확진...31번이 156명 옮겼다 (Gyeongsangbuk-do, 221 out of 1612 tested positive, the 31st patient responsible for infecting 156 people)

¹⁵⁷ 서울시, 신천지 집회 시설 폐쇄 결정 (Seoul shuts down Shincheonji-related facilities)

¹⁵⁸ 제주 신천지 신도 전원 능동감시 종료...집회 금지는 유지 (Shincheonji believers now free from monitoring, still religious gatherings prohibited)

¹⁵⁹ 경기도, 신천지 353개 시설 14일간 강제폐쇄·집회금지 조치 내려 (Gyeonggi-do shuts down 353 Shincheonji facilities for 14 days)

¹⁶⁰ 광주일보 "전남도, 신천지 교회·시설 58곳 강제폐쇄 행정명령 발동" (Jeollanam-do shuts down 58 Shincheonji-related facilities)

¹⁶¹ 여성조선 "경남, 신천지 시설폐쇄 및 집회 금지 행정명령 발동" (Gyeongsangnam-do shuts down Shincheonji facilities and forbids religious gatherings)

¹⁶² 인천시, 신천지교회 종교시설 추가 폐쇄조치 시행 | 기관 소식 | 정책·정보 (Incheon shuts down more Shincheonji facilities)

¹⁶³ 울산시, 신천지교회 및 부속기관 폐쇄 조치 (Ulsan shuts down Shincheonji facilities)

¹⁶⁴ [코로나19] 부산, 신천지 시설 폐쇄·집회 금지 2주 추가 연장 (Busan shuts down Shincheonji facilities for two weeks more)

¹⁶⁵ 전북 신천지 시설 폐쇄·집회 금지 연장... (Jeollabuk-do extends the period of shutting down Shincheonji facilities)

¹⁶⁶ 충북 신천지 시설 38개소 폐쇄·방역 완료 (Chungcheongbuk-do shuts down 38 Shincheonji facilities)

¹⁶⁷ 광주광역시, 신천지 시설 폐쇄 행정명령 (Gwangju shuts down Shincheonji facilities)

¹⁶⁸ 충남도, 신천지 관련 시설 58개소 폐쇄 (Chungcheongnam-do shuts down 58 Shincheonji facilities)

¹⁶⁹ 대전광역시 신천지 시설 방역 및 폐쇄조치 현황입니다. 신천지 신도 및 교육생 현황입니다. (Status report on Shincheonji facilities shutdown, and Shincheonji believers and trainees)

Many provinces have also canceled public events organized by local administrative agencies. We code this policy in the variable *event_cancel*. The regions which exercised this policy are: Seoul,¹⁷⁰ Daegu,¹⁷¹ Gangwon-do,¹⁷² Chungcheongbuk-do,¹⁷³ Chungcheongnam-do,¹⁷⁴ Sejong,¹⁷⁵ Daejeon,¹⁷⁶ Gyeongsangbuk-do,¹⁷⁷ Gyeongsangnam-do,¹⁷⁸ Jeju,¹⁷⁹ Gyeonggi-do,¹⁸⁰ Ulsan,¹⁸¹ Gwangju,¹⁸² Busan,¹⁸³ Incheon,¹⁸⁴ Jeollanam-do,¹⁸⁵ and Jeollabuk-do.¹⁸⁶

The policy variable *school_closure* has been turned on for the entirety of the Korean time series dataset. This is because all schools were already on vacation during the beginning of the outbreak, and the government then postponed their start dates. At the time of writing, the Ministry of Education announced that schools would be kept closed until April 3, 2020.¹⁸⁷ Therefore, this policy variable is always equal to 1 in the dataset.

social_distance_opt has been turned on from February 29, 2020, when KCDC recommended social distancing as one of the main tools to deal with the outbreak. In their press release, they recommended that people maintain personal hygiene and practice ‘social distancing’ until the beginning of March, an important point of this outbreak.¹⁸⁸ In the case of Daegu, the hardest-hit region in the country, we coded the variable as 1 starting from February 22, 2020, based on the statement, “It is recommended for residents in Daegu to minimize gathering events and outdoor activities.”¹⁸⁹

The first travel restriction for incoming travelers (*travel_ban_intl_in_opt*) was implemented on January 28, 2020. It is worth noting that it was not a total prohibition of incoming visitors; rather, it means inbound travellers were subject to COVID-19 specific emergency measures. KCDC mentioned that starting on January 28, 2020 “any travellers depart[ing] from China [would] be a subject to strengthened screening and quarantine measures.”¹⁹⁰ On February 12, 2020, KCDC broadened the list

¹⁷⁰ ‘신종 코로나 확산’ 2월 취소 행사 확인하세요! (Event cancellation in February due to COVID-19)

¹⁷¹ 코로나19(Covid-19) 확산 2~3월 취소 행사 확인하세요! (Event cancellation in February and March due to COVID-19)

¹⁷² 연합뉴스 "강원5명 코로나19 확진..공공시설 출입제한·행사 연기·취소" (Five confirmed cases in Gangwon-do, public facilities shutdown, events delayed or canceled)

¹⁷³ 충북도, 코로나19 확산될라...행사 줄줄이 취소 (Chungcheongbuk-do cancels events due to COVID-19)

¹⁷⁶ '심각단계' 격상 코로나19 대응 시정브리핑 (The alert level raised, COVID-19 daily briefing)

¹⁷⁴ '신종코로나 유입 막자'...충남 대규모 체육·문화 행사 줄취소(종합) (Chungcheongnam-do cancels events due to COVID-19)

¹⁷⁵ 세종시 신종 코로나여파 각종 행사 취소 및 자제요청 (Sejong urges cancellation of events amid COVID-19 outbreak)

¹⁷⁷ 신종 코로나바이러스 여파로 경북도내 각종 축제·행사 취소 또는 연기 (Gyeongsangbuk-do cancels or delays events due to COVID-19)

¹⁷⁸ 신종 코로나 확산에 경남 지역행사 등 줄줄이 취소 (Gyeongsangnam-do cancels events due to COVID-19)

¹⁷⁹ 제주도내 행사 등 전면 취소, "코로나19 확산 방지 우선" (Jeju cancels events due to COVID-19)

¹⁸⁰ 신종 코로나바이러스 감염증 대응을 위한 도내 각종 행사 취소, 축소 방침 (Gyeonggi-do cancels events due to COVID-19)

¹⁸¹ 울산지역 주요행사 잇따라 취소·연기 (Ulsan cancels or delays events due to COVID-19)

¹⁸² 코로나바이러스감염증-19 대응 관련 취소 행사 현황 (2.28. 현재) (The list of events canceled due to COVID-19)

¹⁸³ 2020년 정월대보름 관측행사 취소 안내 (Daeboreum events canceled)

¹⁸⁴ 인천시, 코로나19 확산방지 강력조치 (Incheon strict policies for limiting the spread of the virus)

¹⁸⁵ 신종코로나 확산...전남 지자체, 행사 줄줄이 취소 (Jeollanam-do cancels events)

¹⁸⁶ 송하진 도지사, 코로나바이러스 대응 ‘올인’ (Governor of Jeollabuk-do makes every effort to fight against the virus)

¹⁸⁷ 전국 모든 유·초·중·고·특 개학 2주간 추가연기 결정 (코로나19) (All kindergarten, elementary schools, middle schools, and high schools are closed for two more weeks)

¹⁸⁸ The updates of COVID-19 in Korea, February 29

¹⁸⁹ The updates of COVID-19 in Korea as of 22 February

¹⁹⁰ The case definition of 2019 novel coronavirus will be expanded | Press Release | News Room : KCDC

of countries subject to the stricter measures to include Hong Kong and Macau.¹⁹¹ Subsequently, KCDC added Italy and Iran (on March 11, 2020);¹⁹² France, Germany, Spain, UK, and Netherlands (on March 15, 2020);¹⁹³ and any remaining European countries (March 15, 2020)¹⁹⁴ to their country list. On March 19, 2020, the policy was expanded to include all travelers arriving at port regardless of country of origin.¹⁹⁵

This restriction was not limited to inbound travellers. The government also issued advisories on countries where the number of infections had increased, which has been encoded as the variable *travel_ban_intl_out_opt*. The first outbound travel alert due to COVID-19 was announced on January 28, 2020: The Ministry of Foreign Affairs (MOFA) issued a Level 2 (Yellow) alert for any travel to mainland China, Hong Kong, and Macau.¹⁹⁶ Later, MOFA added Italy on February 28, 2020,¹⁹⁷ Japan on March 9, 2020,¹⁹⁸ and all European countries on March 16, 2020.¹⁹⁹ On March 18, 2020, KCDC strongly called for the cancellation or delay of all international travel on non-urgent matters.²⁰⁰ It should be noted that the Level 2 alert does not enable the government to prohibit travel to these destinations, which is why the policy was coded as optional.

There are four types of travel advisories distributed by the South Korean government: Level 1, Navy; Level 2, Yellow; Level 3, Red; and Level 4, Black.²⁰¹ Travel under the Level 4 alert is prohibited, and the government utilizes legal instruments to enforce the restriction. If people leave the country under the black alert, they will be subject to fines up to ten million KRW, or imprisonment up to a year. However, there is no enforcement instrument for the advisories up to Level 3. In that sense, we stated above that the banning policy does not mean prohibiting travel. Nevertheless, we coded the yellow alert as the first travel ban in our dataset, since Level 2 alerts are issued relatively rarely, such as during a significant demonstration²⁰² or military coup.²⁰³ As a result, we coded the Level 2 alert due to COVID-19 into the dataset for the policy analysis.

The policy variable *work_from_home_optional* indicates when KCDC began recommending that people work from home. On March 15, 2020, the KCDC press release stated: “Since contact with confirmed cases in an enclosed space increases the possibility of transmission, it is recommended to work at home or adjust desk locations so as to keep a certain distance among people in the office. More detailed guidelines for local governments and high-risk working environments will be distributed soon.”²⁰⁴

¹⁹¹ [Expand strict quarantine screening of 2019-nCoV to Hong Kong, Macao | Press Release | News Room : KCDC](#)

¹⁹² [The updates on COVID-19 in Korea as of 11 March | Press Release | News Room : KCDC](#)

¹⁹³ [Updates on COVID-19 in Korea \(as of 12 March\)](#)

¹⁹⁴ [The updates on COVID-19 in Korea as of 16 March | Press Release | News Room : KCDC](#)

¹⁹⁵ [목록 | 보도자료 | 알림·자료 \(The list of press release\)](#)

¹⁹⁶ [최신 여행경보단계 조정 \(The latest adjustment on the travel alert levels\)](#)

¹⁹⁷ [최신 여행경보단계 조정 \(The latest adjustment on the travel alert levels\)](#)

¹⁹⁸ [일본 전 지역\(후쿠시마 원전 주변지역 제외\)에 여행경보 2단계\(황색경보, 여행자제\)로 상향 조정 \(All Japanese region, other than the Fukushima nuclear reactor area, now under the level 2 travel alert\)](#)

¹⁹⁹ [최신 여행경보단계 조정 \(The latest adjustment on the travel alert levels\)](#)

²⁰⁰ [The updates on COVID-19 in Korea as of 18 March](#)

²⁰¹ [여행경보제도 소개 \(The description on the travel alert policy\)](#)

²⁰² [홍콩 여행경보 2단계\(여행자제\)로 상향 조정 \(Now travels to Hong Kong under the level two alert\)](#)

²⁰³ [기니의 여행경보단계 상향 조정 \(The alert level is raised against travels to Guinea\)](#)

²⁰⁴ [The updates on COVID-19 in Korea as of 11 March | Press Release | News Room : KCDC](#)

On March 22, 2020, the KCDC announced that all inbound travelers from Europe would be tested at the airport and subject to quarantine measures.²⁰⁵ Korean citizens and long-term visitors returning from abroad needed to home-quarantine for 14 days (even if they test negative for COVID-19), while short-term visitors would be actively monitored. Inbound travelers with no symptoms were required to stay at temporary facilities while awaiting their test results.²⁰⁶ We coded this as the policy variable *pos_cases_quarantine* modulated by *policy_intensity* = 0.25. When this policy was expanded on March 27, 2020 to include all symptomatic travelers arriving from the US,²⁰⁷ we coded this variable with a *policy_intensity* = 0.5. On April 1, 2020, these quarantine measures were extended to include inbound travelers arriving from all countries, with exceptions allowed only for limited cases (diplomatic missions etc.).²⁰⁸ This variable was then coded with *policy_intensity* = 0.75. Lastly, starting on April 5, 2020, the KCDC announced that inbound travelers who fail to comply with quarantine regulations are subject to imprisonment of up to 1 year or a fine of up to 10 million won for the violation of the Infectious Disease Control and Prevention Act. In addition, persons of foreign nationality who fail to comply may be subject to measures including deportation and entry ban in accordance with the Immigration Act.²⁰⁹ We then coded this variable with the *policy_intensity* = 1.

Italy

We have obtained data on Italy's policy responses to the COVID-19 pandemic primarily from the English version of the COVID-19 dossier "Chronology of main steps and legal acts taken by the Italian Government for the containment of the COVID-19 epidemiological emergency"²¹⁰ written by the Department of Civil Protection (*Dipartimento della Protezione Civile*), most recently updated on March 12, 2020. This dossier details the majority of the municipal, regional, provincial, and national policies rolled out between the start of the pandemic to present-day. We have supplemented these policy events with news articles that detail which administrative areas were specifically impacted by the additional policies.

The first major policy rollout was on February 23, 2020, when 11 municipalities across two provinces in Northern Italy were placed on lockdown. These policies included closing schools, cancelling public and private events and gatherings, closing museums and other cultural institutions, closing non-essential commercial activities, and prohibiting the movement of people into or out of the municipalities.

The second major policy rollout was on March 1, 2020, when two provinces and three regions in Northern Italy were placed on partial lockdown. These policies also included closing schools,

²⁰⁵ [The updates on COVID-19 in Korea as of 22 March | Press Release | News Room : KCDC](#)

²⁰⁶ [The updates on COVID-19 in Korea as of 23 March | Press Release | News Room : KCDC](#)

²⁰⁷ [The updates on COVID-19 in Korea as of 27 March | Press Release | News Room : KCDC](#)

²⁰⁸ [The updates on COVID-19 in Korea as of 31 March | Press Release | News Room : KCDC](#)

²⁰⁹ [The updates on COVID-19 in Korea as of 5 April | Press Release | News Room : KCDC](#)

²¹⁰ [Chronology of main steps and legal acts taken by the Italian Government for the containment of the COVID-19 epidemiological emergency](#)

cancelling public and private events and gatherings, closing museums, closing non-essential commercial activities, as well as limiting the number of people at places of worship, restricting operating hours of bars and restaurants, and encouraging people to work remotely.

The third major policy roll-out was on March 5, 2020, when all schools across the country were closed.

The fourth major policy roll-out was on March 8, 2020 when the region of Lombardy and 13 provinces in Northern Italy were placed on lockdown. These policies included the cancellation of public and private events and gatherings, closing of museums, encouraging people to work remotely, limiting the number of people at places of worship, restricting opening hours of bars and restaurants, mandating quarantine of people who tested positive for COVID-19, prohibiting the movement of people into or out of the affected area, and restricting movement within the affected area to only work or health-related purposes. Commercial activities were still allowed, as long as they maintained a safety distance of one meter apart per person within the establishment. All civil and religious ceremonies, including weddings and funerals, were suspended. During this same policy roll-out, the rest of the country faced less stringent policies: cancelling public and private events, closing museums, and requiring restaurants and commercial establishments to maintain a safety distance of one meter apart per person within the establishment.

The fifth major policy roll-out was announced on March 9, 2020, and went into effect on March 10, 2020, when lockdown policies applied to Northern Italy were rolled out to the entire country. Lastly, on March 11, 2020, the lockdown was changed to also cover the closing of any non-essential businesses and further restricted people from leaving their home.

After the death toll in Italy surpassed that of China on March 21, 2020, the Italian government increased the severity of their existing policies. Effective March 22, 2020, all non-essential industrial production and factories would be shut down across the country.²¹¹ Domestic travel was further restricted; people were not permitted to leave the municipality they were currently in except for urgent matters or emergencies.²¹² Lastly, in the hard-hit northern region of Lombardy, the regional government increased lockdown restrictions by banning all individual outdoor exercise or sporting activity.²¹³

Policy Intensity: We have modified the policy intensity of three different policy variables: *home_isolation*, *business_closure*, and *travel_ban_local*.

The *home_isolation* policy underwent three policy revisions:

1. The least restrictive version of the policy applies to when people were allowed to leave the house for work, health, and essential reasons (*policy_intensity* of *home_isolation* = 0.33).

²¹¹ [Virus News: Italian Industry Shuts Down](#)

²¹² [Italy Coronavirus News: Travel Ban Inside Country](#)

²¹³ [Coronavirus: Lombardy region announces stricter measures](#)

2. The moderate version of the policy applies to when people were allowed to leave the house only for health and essential reasons (which includes the ability to go outdoors for individual exercise/sporting activities) (*policy_intensity* of *home_isolation* = 0.67).
3. The most restrictive version of the policy applies to when people were allowed to leave the house only for health and essential reasons, but were no longer allowed to leave the house for individual exercise/sporting activities (*policy_intensity* of *home_isolation* = 1).

The *business_closure* policy underwent three policy revisions:

1. The least restrictive version of the policy applies to the limiting of restaurant hours (but other commercial activities were permitted) (*policy_intensity* of *business_closure* = 0.33),
2. the moderate version of the policy applies to the closing of all non-essential businesses, (*policy_intensity* of *business_closure* = 0.67),
3. and the most restrictive version of the policy applies to the closing of all non-essential industrial production and factories, in addition to the closing of non-essential businesses (*policy_intensity* of *business_closure* = 1).

Lastly, the *travel_ban_local* policy underwent two policy revisions:

1. The least restrictive version of the policy applies to when people were not allowed to enter/exit the affected administrative area, (*policy_intensity* of *travel_ban_local* = 0.5),
2. and the most restrictive version of the policy applies to a more restrictive ban on domestic travel that mandated that people had to stay in the municipality they were currently in (*policy_intensity* of *travel_ban_local* = 1).

Iran

For Iran's policy response to the COVID-19 pandemic, we relied on news media reporting as the primary source of policy information (mostly due to translation restrictions). We also relied on two timelines of pandemic events in Iran to help guide the policy search.^{214, 215}

The first major outbreak in Iran was connected to a major Shia pilgrimage in the city of Qom that brought Shiite pilgrims from Iran and throughout the Middle East, where they came to kiss the Fatima Masumeh shrine. It is possible that the disease was brought to Qom by a merchant traveling from Wuhan, China.²¹⁶ In addition, it is believed that the Iranian government knew of the COVID-19 outbreak prior to its February 21, 2020 parliamentary elections, but downplayed the risks associated with the disease as not to suppress voter turnout (given concerns that a low turnout would reflect poorly on its legitimacy).²¹⁷ The disease, initially centered in Qom and neighboring Tehran, spread rapidly throughout the country.

²¹⁴ [UPDATED: Timeline of the Coronavirus](#)

²¹⁵ [2020 coronavirus pandemic in Iran](#)

²¹⁶ [How Iran Became a New Epicenter of the Coronavirus Outbreak](#)

²¹⁷ [How Iran Became a New Epicenter of the Coronavirus Outbreak](#)

As the number of cases grew, the Iranian government started to increase the stringency of its response. The first case was reported on February 19, 2020 (two individuals who both were reported to have died that day). The next day, school closures were announced in the province of Qom and travel in the region was discouraged. By February 22, 2020 the government closed schools in 14 provinces and closed down major gathering sites such as football matches and theaters. By March 5, 2020 schools were closed nationwide and government employees were required to work from home. Home isolation was implemented by the military on March 13, 2020, which the media described as “the near-curfew follows growing exasperation among MPs that calls for Iranian citizens to stay at home had been widely ignored, as people continued to travel before the Nowruz New Year holidays.”

²¹⁸

France

We obtain data on France’s policy response to the COVID-19 pandemic from the French government website, press releases from each regional public health site, and Wikipedia.

The French government website contains a timeline of all national policy measures.²¹⁹ Each regional public health agency (*l’Agence Régionale de Santé*) in France posts press releases with information on the policies the région or départements within the région will implement to mitigate the spread and impact of the COVID-19 outbreak.²²⁰ The Wikipedia page on the 2020 coronavirus pandemic in France has collated information on the major policy measures taken in response to the COVID-19 pandemic.

²²¹

Starting February 29, 2020, France banned mass gatherings of more than 5,000 people nationwide, while some major sporting events were cancelled and a handful of schools closed to mitigate the spread of the virus. As more COVID-19 cases were confirmed during the following week, additional sporting events were canceled, more schools decided to close, and certain cities and *départements* limited mass gatherings to no more than 50 people, excluding shops, business, restaurants, bars, weddings, and funerals. Some régions closed early childhood establishments (e.g. nurseries, daycare centers) and prohibited visitors to elderly care facilities. On March 8, 2020, France banned mass gatherings of more than 1,000 people nationwide. Other schools, cities, and départements followed suit with additional school closures and limiting mass gatherings. On March 11, 2020, France prohibited all visits to elder care establishments. Starting March 16, 2020, France closed all schools nationwide. Between March 17, 2020 – March 23, 2020, governments at both the national level and région level implemented more restrictive lockdown policies, which included shelter-in-place measures,²²² the closing of public places,²²³ and banning of outside markets and severely restricting movement outside of the house.²²⁴

²¹⁸ [Revolutionary Guards to enforce coronavirus controls in Iran](#)

²¹⁹ [Info Coronavirus COVID-19](#)

²²⁰ [Agence régionale de santé | Agir pour la santé de tous](#)

²²¹ [Pandémie de Covid-19 en France](#)

²²² [Info Coronavirus COVID-19](#)

²²³ [Confinement de 2020 en France](#)

²²⁴ [Décret n° 2020-293 du 23 mars 2020 prescrivant les mesures générales nécessaires pour faire face à l'épidémie de covid-19 dans le cadre de l'état d'urgence sanitaire](#)

We have coded various policies that cancel events and large gatherings as such: any cancellations of professional sporting and other specific pre-scheduled events as the policy variable *event_cancel*. The *no_gathering* policy variable represents policy measures that banned all events or mass gatherings of a certain size, e.g. no gatherings of over 1,000 people. The *social_distance* policy variable includes measures preventing visits to elder care establishments, closures of public pools and tourist attractions, and teleworking plans for workers.

United States

For the United States' policy response to the COVID-19 pandemic, we relied on a number of sources, including the U.S. Center for Disease Control (CDC), the National Governors Association, individual state health departments, as well as various press releases from county and city-level government or media outlets. The CDC has posted and continually updated a Community Mitigation Framework that encompasses both mandatory and recommended policies at a national level.^{225, 226} The CDC's framework along with the guidelines released by the White House on March 16, 2020 named "15 Days to Slow the Spread" included a set of social distancing guidelines that urged Americans to work from home, engage in remote schooling, avoid unnecessary shopping trips, refrain from visiting restaurants and bars, and not gather in groups of more than 10 people.²²⁷ These guidelines were interpreted by individual governors as they each declared their own States of Emergency at various subsequent dates, and later released their own community mitigation plans or executive orders.

Some of the first states to release such plans included Massachusetts, California, Florida, Washington, and New York, which included both mandatory and optional policies to prevent the COVID-19 spread.²²⁸ In addition the National Governors Association has served as a resource for individual states' policies in response to COVID-19, updating each states' policy rollout timelines as well as providing links to states' Executive Orders and other official policy documentation.²²⁹ To supplement both national and state level policies and recommendations, data was collected, when possible, for cities and counties that have also taken on the role of providing guidance and implementing policies to mitigate the spread of COVID-19.

There have been a wide range in regulatory responses since the first case of COVID-19 was announced in Washington State on January 14, 2020. As a result, the CDC and the Department of State began releasing guidance to those at risk of being exposed to the virus. The initial recommendations included travel warnings for specific countries with confirmed cases and sustained COVID-19 spread. Over the course of our dataset, these warnings increased in intensity, changing from warning against inbound and outbound travel to specific countries in both Europe and Asia to warning against travel at all.²³⁰ International travel restrictions were coded as *travel_ban_int_out* for outbound travel, and

²²⁵ [What's New | COVID | CDC](#)

²²⁶ [Community Mitigation Strategies | CDC](#)

²²⁷ [15 Days to Slow the Spread](#)

²²⁸ [Schools, Workplaces & Community Locations | CDC](#)

²²⁹ [National Governors Association](#)

²³⁰ [CDC Travel Guidance & Warnings](#)

travel_ban_int_in for inbound travel, with lists of the places to and from which travel was restricted also included. On March 31, 2020, the US changed its global travel warning to Level 4, the highest warning level, which the US Department of State defines as avoiding “all international travel due to the global impact of COVID-19.”²³¹ In addition to the national travel restrictions, individual states also implemented local travel bans, coded as *travel_ban_local* as the spread continued to grow, such that anyone entering specific states in which this policy was in effect were required to self-quarantine for 14 days. This ultimately reflected the national policy as well, in that people could still technically travel under a Level 4 warning, but upon arrival to the US, they would be put in a mandatory quarantine for 14 days.

In addition to travel restrictions, as COVID-19 prevalence increased within the US, the CDC began to release additional guidance for healthcare workers, individuals at higher risk, as well as for state-level action (e.g. travel or social distancing policies),²³² and these policies have largely been implemented at the state-level rather than at the national level.

Social distancing guidelines were among the first and most widely implemented policies across states, but there exists a fair amount of variation in the specifics of each policy between states. As a result, *social_distancing* encompasses a wide range of mandatory and optional guidelines that all were focused on keeping people apart. Specifically, this policy variable was coded as a number of subcategories to capture the fact that it was implemented in a piecemeal manner within and across states.^{233, 234}

These are the policies included within each of the *social_distance* subcategories:

- *Isolate certain populations*: recommend or mandate the isolation of populations such as the elderly, immunocompromised or those who have recently returned from a cruise
- *If outside the home, must abide by social distancing standards*: require a six foot minimum distance from others outside the home, maintain distance when riding public transportation, ask that businesses restrict the number of people within storefront at a time as well as restricting certain types of activities that involve physical interaction with customers (e.g., bagging groceries, taking cash payment)
- *Mandate mask wearing*: require people to wear a mask outside the home
- *Close public facilities*: close libraries, museums, flea markets, historic sites, memorials, and polling locations
- *Close outdoor facilities*: close beaches, state parks, public parks, public toilets, lakes, and campgrounds.
- *Social distance restriction of visitation to certain facilities*: restrict visitation to prisons, long term care facilities, child care facilities, and homeless shelters, stop elective medical and veterinary procedures, and bar short term rental accommodations.

²³¹ [US Travel Warnings](#)

²³² [CDC COVID-19 Guidance Documents](#)

²³³ [NYT Article | "Wondering About Social Distancing?"](#)

²³⁴ [COVID-19 | Get Your Mass Gatherings or Large Community Events Ready for Coronavirus Disease 2019](#)

- *Suspend non-critical state operations/government services*: close government buildings, stop in person meetings of people working for the state, suspend court operations, waive or extend licensing, and permit certain types of work to be carried out remotely, when normally could not (e.g.; notaries, police work, licensing)

Along the same line of social distancing policies, a separate variable was coded as *no_gathering* to represent policy measures that banned all events or mass gatherings of a certain size, i.e. no gatherings over a certain number of people (where this number has varied by region).

In addition, many governors mandated statewide school closures at the private and public K-12 and higher education levels, while others have left it up to each school district to decide.²³⁵ School closures have been coded as *school_closure* and once implemented, have been “turned on” for the remainder of our time series, as no schools have reopened since these policies have been implemented.

Business closures, coded as *business_closure*, have also been recommended or mandated at the state level. These policies have ranged from shutting down all non-essential businesses, reducing the number of hours a business may be in operation, severely restricting the number of customers that are allowed inside at one time, to prohibiting customers to enter a business, such as in the case of bars and restaurants, where they were only allowed to operate or take-out and delivery services. When business closures have involved shutting down all non-essential operations, “essential” has been defined by each state but is largely similar between states, generally defining essential as food or healthcare providers, as well as basic government operations (i.e. trash collection, mail, water monitoring, etc). To support employees working remotely or staying home when sick, a number of states have also mandated paid sick leave for those who are affected by COVID-19, which has been coded as *paid_sick_leave*. There is a separate *work_from_home* category that includes measures that require businesses to allow employees to telework, if possible, such that no workers except for those who have essential functions are allowed to work in an office.

At the subnational level, many governors have implemented a statewide mandatory shelter-in-place policy, requiring all individuals to self-isolate within their home or place of residence and limit outdoor activity to essential functions only, which is defined by each state. Shelter-in-place laws have been coded as *home_isolation*, which implies all of the policies enacted alongside shelter-in-place policies, including work from home, business closures, and no gathering policies.

Both the *emergency_declaration* and *event_cancel* variables were included in the raw data, but were not included in the final analysis.

Population Data

In order to construct population weighted policy variables and to determine the susceptible fraction of the population for disease projections under the realized and the no-policy counterfactual scenarios, we obtained the most recent estimates of population for each administrative unit included in our analysis. The sources of that population data are documented below.

²³⁵ [US School Closures due to COVID-19](#)

China

City-level population data have been extracted from a compiled dataset of the 2010 Chinese City Statistical Yearbooks. We matched the city level population dataset to the city level COVID-19 epidemiology dataset. As the two datasets use slightly different administrative divisions, we only matched 295 cities that exist in both datasets, and grouped the remaining 39 cities in our compiled epidemiology dataset into "other" for prediction purposes. Cities grouped into "other" because of mismatches have a total population of 114,000,000, or 8.5% of the total population in China.

For these 39 cities that we could not match, we imputed the population by taking the total remaining population (114,000,000) and divided it evenly between these remaining cities. We flag the imputed populations by using the binary variable *pop_is_imputed*.

South Korea

We downloaded the number of population by provinces from a webpage administered by the Korean Statistical Information Service (KOSIS).²³⁶ The government agency recently updated the population information of February, 2020, which we used for our analysis.

Italy

Region and province level population data come from the Italian National Institute of Statistics (Istat), estimating total population on January 1, 2019. The datasets for all Italian regions and provinces are scraped from [Istat's website](#) in *get_adm_info.ipynb*.

Iran

Province level population data for Iran comes from the 2016 Census, as listed on the *City Population* website.²³⁷ It is scraped in *get_adm_info.ipynb*.

France

Département-level populations are obtained from the National Institute of Statistics and Economic database.²³⁸ We used the most up to date estimation of the population in France as of January 2020.

United States

State- and county-level population data come from the 2017 American Community Surveys dataset, and is downloaded via the *census* Python package²³⁹ in *get_adm_info.ipynb*.

²³⁶ [행정구역\(시군구\)별, 성별 인구수 \(Population by county and gender\)](#)

²³⁷ [Iran: Administrative Division \(Provinces and Counties\) - Population Statistics, Charts and Map](#)

²³⁸ [Population de 1999 à 2020](#)

²³⁹ [datamade/census: A Python wrapper for the US Census API.](#)

Supplementary Methods

In our Supplementary Methods, we describe several sensitivity analyses performed to assess the robustness of our growth rate impacts and projections of cases averted/delayed.

This section is divided into five analyses:

1. Testing the sensitivity of projected averted/delayed cases to varying epidemiological parameters
2. Testing the sensitivity of our regression model to varying epidemiological parameters
3. Testing the sensitivity of our regression model to changes in lag structure
4. Testing the sensitivity of our regression model to withholding of data
5. Testing the sensitivity of our regression model to changes in policy groupings

1. Sensitivity of projected averted/delayed cases to the removal rate γ , the use of an SEIR framework, and the infection-fatality ratio

We compute the empirical removal rate using aggregated data from the countries for which we observe active cases (i.e., China and South Korea) and estimate a value of $\gamma = 0.079$ (see the Methods section of the main text). This value measures the inverse of the mean duration from being *reported* as infected to being *reported* as recovered (or dead) and may differ from the fundamental epidemiological parameter describing the rate of removal from the pool of infectious individuals. While our estimate implies a recovery period (symptom onset to recovery) that is comparable to some estimates in the literature (median time of 19-23 days, varying based on age group, sex, severity, and mode of detection²⁴⁰), we test the extent to which our simulation results in Figure 4 depend on this value (Extended Data Figure 7). One motivation for this exercise is that there may be an unknown delay between the time when a patient becomes non-infectious and the time in which they are recorded in the aggregate data as recovered. For example, assuming a conservatively high value of 14 days for the average delay between “no longer infectious” and “confirmed recovered,” our empirical estimation yields $\gamma = 0.18$.

In addition, the use of an SIR framework may misrepresent the true underlying disease dynamics, and a more general SEIR framework, which includes representation of people exposed to the infection without yet being infectious, may produce more realistic simulations of cases averted/delayed by policy. We also test sensitivity to the use of the SEIR framework, as well as a key parameter in this alternative framework -- the assumed rate of transition from exposed to infectious (σ).

We replicate the simulation underlying Figure 4 using an SEIR framework with values of $\gamma = \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$ and $\sigma \in \{0.2, 0.33, 0.5, \infty\}$, with $\sigma = \infty$ corresponding to the SIR framework we employ in the main simulation. We present our estimates of the total number of cases

²⁴⁰ [Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts](#)

under both the no-policy and policy scenarios, as well as the total number of cases averted/delayed by policy. We sum simulated cases across all countries on the last dates of the countries' respective samples. Note that the simulation uses the growth rates derived from our empirical model such that changes in γ and σ correspond to changes in the transmission rate β . β must vary with γ and σ as our data determine the underlying exponential growth rate. We show the results of this sensitivity analysis in Extended Data Figure 7.

Panels (a) and (b) respectively show the simulated number of cases in the no-policy and with-policy scenarios. The number of simulated no-policy cases is decreasing in γ for high σ and increasing in γ for low σ . The number of simulated no-policy cases is increasing in σ for low γ and nonmonotonic in σ for high γ . The number of simulated with-policy cases is increasing in γ and decreasing in σ . Panel (c) shows the number of cases averted due to policy and demonstrates that varying σ or γ can reduce our estimate of cases averted on the order of several million reported cases (up to 10%). Panel (c) shows that higher values of γ produce lower estimates of cases averted for the SIR model ($\sigma = \infty$), but increasing estimates of cases averted for the lower values of σ . Panel (d) plots the content of Panel (c) on the log scale used in Figure 4 of the main text for comparison. For our simulation value of γ , decreasing σ decreases our estimate of cases averted.

Overall, these results show that the order of magnitude of the number of cases averted is preserved within this reasonable range of potential γ and σ values.

We also compute country-specific estimates of infection underreporting to improve the outbreak simulations in Figure 4. To produce these estimates, we use code from Russell et al. and substitute the assumed case-fatality ratio (a key parameter in their model) for an infection-fatality ratio (IFR) of 0.75%.^{241,242} Their analysis produces country-specific estimates of infection underreporting, which we use to scale our estimates of confirmed cases to estimate total infections. We test the sensitivity of our main results to this IFR assumption in Supplementary Table 6, where we also show results for the total number of confirmed cases and infections avoided/delayed using IFR assumptions of 0.5% and 1%. The table shows approximately a 70% increase in the estimated number of confirmed cases avoided/delayed for a doubling in the IFR. However, the estimated number of infections is relatively stable with approximately a 15% decline in the estimated number of infections for a doubling in the IFR.

2. Sensitivity of exponential regression model to varying epidemiological parameters

The model we use to estimate the impacts of policy on growth rates assumes exponential growth, which is typically valid for early-stage disease outbreaks. If growth is not exponential, there exists the potential for bias in estimated coefficients. There are three primary reasons why an early-stage outbreak could exhibit non-exponential growth in the absence of policy intervention:

²⁴¹ [Using a delay-adjusted case fatality ratio to estimate under-reporting](#)

²⁴² [A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates](#)

1. The infection spread may progress quickly, lowering the susceptible fraction of the population to a degree that materially affects the infection spread, transitioning the outbreak away from the exponential “early stage” regime.
2. In a disease with a substantial latent period, the growth of infections is only asymptotically exponential.²⁴³ At any given moment in time, the instantaneous growth rate may differ from a steady-state exponential growth rate.
3. When analyzing growth in cumulative infections, as we do for countries where active infection data are unavailable, growth is similarly only asymptotically exponential.

In our dataset, 95% of administrative units have susceptible fractions above 0.93 on their last analysis day and all have susceptible fractions above 0.78, indicating that the first reason is unlikely to induce substantial bias in our results. When the transmission rate of a disease declines due to anti-contagion policy, the growth rate in infections decreases with a lag due to the dynamics associated with the latent period. Because of this, exponential models estimating the average treatment effect (ATE) of a policy may underestimate the true reduction seen from a policy because they include days in which the growth rate was still higher than the new steady state growth rate. Finally, in the early stages of an outbreak, the number of active cases will dominate the number of recovered/deceased patients and thus the differences in growth of active and cumulative cases is likely to be small.

To test the robustness of our regression approach, we construct simulated outbreaks in which we control demographic, policy, and epidemiological parameters. We then use a variant of the regression model (Eq. 7) from the main text to estimate the no-policy growth rates and the effects of each policy. In this simulation, we do not include any fixed effects to control for day-of-week (δ) and changing testing regime (μ) effects. These variables are not simulated as these are primarily measurement controls and their effects would be directly absorbed by the corresponding regression parameters if simulated. We compare the coefficient estimates to the “true” values used in the simulation.

To capture the impact of disease latency, we use an SEIR model framework to generate synthetic outbreaks. We simulate 12,000 45-day outbreaks at hourly timesteps, starting with a no-policy exponential growth rate of 0.4 (similar to those estimated in our main analysis) and a single exposed individual. We adjust this asymptotic exponential growth rate to account for three synthetic policies that turn on at random times, each with a known effect (-0.05, -0.1, and -0.2). For each subset of 1,000 simulations, we use one of four plausible values for the mean latency period, σ^{-1} (0, 2, 3, and 5 days), and one of three plausible values for the mean infectious period, γ^{-1} (3, 5, and 20 days). We choose a wide range of these variables due to substantial uncertainty over the epidemiological characteristics of the novel coronavirus,²⁴⁴ and a nonexistent latency period is included for comparison to an SIR-like data generating process. The mean transmission rate per infected person per day, β , is derived from the asymptotic growth rate, the mean latency period, and the mean infection period by solving for the eigenvalues of a SEIR system,²⁴⁵ which yields:

²⁴³ [Estimating epidemic exponential growth rate and basic reproduction number](#)

²⁴⁴ [MIDAS Network Online COVID-19 Portal: Parameter Estimates](#)

²⁴⁵ [Estimating epidemic exponential growth rate and basic reproduction number](#)

$$\beta = \frac{(g+\gamma)(g+\sigma)}{\sigma}.$$

We apply exponential noise to β for each simulation and at each timestep, and contribute additional gaussian noise to σ and γ (standard deviations of 0.01 and 0.03, respectively). We add additional gaussian “measurement noise” to the instantaneous growth rates after simulation but before running our regression (standard deviation of 0.1). Cumulatively, this results in an average root-mean-squared-error (RMSE) in regressions across all 12,000 simulations of ~ 0.11 , which matches the RMSE of daily growth rate values across all six countries in our main analysis.

The dynamic model outputs a time series of susceptible (S), exposed (E), infectious (I), and removed (R) individuals, as a fraction of the total population. We use both I and I+R as the left-hand-side variables in our regression framework. The former corresponds to the analysis we run for countries in which we observe active cases and the latter to countries in which we observe only cumulative cases. We assume the majority of the “exposed but not infectious” population will not yet have been tested and will not appear in any of the datasets used in the main analysis. Our right-hand-side variables consist of the binary policy variables, allowing only for contemporaneous effects. This matches our main specification for all countries except China (where data availability allows for the estimate of lagged effects) and provides the most challenging environment in which to estimate the effect of policy in a dynamic system.

Results are presented in Extended Data Figures 8 and 9. While it is possible to simulate outbreaks consisting of innumerable parameter combinations and noise distributions, we display those that seem most relevant for evaluating the robustness of our main analysis. Our associated GitHub repository contains a Jupyter notebook for readers to further examine the effect of simulation configurations on regression model robustness.

Figures 8 and 9 are each split into two panels (a) and (b). Panel (a) of each figure shows simulations in near-ideal data conditions, in which we observe active infections within a large population. This means that the susceptible fraction of the population remains high during the entire sample period. For example, these conditions are similar to those in our real data for Chongqing, China. Panel (b) of each figure shows simulations in a non-ideal data scenario where we are only able to observe cumulative infections in a small population. In these simulations, the susceptible fraction declines to values as low as 33% of the population. For example, these conditions are similar to those in our real sample of data for Cremona, Italy.

Figure 8 demonstrates that our model recovers unbiased estimates of the no-policy growth rate under all conditions simulated. Because the growth rate prior to policy has likely approached its asymptotic rate by the time we begin our regressions, variance in our no-policy growth rate estimates comes from noise in the disease parameters and measurement. The ability to recover unbiased estimates of this value has important implications for our estimate of the total number of cases averted/delayed to date, as this number is primarily driven by the counterfactual number of cases we would expect to see in a world in which no anti-contagion policy was enacted.

Figure 9 demonstrates that our model recovers unbiased estimates of the cumulative effect for a disease with very short latency. As the latency period increases, the model begins to slightly underestimate the true effect of policy (i.e. it predicts a less negative value), due to the decay time over which a shock to transmission rate propagates to a new steady-state growth rate. The underestimate is reduced in situations where we are able to directly observe active infections and is increased when we can only observe cumulative infections. Note that statistical uncertainty in these estimated parameters dominates potential biases, even in “worst case” data conditions.

We conclude that biases (due to the use of an exponential model) in our estimates of the no-policy growth rate are essentially zero and are likely to be small and negative for our estimates of policy effectiveness. If present in the data, such biases would cause us to modestly understate the effectiveness of anti-contagion policies.

3. Sensitivity of estimates to changes in lag structure

Existing evidence has not demonstrated whether policies should affect infection growth rates in the days immediately following deployment. It is therefore not clear *ex ante* whether the policy variables in Equation 7 should be encoded as “on” immediately following a policy deployment. As a robustness check, we estimate “fixed-lag” models in which a fixed delay between a policy’s deployment and its effect is assumed. Specifically, we assume that policies cannot influence infection growth rates for L days, recoding a policy variable at time t as zero if a policy was implemented fewer than L days before t . We re-estimate Equation (7) for each value of L and present results in Extended Data Figure 5 and Supplementary Table 5. If a delay model is more consistent with real world infection dynamics, these fixed lag models should recover larger estimates for the impact of policies and exhibit better model fit.

Panel a of Extended Data Figure 5 displays the R^2 associated with each country-level fixed lag model with fixed lag lengths ranging from no fixed lags up to a 15 day fixed lag. In-sample fit generally declines or remains unchanged if policies are assumed to have a delay longer than 4 days. Panel b of Extended Data Figure 5 plots the estimated effects for no lag (the model reported in the main text) and for fixed-lags between one and five days. Estimates generally are unchanged or shrink towards zero (e.g. *home_isolation* in Iran), consistent with miscoding of post-policy days as no-policy days.

In Supplementary Table 5, we show our estimates of the effect of all policy interventions in each country (analogous to the average difference between red and blue markers in Figure 3 of the main text) using a fixed lag of up to 5 days. The estimated effects are broadly consistent across different lag lengths; however, the magnitude of the effect size declines slightly with increasing lag lengths. If policies take several days to impact infection growth rates, we would expect effect sizes to increase rather than decrease with lag lengths. Our finding of declining effect sizes is more consistent with contamination of the control group, where policies are incorrectly encoded as zeros after they have been deployed.

4. Sensitivity of estimates to withholding of data

To ensure that the estimates from our regression model are robust to the withholding of data, we re-estimate our main model k_c number of times for each country, where k_c is the number of first level administrative units (Adm1, i.e. state or province) in country c . In each of the k_c regressions for country c , we withhold data from one Adm1 unit when we estimate the effects of policy interventions on growth rates. The results of this exercise are displayed in Extended Data Figures 3 and 4.

5. Sensitivity of estimated growth rates to changes in policy groupings

In our main regression model, due to the limited length of our time series data and instances where multiple policies are deployed on the same date, we group certain policy interventions together. We group policies together that have similar objectives (e.g. *travel_ban_local* and *transit_suspension* would be one group, *event_cancel* and *no_gathering* would be another) and keep certain policies separate (i.e. *business_closure*, *school_closure*, *home_isolation*) where possible.

To test the sensitivity of our results to the grouping of policy interventions, we also estimate a model where the policies are estimated without grouping. Extended Data Figure 6 panel a shows the estimated infection growth rates and no-policy counterfactual growth rates using the model with disaggregated policies. Additionally, in Supplementary Table 4, we compare the effect of policy interventions in each country when the effect of all policies are estimated separately (“Disaggregated Model”) and when they are grouped into policy packages as in our preferred specification (“Main Specification”). We find the estimated impact of policy interventions on case growth rates is robust to this alternative specification.

Supplementary Tables

Supplementary Table 1 | Number of unique anti-contagion policies in this study tabulated by administrative divisions of each country.

Country	Adm0	Adm1	Adm2	Adm3	Total
China	0	35	133	0	168
South Korea	20	39	0	0	59
Italy	14	29	94	77	214
Iran	6	17	0	0	23
France	8	1	50	0	59
United States	43	694	426	31	1194
Total	91	815	703	108	1717

This table reports the policy data that have been collected at various levels of administrative divisions in each country. *Adm0* represents the country level, and higher *Adm** numbers indicate smaller administrative subdivisions, which are specific to each country. Each policy is counted at the highest resolution administrative unit in which the policy is applied differentially. For example, if a country has ten regions at the *Adm1* level, and a policy is applied across five of those regions, the policy is counted as five separate *Adm1* policies rather than a single *Adm0* policy. National policies are counted only once as *Adm0*.

Supplementary Table 2 | Estimates of no-policy infection growth rates and case doubling time using different samples from Wuhan, China.

(1)	(2)	(3)	(4)	(5)	(6)
Date (used as start of sample)	Raw case data on date in (1) Source: Wu et al. (2020) ²⁴⁶ Cumulative confirmed cases	Raw case data on date in (1) Active cases	Estimates using 1/22/20 as sample end date Estimated infection growth rate	Estimates using 1/22/20 as sample end date Case doubling time (days)	Notes
12/10/19	1	1	0.12	5.69	← Wu et al. (2020) and Li et al. (2020) ²⁴⁷ sample begins
12/11/19	1	1	0.12	5.56	
12/12/19	1	1	0.13	5.42	
12/13/19	1	1	0.13	5.29	
12/14/19	2	2	0.12	5.96	
12/15/19	4	4	0.10	6.88	
12/16/19	6	6	0.09	7.51	
12/17/19	6	6	0.09	7.30	
12/18/19	6	6	0.10	7.10	
12/19/19	6	6	0.10	6.89	
12/20/19	7	7	0.10	7.00	
12/21/19	8	8	0.10	7.08	
12/22/19	10	10	0.09	7.40	
12/23/19	11	11	0.09	7.40	
12/24/19	11	11	0.10	7.15	
12/25/19	13	13	0.09	7.34	
12/26/19	13	13	0.10	7.07	
12/27/19	13	13	0.10	6.80	
12/28/19	14	14	0.10	6.72	
12/29/19	18	18	0.10	7.17	
12/30/19	21	21	0.09	7.37	
12/31/19	22	22	0.10	7.20	
1/1/20	28	28	0.09	7.78	
1/2/20	34	34	0.08	8.30	
1/3/20	44	44	0.07	9.39	
1/4/20	44	44	0.08	8.87	← Li et al. (2020) sample ends
1/5/20	59	59	0.06	10.71	
1/6/20	59	59	0.07	10.04	
1/7/20	59	59	0.07	9.37	
1/8/20	59	59	0.08	8.70	
1/9/20	41	41	0.12	5.94	← Official cumulative cases decreases
1/10/20	41	38	0.13	5.17	← No new official cases
1/11/20	41	34	0.16	4.37	← No new official cases
1/12/20	41	33	0.18	3.86	← No new official cases
1/13/20	41	33	0.20	3.43	← No new official cases
1/14/20	41	33	0.23	3.00	← No new official cases
1/15/20	41	27	0.30	2.29	← National standards for diagnosis issued
1/16/20	45	28	0.36	1.95	← Testing available in Hubei
1/17/20	62	41	0.35	1.98	Our sample begins (see the Supplementary Notes on Epi. Data in China)
1/18/20	121	94	0.35	1.98	
1/19/20	198	170	0.27	2.58	
1/20/20	258	228	0.26	2.70	
1/21/20	365	329	0.15	4.72	
1/22/20	425	381	-	-	← Wu et al. (2020) and our sample for estimating no-policy growth rate end

This table reports the raw epidemiological data in Wuhan, China and our estimates of infection growth rates and case doubling times prior to the city-wide lockdown on January 22, 2020, utilizing different sample periods to estimate these values. Column (1) provides the dates of observations. Columns (2) and (3) show raw official epidemiological data: the number of cumulative cases and the number of active cases (= *cumulative cases* - *recoveries* - *deaths*), respectively.

²⁴⁶Wu, J.T., Leung, K., Bushman, M. et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nature Medicine* 26, 506–510 (2020). <https://doi.org/10.1038/s41591-020-0822-7>

²⁴⁷Li, Qun, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren et al. "Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia." *New England Journal of Medicine* (2020).

We show data from Wu et al. (2020) which match our data in the overlapping periods; in the interest of space, we do not display raw data used in Li et al. (2020), but they are similar with small differences. Columns (4) and (5), report estimates of pre-policy infection growth rates and case doubling times (in days) using different start dates, respectively. Specifically, we set the sample start date as the date in column (1) and the end date of the sample (used for estimating our no-policy growth rate) fixed at January 22, 2020 (the day prior to city-wide lockdown), using the time series data between these two dates to estimate the parameters in columns (4) and (5). Column (6) provides notes on these data. Using a start date of 12/10/19, the same as Wu et al (2020), we obtain an estimated case doubling time of 5.7 days, similar to their estimated 5.2 days (they use an alternative structural modeling approach). As described in our Supplementary Notes on Epidemiological Data in China, we only use data in China beginning 1/16/20 because the first national guidelines for diagnosis were issued on 1/15/20. Prior to that date, there did not exist a consistent case definition to identify the earliest 41 confirmed cases in Wuhan.²⁴⁸ Additionally, the documented lack of testing capacity in the province of Hubei before 1/16/20 raises concern about data quality during that time period.²⁴⁹ These concerns about data quality appear consistent with irregularities in the official record of cumulative cases (column 2). For example, official cumulative cases decreased on 1/9/20, which should not be possible. Additionally, no new cases were reported between 1/9/20 and 1/15/20, when at least roughly five new cases per day should have been reported if case doubling time actually was 5.2 days. The reliability of these official reports during the 1/9/20-1/15/20 period has been called into question, with news sources suggesting that people who were likely to have been infected by COVID-19 in that period of time (and deaths attributed to the disease) were not counted in the official tally.^{250,251} These data quality concerns motivate our use of the 1/16/20 start data for our sample in China, which provide an infection growth rate of 0.36 and case doubling time of 1.95 days (gray) using only the time series in Wuhan. These estimates are broadly consistent with our estimates from all other countries we examine, except Iran, and the global average growth rate of 0.37 we estimate (see Figure 2A in the main text).

²⁴⁸ [Impact of changing case definitions for COVID-19 on the epidemic curve and transmission parameters in mainland China Affiliation](#)

²⁴⁹ [武汉肺炎：疫情从可控到失控的三十天 \(Wuhan pneumonia: 30 days from outbreak to out of control\)](#)

²⁵⁰ [Warning against cover-up as China virus cases jump](#)

²⁵¹ [As families tell of pneumonia-like deaths in Wuhan, some wonder if China virus count is too low](#)

Supplementary Table 3 | Country level regression results estimating the effect of policy interventions on COVID-19 infection growth rates.

a. China

Dependent variable: growth rate of active cases ($\Delta \log$ cases per day)

Variables	Coefficient	Std Error	P-value		Coefficient	Std Error	P-value	
<i>Policy variables:</i>								
Emergency declaration, Week 1	-0.012	0.043	0.792		Week 1	-0.026	0.046	0.581
Emergency declaration, Week 2	-0.173***	0.046	0.000		Week 2	-0.204***	0.049	0.000
Emergency declaration, Week 3	-0.234***	0.047	0.000		Week 3	-0.279***	0.047	0.000
Emergency declaration, Week 4	-0.250***	0.047	0.000		Week 4	-0.302***	0.046	0.000
Emergency declaration, Week 5	-0.248***	0.047	0.000		Week 5	-0.292***	0.047	0.000
Travel ban, Week 1	0.015	0.016	0.338					
Travel ban, Week 2	-0.009	0.018	0.618					
Travel ban, Week 3	-0.031	0.024	0.191					
Travel ban, Week 4	-0.048	0.029	0.107					
Travel ban, Week 5	-0.079**	0.038	0.043					
Home isolation, Week 1	-0.029**	0.014	0.041					
Home isolation, Week 2	-0.021	0.017	0.223					
Home isolation, Week 3	-0.014	0.021	0.512					
Home isolation, Week 4	-0.003	0.027	0.897					
Home isolation, Week 5	0.035	0.037	0.355					
<i>Testing regime dummy variables:</i>								
Testing regime change on Jan 18, 2020	0.488***	0.046	0.000					
Testing regime change on Jan 28, 2020	0.065***	0.015	0.000					
Testing regime change on Feb 06, 2020	-0.039***	0.012	0.002					
Testing regime change on Feb 13, 2020	-0.002	0.002	0.356					
Testing regime change on Feb 20, 2020	-0.005**	0.002	0.015					
Testing regime change on Mar 05, 2020	-0.006	0.005	0.237					
Observations	3,669							
R-squared	0.60							

*** p<0.01, ** p<0.05, * p<0.1

This regression includes city fixed effects and clustered standard errors at the day level. P-values, are from the two-sided test with the null hypothesis that the coefficient is equal to zero. The sample includes 3,669 observations at the city-day level.

In China, we encode home isolation to imply a local travel ban. The estimate reported for home isolation in Figure 2 is the combined effect of both of these policies. Here, we report the incremental effect of the home isolation order separate from the effect of all policies it implies.

b. South KoreaDependent variable: growth rate of active cases ($\Delta \log$ cases per day)

Variables	Coefficient	Std Error	P-value
<i>Policy variables:</i>			
Religious closure, welfare service facilities closure, no demonstration	-0.304**	0.151	0.049
Work from home, business closure, other social distance (optional)	-0.083**	0.038	0.032
Emergency declaration	-0.127***	0.046	0.009
Quarantine inbound travelers	0.020	0.019	0.295
<i>Testing regime dummy variables:</i>			
Testing regime change on Feb 20, 2020	0.074	0.087	0.398
Testing regime change on Feb 29, 2020	0.040**	0.019	0.044
Testing regime change on Mar 22, 2020	-0.010	0.023	0.655
Testing regime change on Mar 27, 2020	-0.039	0.038	0.309
Observations	595		
R-squared	0.28		

*** p<0.01, ** p<0.05, * p<0.1

This regression includes province fixed effects, day-of-week fixed effects, and clustered standard errors at the day level. P-values are from the two-sided test with the null hypothesis that the coefficient is equal to zero. The sample includes 595 observations at the province-day level.

"Work from home, business closure, other social distance (optional)" includes recommended policies related to social distancing, e.g. limit gatherings, work from home, and closing businesses such as karaoke and cyber cafes.

This regression does not include school closures and event cancellation policies, which were implemented in all provinces before our sample period starting February 17, 2020.

Effect of all policies combined

Coefficient	Std Error	P-value
-0.494***	0.150	0.002

c. ItalyDependent variable: growth rate of cumulative confirmed cases ($\Delta \log$ cases per day)

Variables	Coefficient	Std Error	P-value
<i>Policy variables:</i>			
School closure	-0.109	0.070	0.127
Quarantine positive cases	-0.084	0.055	0.134
Work from home, no gathering, other social distance	0.144**	0.068	0.040
Travel ban, transit suspension	-0.331**	0.137	0.021
Business closure	-0.121	0.098	0.225
Home isolation	0.048	0.051	0.360
Observations	2,898		
R-squared	0.29		

*** p<0.01, ** p<0.05, * p<0.1

This regression includes province fixed effects, day-of-week fixed effects, and clustered standard errors at the day level. P-values are from the two-sided test with the null hypothesis that the coefficient is equal to zero. The sample includes 2,898 observations at the province-day level.

"Work from home, no gathering, other social distance" includes policies for working from home, maintaining one meter distance from others in public, and prohibiting public and private events.

In Italy, we encode home isolation to imply a local travel ban, business closures, work from home, no gathering, and other social distancing measures. The estimate reported for home isolation in Figure 2 is the combined effect of these policies. Here, we report the incremental effect of the home isolation order separate from the effect of all policies it implies.

Effect of all policies combined

Coefficient	Std Error	P-value
-0.453***	0.105	0.000

d. IranDependent variable: growth rate of cumulative confirmed cases ($\Delta \log$ cases per day)

Variables	Coefficient	Std Error	P-value
<i>Policy variables:</i>			
Travel ban (opt), work from home, school closure	-0.334***	0.089	0.001
Home isolation	-0.154***	0.033	0.000
<i>Testing regime dummy variable:</i>			
Testing regime change on Mar 13, 2020	-0.075	0.096	0.442
Observations	548		
R-squared	0.29		

*** p<0.01, ** p<0.05, * p<0.1

This regression includes province fixed effects, day-of-week fixed effects, and clustered standard errors at the day level. P-values are from the two-sided test with the null hypothesis that the coefficient is equal to zero. The sample includes 548 observations at the province-day level.

"Travel ban (opt), work from home, school closure" policies were the national policies enacted March 1-5, 2020, which overlap with missing provincial case data in Iran on March 2-3, 2020.

This regression does not include a national no gathering policy, provincial school closures, and an optional travel ban in Qom, which were all implemented in the respective provinces before our sample period starting February 27, 2020. This regression does not include a religious closure policy in Qom enacted on March 17, 2020, as Iran stopped reporting provincial case data on March 22, 2020.

Effect of all policies combined

Coefficient	Std Error	P-value
-0.487***	0.076	0.000

e. FranceDependent variable: growth rate of cumulative confirmed cases ($\Delta \log$ cases per day)

Variables	Coefficient	Std Error	P-value
<i>Policy variables:</i>			
School closure	-0.009	0.042	0.826
Cancel events, no gathering, other social distance	-0.243***	0.072	0.003
Business closure, home isolation	0.083	0.052	0.127
<i>Testing regime dummy variable:</i>			
Testing regime change on Mar 15, 2020	-0.112***	0.021	0.000
Observations	270		
R-squared	0.30		

*** p<0.01, ** p<0.05, * p<0.1

This regression includes region fixed effects, day-of-week fixed effects, and clustered standard errors at the day level. P-values are from the two-sided test with the null hypothesis that the coefficient is equal to zero. The sample includes 270 observations at the region-day level.

"Cancel events, no gathering, other social distance" policies include banning audiences at professional sport events, prohibiting gatherings of a certain size, and limiting visits to long term care facilities.

In France, we encode home isolation to imply event cancellations, no gathering, and social distancing measures. The estimate reported for the national lockdown in Figure 2 is the combined effect of these policies. Here, we report the incremental effect of the national lockdown separate from the effect of all policies it implies.

Effect of all policies combined

Coefficient	Std Error	P-value
-0.169***	0.020	0.000

f. United StatesDependent variable: growth rate of cumulative confirmed cases ($\Delta \log$ cases per day)

Variables	Coefficient	Std Error	P-value
<i>Policy variables:</i>			
Slow the Spread Guidelines (Mar 16 federal)	0.049	0.037	0.193
Other social distance	-0.246***	0.044	0.000
Paid sick leave	-0.027	0.087	0.761
Quarantine positive cases	-0.061*	0.031	0.058
Travel ban, transit suspension	-0.008	0.033	0.802
School closure	0.026	0.031	0.407
Religious closure	0.008	0.020	0.713
Work from home	-0.047***	0.017	0.009
No gathering	0.011	0.035	0.750
Business closure	-0.055**	0.021	0.014
Home isolation	-0.030*	0.017	0.096
<i>Testing regime dummy variables:</i>			
Testing regime change on Mar 13, 2020 in NY	-0.137***	0.043	0.003
Testing regime change on Mar 16, 2020 in CA	-0.021	0.024	0.388
Testing regime change on Mar 18, 2020 in NC	0.169***	0.042	0.000
Testing regime change on Mar 19, 2020 in CT	0.106**	0.040	0.012
Testing regime change on Mar 19, 2020 in NV	0.241***	0.039	0.000
Testing regime change on Mar 19, 2020 in UT	-0.162***	0.038	0.000
Testing regime change on Mar 20, 2020 in IA	-0.211***	0.024	0.000
Testing regime change on Mar 21, 2020 in TN	0.203***	0.037	0.000
Testing regime change on Mar 22, 2020 in AL	-0.001	0.027	0.966
Testing regime change on Mar 23, 2020 in HI	-0.034	0.024	0.162
Testing regime change on Mar 24, 2020 in KS	-0.041**	0.020	0.046
Testing regime change on Mar 24, 2020 in NJ	0.022	0.026	0.406
Testing regime change on Mar 25, 2020 in OH	0.010	0.027	0.717
Testing regime change on Mar 27, 2020 in AZ	0.047	0.033	0.162
Testing regime change on Mar 28, 2020 in MD	0.006	0.029	0.833
Testing regime change on Mar 28, 2020 in MO	-0.034	0.051	0.502
Testing regime change on Mar 30, 2020 in DE	-0.008	0.033	0.808
Observations	1,238		
R-squared	0.29		

*** p<0.01, ** p<0.05, * p<0.1

This regression includes state fixed effects, day-of-week fixed effects, and clustered standard errors at the day level. P-values are from the two-sided test with the null hypothesis that the coefficient is equal to zero. The sample includes 1,238 observations at the state-day level.

"Other social distance" includes policies such as closing libraries, maintaining 6 feet distance from others in public, and limiting visits to long term care facilities.

In the United States, we encode home isolation to imply work from home, no gathering, and closure of non-essential businesses. The estimate reported for home isolation in Figure 2 is the combined effect of these policies. Here, we report the incremental effect of the home isolation order separate from the effect of all policies it implies.

Effect of all policies combined

Coefficient	Std Error	P-value
-0.380***	0.095	0.000

These regression tables **a-f** display the results from our main model estimating the effect of policy on daily COVID-19 infection growth rates in China, South Korea, Italy, Iran, France, and the United States. The regression model is estimated separately for each country, allowing for the policy type to have different average treatment effects for each country.

Supplementary Table 4 | Estimated Effect of Actual Policies on Infection Growth Rates With and Without Grouping Policies

Country	Main Specification (policies grouped)		Disaggregated Model (policies separate)	
	Effect size	95% CI	Effect size	95% CI
China	-0.252	(-.339, -.164)	-0.252	(-.339, -.164)
South Korea	-0.248	(-.423, -.073)	-0.21	(-.377, -.044)
Italy	-0.24	(-.372, -.107)	-0.237	(-.369, -.104)
Iran	-0.355	(-.48, -.231)	-0.291	(-.437, -.146)
France	-0.123	(-.16, -.086)	-0.123	(-.176, -.07)
United States	-0.084	(-.144, -.025)	-0.084	(-.144, -.024)

Average effect of all policies on infection growth rates in our sample (analogous to the average difference between red and blue markers in Figure 3 of the main text) using our main specification, where the impact of grouped policy interventions are estimated jointly, compared to a disaggregated model where the effect of all policies are estimated simultaneously but without grouping them (See Supplementary Methods 5). In our main specification, policies are grouped due to limited data. We group policies that have similar objectives (e.g. *travel_ban_local* and *transit_suspension*) or are deployed very close in time. We keep other policies separate where possible (e.g. *school_closure* and *home_isolation*). In the disaggregated model, no policies are grouped and individual policies estimates are noisier, however combined effects are similar. The sample size is 3,669 in China, 595 in South Korea, 2,898 in Italy, 548 in Iran, 270 in France, and 1,238 in the United States.

Supplementary Table 5 | Estimated Effect of Actual Policies Combined on Infection Growth Rates Using Alternative Models that Assume Lagged Effects Of Policies

Country	Statistic	0 Day Lag	1 Day Lag	2 Day Lag	3 Day Lag	4 Day Lag	5 Day Lag
China	Effect size	-0.252	-0.239	-0.242	-0.275	-0.268	-0.269
	95% CI	(-.339, -.164)	(-.313, -.166)	(-.299, -.184)	(-.349, -.2)	(-.334, -.202)	(-.325, -.214)
	R ²	0.597	0.581	0.563	0.559	0.562	0.552
South Korea	Effect size	-0.248	-0.246	-0.23	-0.221	-0.197	-0.171
	95% CI	(-.423, -.073)	(-.372, -.119)	(-.337, -.123)	(-.312, -.13)	(-.282, -.113)	(-.246, -.097)
	R ²	0.281	0.298	0.301	0.314	0.309	0.29
Italy	Effect size	-0.24	-0.252	-0.219	-0.207	-0.192	-0.176
	95% CI	(-.372, -.107)	(-.344, -.16)	(-.288, -.149)	(-.26, -.153)	(-.233, -.151)	(-.222, -.13)
	R ²	0.294	0.296	0.29	0.291	0.295	0.3
Iran	Effect size	-0.355	-0.34	-0.324	-0.347	-0.348	-0.278
	95% CI	(-.48, -.231)	(-.416, -.264)	(-.44, -.207)	(-.435, -.259)	(-.416, -.28)	(-.324, -.231)
	R ²	0.289	0.311	0.298	0.328	0.351	0.356
France	Effect size	-0.123	-0.105	-0.089	-0.077	-0.071	-0.066
	95% CI	(-.16, -.086)	(-.139, -.07)	(-.127, -.051)	(-.103, -.05)	(-.094, -.048)	(-.086, -.045)
	R ²	0.299	0.292	0.282	0.266	0.258	0.245
United States	Effect size	-0.084	-0.078	-0.091	-0.079	-0.089	-0.083
	95% CI	(-.144, -.025)	(-.119, -.037)	(-.12, -.062)	(-.105, -.053)	(-.115, -.064)	(-.105, -.062)
	R ²	0.291	0.293	0.288	0.291	0.293	0.288

Average effect of all policies on infection growth rates in our sample (analogous to the average difference between red and blue markers in Figure 3 of the main text) using models with different assumed delays of policy impact, ranging from a one day lag to a five day lag. The “0 Day Lag” column corresponds to the estimates reported in the main text. For models with non-zero lag, the lag structure of the estimating equation is such that for a model with an L day lag, the policy intervention implemented on day t is assumed to have no impact on infection growth rates until day $t+L$. As in the main text, for China and Korea, the reported effect size is the policy impact on the growth rate of active cases, where in the other countries cumulative cases are used. See Supplementary Methods 3 for a discussion of these results. The sample size is 3,669 in China, 595 in South Korea, 2,898 in Italy, 548 in Iran, 270 in France, and 1,238 in the US.

Supplementary Table 6 | Sensitivity of cases and infections delayed/avoided to assumed infection fatality rate

	IFR = 0.5%		IFR = 0.75% (Main results)		IFR = 1.0%	
	Confirmed cases avoided/delayed	Infections avoided/delayed	Confirmed cases avoided/delayed	Infections avoided/delayed	Confirmed cases avoided/delayed	Infections avoided/delayed
China	28,316,000	321,771,000	37,013,000	284,718,000	45,682,000	253,790,000
South Korea	7,955,000	39,775,000	11,547,000	38,491,000	14,966,000	37,414,000
Italy	1,404,000	48,423,000	2,122,000	49,351,000	2,828,000	49,616,000
Iran	3,540,000	58,036,000	4,900,000	53,843,000	6,071,000	50,591,000
France	934,000	44,480,000	1,401,000	45,182,000	1,846,000	45,022,000
United States	3,727,000	70,323,000	4,789,000	59,867,000	5,752,000	52,292,000
Total	45,876,000	582,808,000	61,772,000	531,452,000	77,145,000	488,725,000

Total estimated confirmed cases and infections avoided/delayed for assumed infection fatality ratios (IFRs) of 0.5%, 0.75% (main results) and 1.0%. Table shows values on the respective final days of our samples for each of the six countries presented in our analysis. We use the assumed IFRs to compute underreporting estimates following Russell et al.²⁵²

²⁵² [Using a delay-adjusted case fatality ratio to estimate under-reporting](#)

Chapter 4

Past, Present, and Future Economic Risks from Hurricanes in the United States

Hurricanes are one of the costliest natural disasters in the United States, and climate change is expected to alter the frequency, intensity, and spatial distribution of these events. These changing physical characteristics will likely alter economic risks in coastal regions. However, the magnitude of the contribution of climate change to past economic losses and to potential future losses remains uncertain. To estimate the impact of climate change on coastal risk, we construct and apply a tropical cyclone integrated assessment model that simulates economic risk for a variety of climate scenarios. We first derive the relationship between storm surge, high winds, and economic damages using physical simulations of historical storms. We then use this relationship to project damages from synthetic storms representative of 21st century climate under two different emissions scenarios. We show that risk increases nonlinearly through the 21st century. There is a notable difference in risk between the high- and low-emissions scenarios studied, indicating a potential to minimize this increase in risk through greenhouse gas emissions reduction. While analyses of coastal climate risk typically focus on sea level rise, these results suggest that changing hurricane patterns represent another substantial component of this risk.

This chapter represents work led by the author of this dissertation, in partnership with a number of collaborators from the Climate Impact Lab. These include Daniel Allen, Michael Delgado, Michael Greenstone, Ali Hamidi, Trevor Houser, Solomon Hsiang, Robert Kopp, and Maya Norman. Additional contributors from other institutions include Kerry Emanuel, Daniel Gilford, Jaecheol Lee, Moon Limb, Kyle Mandli, and Jay Sayre. The project was conceptualized by Solomon Hsiang, Robert Kopp, Trevor Houser, and Michael Greenstone. Analysis was conducted by the dissertation author, in addition to Daniel Allen, Michael

Delgado, Ali Hamidi, Moon Limb, Jaecheol Lee, Maya Norman, and Jay Sayre. The chapter was written by the dissertation author.

4.1 Introduction

Tropical cyclones (TCs) are massive collections of organized atmospheric convection that can cause large economic impacts upon landfall (Klotzbach et al. 2018). In the United States alone, the average annual loss (AAL) from North Atlantic TCs – also known as hurricanes – has been estimated at \$54 billion (Congressional Budget Office 2019). Despite this large societal cost, the influence of past and future anthropogenic climate change on these impacts remains an open question. While historical trends in TC landfall frequency and damage are uncertain (Gabriel A. Vecchi and Knutson 2011; Gabriel A Vecchi, Swanson, and Soden 2008; Roger A. Pielke et al. 2008; K. Emanuel 2005; Schmidt, Kemfert, and H  ppe 2009; Schmidt, Kemfert, and H  ppe 2010), recent research suggests that changes to the large-scale climate have measurably affected recent severe storms (Patricola and Wehner 2018). Regardless of climate-driven changes to date, evidence suggests that TC frequency (K. Bhatia et al. 2018), intensity (Sobel et al. 2016), rainfall (P. Lu et al. 2018; Kossin 2018), and spatial pattern (Kossin, K. A. Emanuel, and Gabriel A. Vecchi 2014) are all expected to change with future climate change (Knutson et al. 2010). These changes in TC characteristics will alter physical hazards, such as storm surge and high wind speeds, experienced by coastal populations. The physical capital and human populations exposed to these hazards will also be changing over the coming years, as will the vulnerability of coastal regions.

All three of these factors – hazard, exposure, and vulnerability – affect the socioeconomic risk of coastal regions. Understanding how this risk will change over the next century can benefit decision-making across contexts, from municipal and national policy to institutional investments to residential real estate purchases to insurance pricing. Importantly, knowing how climate change contributes to this evolving risk can improve our understanding of the costs of greenhouse gas emissions. In this paper, we develop and apply a physical-econometric modeling framework that explicitly addresses each of these three contributing factors. We forecast U.S. hurricane-driven economic risk throughout the 21st century under two separate climate scenarios (Representative Concentration Pathways 4.5 and 8.5, Vuuren et al. 2011). We then use these forecasts to identify the impact of climate change on this risk.

Notably, we focus on direct economic losses - defined as damage to physical infrastructure that occurs as an immediate result of a hurricane. Upcoming work will seek to identify the impact of climate on *indirect* hurricane losses, including long-term economic growth impacts as well as loss of life. While the causal relationships between each hurricane event and these outcomes may be more complex, the potential costs are also substantial (S. Hsiang and Jina 2014).

Several previous studies have estimated the socioeconomic impacts of climate-driven TC changes (S. Hsiang and Jina 2014; Gettelman et al. 2018; Mendelsohn et al. 2012; HOUSER et al. 2015; Neumann, K. Emanuel, et al. 2015; United States Global Change Research

Program 2018). Many follow the framework of a Tropical Cyclone Integrated Assessment Model (TCIAM), formalized by Mendelsohn et al. 2012, to map outputs from climate models to TC damages. However, a number of complex physical and econometric processes link large-scale climatic change to change in coastal economic risk, and not all of them have been fully characterized in these previous models. This has the potential to bias estimates of the climate’s impact.

Studies that use a single index of storm hazard, such as wind speed (S. Hsiang and Jina 2014; Gettelman et al. 2018), central pressure (Mendelsohn et al. 2012), or storm surge (Neumann, K. Emanuel, et al. 2015), may not accurately account for damages from other hazards. If climate change has a differential effect on these hazards, e.g. because sea level rise (SLR) interacts with storm surge heights but not wind speed, then these single-hazard models may produce biased estimates of future TC-driven economic risk. A limited number of studies simultaneously account for wind and surge damages, and incorporate SLR. HOUSER et al. 2015 and Dinan 2017 both rely on damage projections using a model from Risk Management Solutions, Inc. (RMS) that simulates surge and wind damages directly for the present day climate. However, both assume simplistic changes in hurricane characteristics over time by varying a hurricane category-specific frequency parameter. This parameterized approach to modeling hurricane changes fails to account for some predicted trends, such as slower moving storms (Zhang et al. 2020), that may contribute to altered hazard patterns. In addition, the closed-source nature of RMS’ damage model limits external validation and modification. Other studies provide more physically based representations of changing TC hazards (K. Yang, Paramygin, and Sheng 2019; Marsooli et al. 2019), yet stop short of coupling econometric analysis to map this physical hazard to societal impact. Still others model the damages associated with shifting the distribution of extreme sea levels due to SLR (Lorie et al. 2020), yet do not account for the changing shape of these distributions due to climate-driven storm pattern changes.

In our analysis, we address each of these challenges by simulating hurricanes across potential future climates, estimating spatial fields of both storm surge and wind for each storm using physically based models, and then jointly estimating damages from both of these hazards.

A second challenge in the literature stems from the reliance of each of these models on a dose-response function, also known as a *damage function* or *vulnerability curve*, that may not be rigorously derived from empirical analysis or representative for the full geographic region of analysis. For example, the United States Army Corps of Engineers’ Depth-Damage Relationships (U.S. Army Corps of Engineers 2000), which drive the storm surge damages in Neumann, K. Emanuel, et al. 2015; HOUSER et al. 2015, and Dinan 2017, were developed from a handful of flood study sites over three years and provide dose-response functions for a variety of residential structure types. When applied in a modeling context, if a storm surge model is biased relative to observations or if an underlying exposure dataset contains commercial or industrial infrastructure, the application of such an externally derived dose-response function may lead to biased damage forecasts. In our analysis, we address this issue by learning an internally consistent empirical dose-response function that translates our

estimated exposure and hazard to damages, using observed insurance claims from historical events.

A third challenge is the representation of tropical cyclones in unprecedented climates. The possibility for so called “black swan” events, with magnitudes that have not been experienced in the historical record, means that we cannot rely on historical analysis alone to forecast future risks. While Earth System Models (ESMs) provide much of our general understanding of future climates, most ESMs applied to climate change at the global scale have too coarse of a resolution to resolve tropical cyclones (Camargo 2013; K. Bhatia et al. 2018). Some TC damage projections rely on higher resolution “convection-permitting” models (Gettelman et al. 2018); however, such models have a high computational cost that limits the number of simulations that can be run. TCs are highly stochastic events and the dose-response function is generally assumed to be highly nonlinear (Mendelsohn et al. 2012; W. Nordhaus 2006; W. D. Nordhaus 2010; Hallegatte 2007; Roger A Pielke 2007). This contributes to a fat-tailed distribution of damages. Mendelsohn et al. 2012, for example, estimates that the worst 1% of TCs cause over 50% of the damages. Without large samples of potential future storms to draw from, projections based on only a handful of realizations may mischaracterize risk by failing to accurately capture the frequency of rare but impactful storms. In our analysis, we address this “tail risk” challenge by using a physically based TC simulation model (K. Emanuel, Sundararajan, and Williams 2008) to produce 120,000 North Atlantic storm tracks representative of different future climates, and by using Monte Carlo sampling to simulate thousands of realizations of each year in our sample period.

Despite these technical challenges, the large economic impact of hurricanes and their projected climate-driven changes make estimating future hurricane risk a key objective for a diverse array of stakeholders. Financial institutions have begun to incorporate this risk in the allocation of their portfolio (Blackrock Investment Institute 2019; Fink 2020), with implications for incentivizing climate adaptation through shifting investments. Regulators and central banks are working to ensure that insurance and reinsurance firms are accurately pricing risk when underwriting their policies (Prudential Regulation Authority 2019). Individuals are leveraging novel data sources to evaluate flood risks when navigating the real estate market (First Street Foundation 2020). And municipal and national policymakers are searching for ways to address coastal climate risks through mitigation or adaptation. Numerous adaptation options exist, and previous work has suggested that many of these approaches may be cost-effective methods for partially offsetting coastal climate damages (Desmet et al. 2018; Lorie et al. 2020; Tiggeloven et al. 2020; Tamura et al. 2019). However, deep uncertainty surrounding future environmental hazards, economic development, and demographic trends complicates advanced decision-making and action (Srивer et al. 2018). Robust adaptation thus depends crucially on accurately quantifying future costs in the absence of action, fully accounting for existing uncertainties.

This study combines recent modeling advances in climate science, tropical cyclone dynamics, SLR projection, hydrodynamic simulation, and climate econometrics. We pair models from across these disciplines with high-resolution real estate data and digital elevation models (DEMs) to provide a new generation of TCIAM that can better quantify future coastal

economic risk and serve the needs of each of these stakeholders. The risk forecasting pipeline we develop is modular, allowing for future integration of improved data or alternative component models. In the process of generating and applying this model, we develop the first database of gridded and consistently modeled wind and flood hazards for all historical storms that caused recorded damages in the continental United States (CONUS). We additionally create the first three-dimensional, high-resolution map of physical capital across CONUS (Section 4.2). We then produce the first nationally representative, empirically-derived dose response functions for hurricane winds and storm surge (Section 4.3). Our primary contribution, however, consists of the first probabilistic projections of hurricane damage in the U.S. to jointly quantify locally modeled wind and storm surge losses and account for spatiotemporal uncertainty in future SLR (Section 4.4).

4.2 Modeling Hazard and Exposure from Historical Events

To identify the relationship linking hurricane-driven flooding and high winds to damages, we first generate spatially resolved “hindcasts” of these hazards from historical events. We apply physically based wind (LICRICE, S. Hsiang and Jina 2014) and storm surge (GEOCLAW, Mandli and Dawson 2014) models to all North Atlantic hurricanes in a comprehensive tropical cyclone track dataset (Knapp et al. 2010). To estimate storm surge from each event, we further integrate topography (*The National Map—New data delivery homepage, advanced viewer, lidar visualization* 2019), bathymetry (Tozer et al. 2019), and local sea level observations and astronomical tidal estimates from NOAA tide gauges (National Oceanographic and Atmospheric Administration 2009). The result is a grid of the maximum wind speed, s_{max} , and water level, TWL_{max} , experienced during each historical event. Figure 4.1 (Column 1) summarizes this data by depicting the predicted 50-year return values for both flood height and wind speed within the historical record.

Importantly, this approach provides model-consistent estimates of spatially resolved hazard that are uncorrelated with any socioeconomic trends. Previous studies have investigated climate-driven trends in hurricane damages by searching for trends in reported damage. However, trends in damage can occur both through changing environmental factors (*hazard*) and through socioeconomic growth (*exposure*). To remove the large socioeconomic component of historical damage trends, these studies have “normalized” damages as a function of economic and population growth (Roger A Pielke and Landsea 1998; Collins and Lowe 2001; Roger A. Pielke et al. 2008; W. D. Nordhaus 2010; Mendelsohn et al. 2012). However, national economic trends are used in each of these cases that may not accurately capture differential trends in per-capita physical asset value across coastal and non-coastal regions.

¹The damping of the 1-in-50 year wind speeds in the synthetic track sets compared with the observed track set is due to an earlier error in wind field simulation that has been corrected and used to re-generate historical wind speeds. This figure will be updated when the wind model has been re-run on our synthetic track sets.

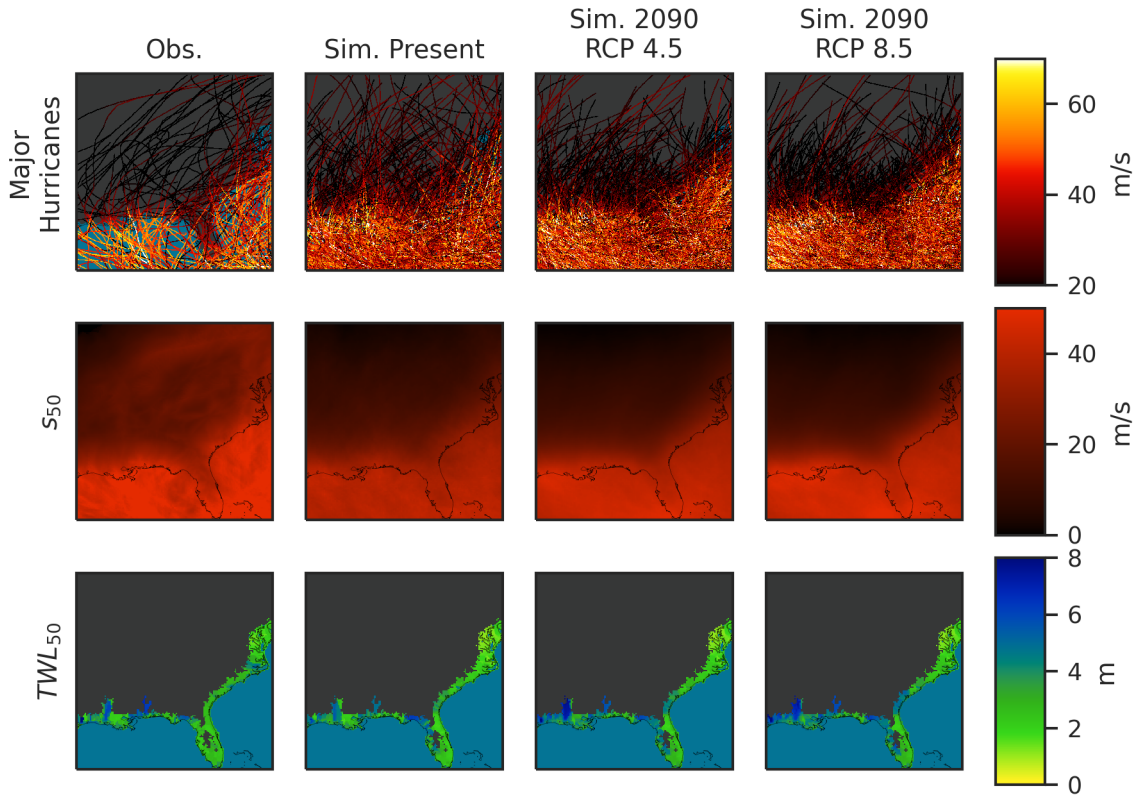


Figure 4.1: **Modeled hazard return values for observed and synthetic hurricanes.** In order, the columns refer to the historical record; synthetic storms generated from the 2008-2018 NCEP/NCAR Re-analysis; synthetic storms generated from seven GCMs in a 2085-2095 climate under RCP 4.5; and synthetic storms generated from seven GCMs in a 2085-2095 climate under RCP 8.5. Row 1 shows the track and the maximum velocity of major hurricanes (maximum wind speed > 50 m/s). The historical figure shows the entire record, beginning in 1851. For the synthetic storm contexts, 1000 simulated “seasons” are shown. Row 2 depicts estimated maximum wind speeds (s_{max}) for the 1-in-50 year wind speed in each pixel. Row 3 depicts estimated flood heights for the 1-in-50 year storm surge-driven flood (TWL_{max}) in each pixel.¹

Instead, our approach separately models physical and socioeconomic components of damage. This enables us to hold either component fixed and independently identify the contribution of the other component.

To model the socioeconomic component, we begin with a novel, property-level database of estimated values, sales prices, and tax assessment values (ATTOM Data Solutions 2019). We apply a set of logical rules to generate best estimates of the present-day value of each of 156 million properties nationwide, and aggregate property value into grid cells at 0.125° horizontal resolution and 10cm vertical resolution. To estimate historical exposures, we then scale these back using a gap-filled index of county-level housing prices (*House Price Index* — *Federal Housing Finance Agency*) and population data from the U.S. Census (US Census Bureau 2017). The output of this model is a time series of 3-dimensional grids of property

value across the continental U.S.

For each historical storm, we combine modeled wind and surge hazards with the estimated exposure grid and datasets of historical private (Verisk Analytics 2019) and public (U.S. Department of Homeland Security 2019) insurance claims. We include both datasets because in the U.S. the majority of wind damage is covered by private insurance while the majority of flood damage is covered by the National Flood Insurance Program (NFIP). The process to generate and combine each of these data sources is described in Figure 4.2, and the derived input datasets for three example storms are shown in Figure 4.3. We then apply a multiple linear regression to uncover a dose-response relationship between fractional losses and surge and wind hazards.

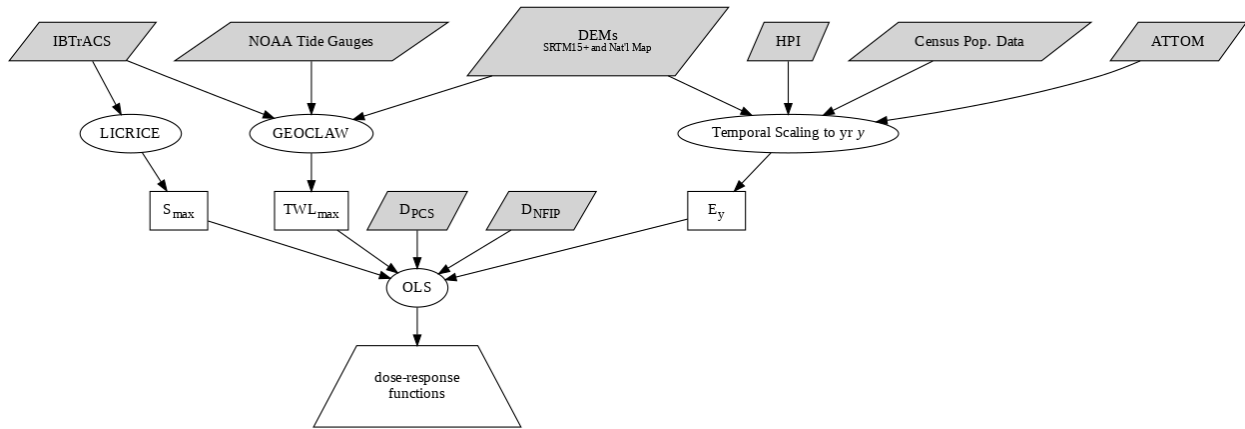


Figure 4.2: **Flowchart of input data and models used to estimate empirical dose-response functions.** Filled nodes and unfilled circles correspond to input data sources and models, respectively. Squares correspond to intermediate data products, and the trapezoid corresponds to the output dose-response relationships. (b) depicts the input and intermediate and intermediate data products used to estimate dose-response relationships. Examples are shown for three high-damage storms that span the majority of our analysis period. The first column visualizes the raw track data from *IBTrACS*, with storm size equivalent to the observed or estimated RMW (Section 4.6.1). The second and third columns visualize the outputs of LICRICE (Section 4.6.1.1) and GEOCLAW (Section 4.6.1.2), respectively. The fourth shows our spatiotemporally scaled exposure product (Section 4.6.1.3). The fifth and sixth visualize damage data from the PCS and NFIP datasets, respectively (Section 4.6.1.4). No NFIP claims exist for Hurricane Donna, as it occurred before the start of the program and is thus excluded from the NFIP dose-response regression.

4.3 Estimating Wind and Coastal Flood Dose-Response Functions

Several functional forms have been proposed for the relationships between economic losses and hurricane-driven wind or flood hazards, but no consensus emerges from the literature. For wind, many previous studies assume power law relationships of order two to five (*Anthropogenic Effects on Tropical Cyclone Activity*; Roger A Pielke and Landsea 1998; Munich

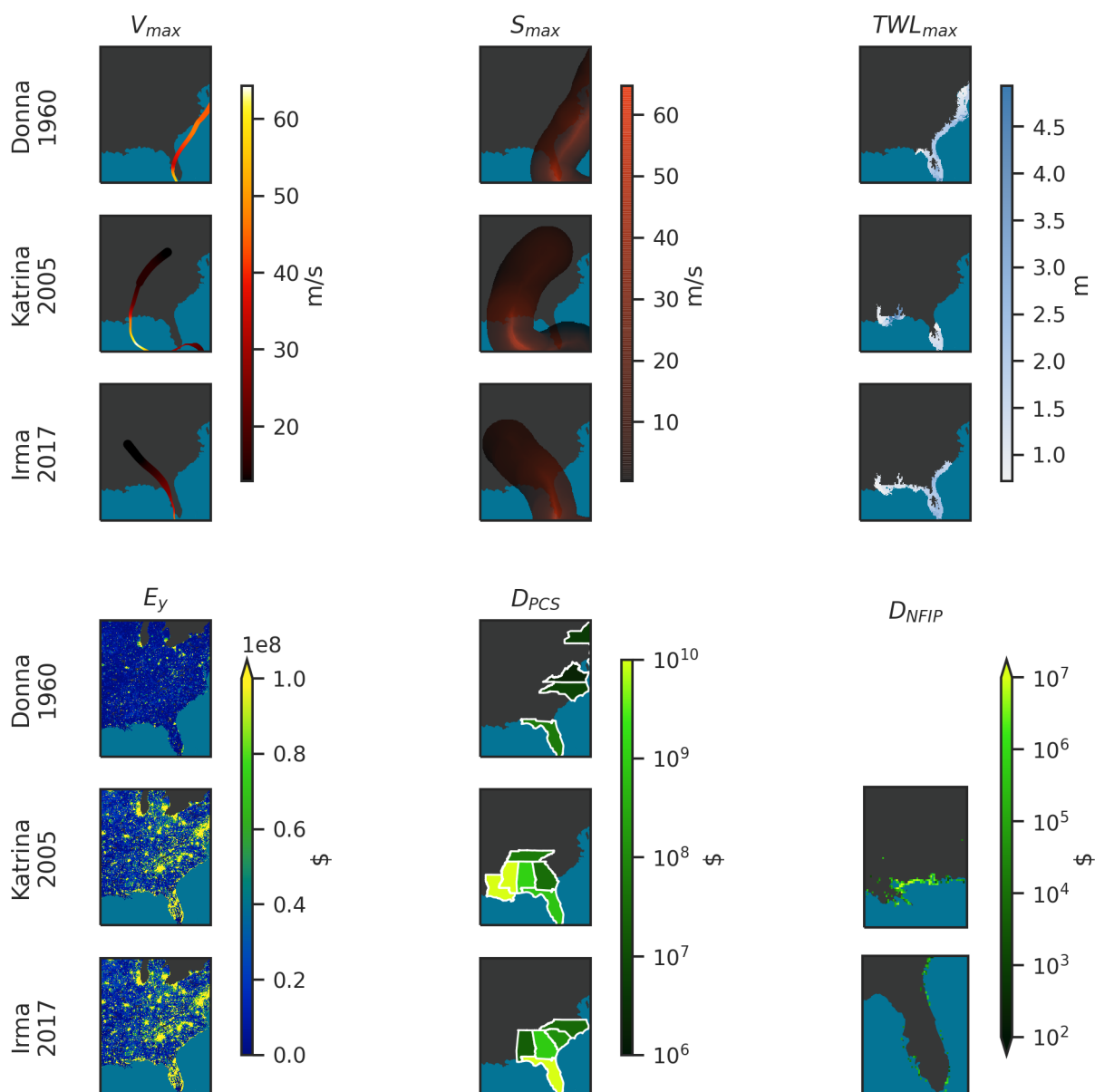


Figure 4.3: **Data and models generating hazard, exposure, and damage datasets for each historical hurricane** Each set of maps depicts the input and intermediate data products used to estimate dose-response relationships. Examples are shown for three high-damage storms that span the majority of our analysis period. Moving from left to right, top to bottom: The first column visualizes the raw track data from *IBTrACS*, with storm size equivalent to the observed or estimated RMW (Section 4.6.1). The second and third columns visualize the outputs of LICRICE (Section 4.6.1.1) and GEOCLAW (Section 4.6.1.2), respectively. The fourth shows our spatiotemporally scaled exposure product (Section 4.6.1.3). The fifth and sixth visualize damage data from the PCS and NFIP datasets, respectively (Section 4.6.1.4). No NFIP claims exist for Hurricane Donna, as it occurred before the start of the program and is thus excluded from the NFIP dose-response regression.

Re Group 2002), based on a relationship to derived physical energetic properties (Bell et al. 2000; K. Emanuel 2005). W. D. Nordhaus 2010 found a substantially higher order relationship, of order nine, through an empirical analysis using all recorded damages from U.S. landfalling hurricanes in the 20th century. One reason for this wide discrepancy could be the use of different wind metrics. While the physically based indices are functions of *experienced wind speed at a given point*, the latter study used a single storm-level scalar: *Maximum wind speed at landfall*. W. D. Nordhaus 2010 also did not control for storm surge; while flood damages were included in the outcome, only wind speed at landfall was included as a predictor. This could bias the observed wind speed-damage relationship given a correlation between wind and storm surge. Prahl et al. 2015 explored several exponential, quadratic, and cubic wind speed-damage relationships for winter wind storms in Europe, including various minimum speed thresholds. They find mixed results for the predictive skill of different functional forms, providing further evidence of the lack of consensus.

In addition to this uncertainty in functional form, it is likely that wind-damage relationship vary across locations based on local building standards and other factors that control vulnerability. This suggests that relationships derived from one location may not be broadly representative. Together, this uncertainty of functional form and heterogeneity argue for an empirically driven approach where a relationship is fit using data from the entire region in which predictions will be made. Importantly, such a function should relate damage to *experienced wind speed* at each point, rather than a storm-level wind intensity scalar such as maximum speed at landfall. This better represents the true data generating process of wind damages and can more accurately represent the differing damage characteristics of small- and large-radius storms.

Many fewer national or global hurricane damage studies have utilized *flood* dose-response relationships, possibly due to the complexity of modeling storm surge. Several of the aforementioned studies used storm-level scalar characteristics, such as the maximum wind speed at landfall, to represent wind hazard. However, storm surge magnitude is dependent on a variety of local characteristics, such as coastal bathymetry, and path-dependent characteristics, such as the size and intensity of the storm prior to landfall. While one could in theory develop a parameterized surge prediction that accounts for these features, the complexity of that relationship motivates instead the use of a physically based model to simulate surge.

The few studies that have examined surge-driven economic losses from hurricanes using numerical simulations (e.g. Neumann, K. Emanuel, et al. 2015) typically rely on depth-damage relationships from the U.S. Army Corps of Engineers (U.S. Army Corps of Engineers 2000; U.S. Army Corps of Engineers 2003). Because of the challenges in surge modeling, biases in modeled versus observed TWL could exist, and thus these observation-based depth-damage curves may be inappropriate for use in a hurricane damage forecast that uses modeled surge. Additionally, these curves are developed using data from a variety of types of flooding. Storm surge from coastal storms may exhibit different characteristics than inland flooding, including different velocities and salinities, which may alter the depth-damage relationship. Furthermore, these curves can often be guided by expert elicitation (Lehman and Nafari 2016), which may bias the estimates away from empirically valid relationships. Indeed,

a recent assessment of NFIP claims data suggests that common notions of monotonically increasing depth-damage relationships may not be empirically valid (Wing et al. 2020). Despite myriad challenges, the influence of SLR on storm surge heights necessitates that accurate forecasts of future hurricane damages simulate surge-driven damages in addition to wind damages. For this reason, we jointly estimate an empirical depth-damage function alongside a wind-damage function using our modeled surge from historical storms.

Using high-resolution modeled fields for both wind and surge, paired with newly available disaggregated data for both private insurance claims (at the state level) and public insurance claims (at the claim level), our empirical analysis exploits spatial (co)variation in hazard and exposure. To do so, it follows a process formalized by S. Hsiang and Jina 2014 for combining local hazard and exposure information with aggregated outcome data. Here, we calculate nonlinear transformations of the hazard variables (maximum sustained wind speed and maximum depth of inundation) at the building-level, representing the dose-response function as a linear combination of these transformed parameters plus an error term:

$$D_i := D(h_i, E_i) = (\gamma(h_i)^T \beta + \epsilon_i) E_i \quad (4.1)$$

where D_i is the damage for a given building i , E_i is the value (a.k.a. *exposure*) of that building, h_i is the vector of hazard variables experienced at that building, γ is the vector-valued nonlinear transformation operator, β is a vector of coefficients, and ϵ_i is a random noise term. Given this “linear-in-the-parameters” model, we can sum the transformed and value-weighted hazard vector $\gamma(h_i)E_i$ across the spatial unit of the outcome data. We then run the regression at the aggregated scale to recover β (see Section 4.6.1.4 for a detailed description of this procedure).

Figure 4.4a shows the resulting dose-response functions. The value of joint estimation of wind and surge damages can be seen when comparing the curves generated from NFIP claims. After controlling for inundation depth, we see little correlation between wind speeds and losses. This matches expectations, as NFIP only covers flood-driven damages. Inundation depth does show a strong correlation with PCS claims. This could be due to one or both of two potential factors. First, the small amount of physical capital covered by private flood insurance in that dataset may consist of high-risk, often damaged buildings. Second, individuals may successfully claim flood damages on their policies when the experience of compound hazards makes it difficult for assessors to disaggregate the driver of losses. Regardless of cause, we sum the empirically estimated dose-response functions across our two claims datasets to use in projecting future losses.²

²There is ample evidence for heterogeneity in dose-response relationships over space and time (S. Hsiang and Jina 2014; W. D. Nordhaus 2010). To address this, ongoing work explores interactions in the dose-response function with local income levels and historical hazard characteristics. The theoretical justification for these interaction terms comes from the concept that adaptation is driven by resource availability and perceived risk. Increased income is likely correlated with increased financial resources needed to implement adaptation measures, and increased historical hazard is likely correlated with a higher perceived risk.

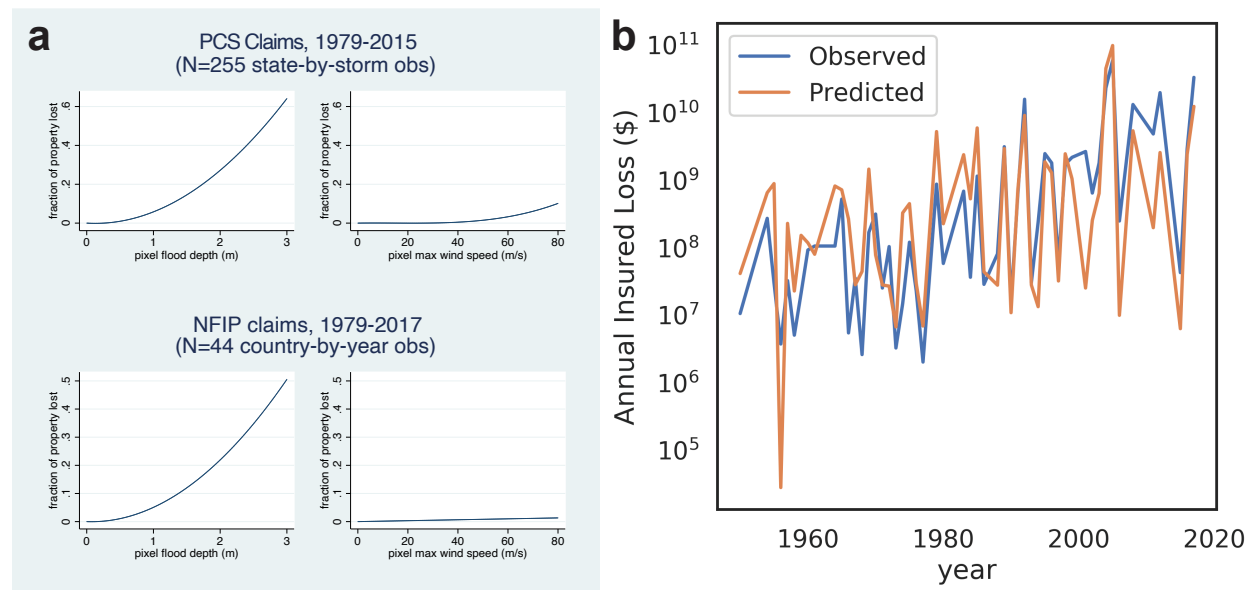


Figure 4.4: **Dose-response relationships estimated from historical hurricanes.** (a) shows the empirically estimated dose-response relationships emerging from both private (PCS) and public (NFIP) insurance claims. Predicted values are uncorrected for uninsured properties. (b) shows out-of-sample validation on the PCS (private) claims dataset. Values are cumulative annual losses, with predictions made using grouped 5-fold cross validation.

4.4 Projecting Risk Under Climate Change

Once we have identified dose-response from the historical data, we can estimate the impact of climate on economic risk by projecting damages in simulated hurricane seasons under different climates. While analyses of historical damage trends have not conclusively demonstrated an environmentally forced signal (Roger A Pielke and Landsea 1998; Roger A. Pielke et al. 2008), the character of annual hurricane damages may be too stochastic to reliably observe the underlying risk over a relatively short historical record (Hallegatte 2007). The alternative approach that we employ is to use a stochastic tropical cyclone simulator (K. Emanuel, Sundararajan, and Williams 2008), forced with different large-scale climates, and then to predict damage from the simulated storms. Mendelsohn et al. 2012 demonstrated this approach using an indexed dose-relationship pinned to storm-level characteristics including minimum pressure, maximum wind speed, and landfall location. Neumann, K. Emanuel, et al. 2015 extended the modeling framework to directly simulate storm surge in an analysis of future coastal flood risks. We further improve on the approach by using the aforementioned empirically estimated, hazard-based dose-response functions; by using explicit simulation of physical hazards from the synthetically generated storms; and by jointly simulating both wind and surge damages.

Our risk projection model is divided into five steps. In the first, we extract climates from seven different global climate models, under two assumed global emissions trajectories (RCP

4.5 and RCP 8.5), and at four 11-year time intervals over the 21st century: Present (2015-2025), near-term (2035-2045), medium-term (2055-2065), and end-of-century (2085-95). We also extract information on past and present climates using the NCEP/NCAR Reanalysis for the 1979-89 and 2008-2018 periods. We feed each of these climates into the hurricane simulator to generate 120,000 synthetic hurricane tracks (Section 4.6.2.1). The simulator also outputs a mean hurricane frequency for each simulated climate.

In the second step, we again apply wind and storm surge models to estimate gridded s_{max} and TWL_{max} for each simulated storm track (Section 4.6.2.2). Figure 4.1 shows predicted return values estimated using synthetic tracks generated from the NCEP reanalysis (Column 2) and from the seven GCMs at the end-of-century (Columns 3 and 4). Figure 4.5 disaggregates the multi-model mean end-of-century results across each of the seven models.

In the third step, we incorporate spatiotemporal variance in SLR projections (Section 4.6.2.3). In the fourth, we utilize the previously estimated dose-response functions to predict damage from multiple realizations of hurricane seasons under each climate and SLR projection (Section 4.6.2.4). And in the fifth, we compare the distributions of these simulated damages across climates and assess the contribution of climate to differences in risk (Section 4.6.3).

Our primary result is an estimate of the average annual loss from hurricanes across multiple time periods in the 21st century. These results are presented in Figure 4.6.

These results suggest that cumulative *insured* losses are expected to increase from \$12-14B/yr to \$20-57B/yr under RCP 8.5 and to \$13-23B per year under RCP 4.5 at end of century. The majority of that increase is driven by increased surge damage in the later part of the 21st century, which occurs due to changing storm frequency but also rising sea level (Figure 4.7 shows the correlation between expected losses and SLR). At the same time, the majority of the cross-model uncertainty in total damage is due to projections of wind hazard³. Overall, cross-model uncertainty grows over time, particularly in the RCP 8.5 scenario, reaching ranges in the end-of-century period that are of similar magnitude to the projections of the median model.

While estimates of the fraction of total economic losses that are insured varies, a ratio of 2 appears to be a commonly used, yet conservative, estimate (Roger A Pielke and Landsea 1998).⁴ Assuming this multiplier to estimate economic losses from predicted insured losses, a 3% discount rate, a 2% GDP growth rate, and an elasticity of property value of 0.45 (Neumann, Hudgens, et al. 2010), the multi-model mean in the RCP 8.5 scenario projects the cost of climate driven TC damages to be \$1.7T in present discounted value (PDV). This is roughly 9% of current US GDP and roughly 2.5 times a recent estimate of the loss in property value associated with nuisance flooding due to SLR (using the same discounting rate and growth assumption) (Union of Concerned Scientists 2018).

³Our SLR projections are conditioned on emissions scenario and year, and thus do not vary across models and do not contribute to the cross-model variance

⁴In ongoing updates to this analysis, adjustments for insurance coverage rates are performed prior to dose-response identification and are performed separately for surge and wind based on disaggregated coverage information from U.S. Department of Homeland Security 2019 and Congressional Budget Office 2019.

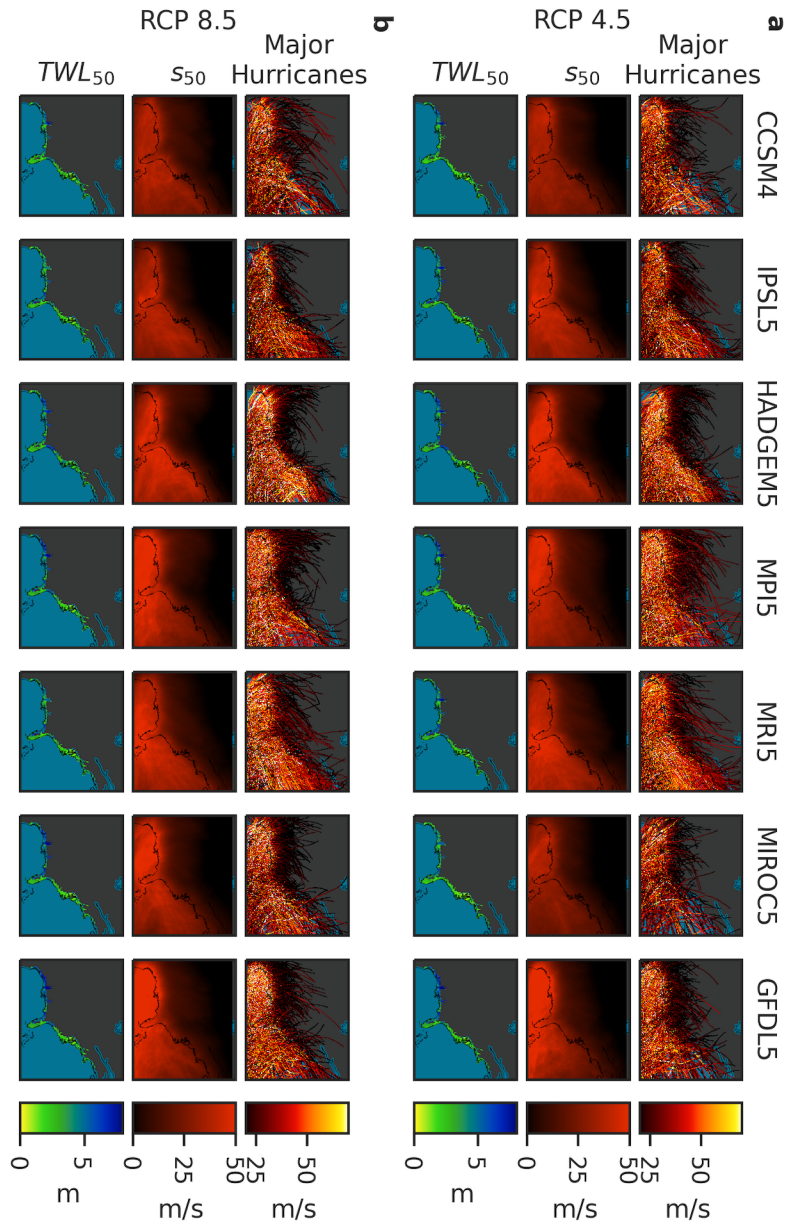


Figure 4.5: **Modeled hazard return values for the 2085-2095 period predicted by each of seven GCMs.** Figure interpretation is as in Columns 3 and 4 of 4.1. **(a)** shows results from the moderate-emissions scenario (RCP 4.5) while **(b)** shows results from the high-emissions scenario (RCP 8.5).

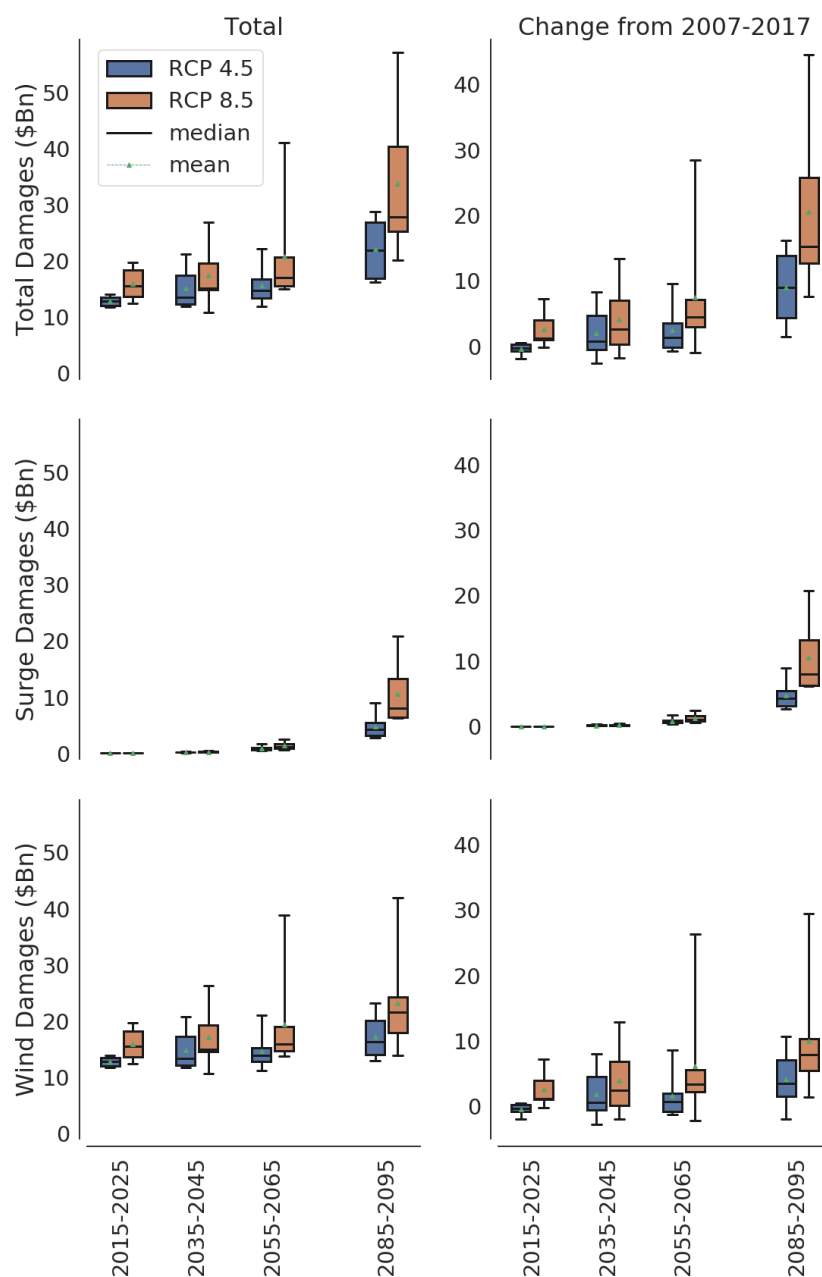


Figure 4.6: **Projections of nationally aggregated Average Annual Insured Loss (AAIL)**. Boxplots show (left) Average annual insured loss by time period and (right) change in average annual insured loss from a 2007-2017 base period. Boxes show inter-quartile range (IQR) and whiskers show the full range across 7 climate models. The top row shows total estimated damages, while the bottom two break this into storm surge- and wind-driven damage. Predictions are made for four 11-year time periods within the 21st century, for both RCP 4.5 and 8.5.

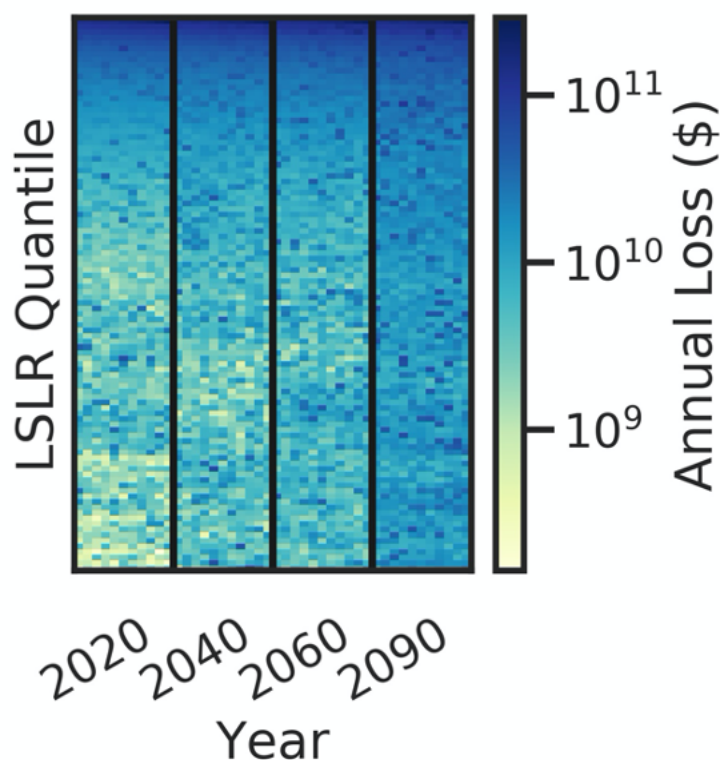


Figure 4.7: **Modeled AAIL for RCP 8.5, by SLR quantile.** Each cell represents mean damages for a single year, averaged across seven GCMs and 100 simulated hurricane seasons within each GCM. Results are ordered by the quantile of SLR uncertainty assumed in simulations. X axis labels depict the midpoint year for each of four 11-year intervals.

In Figure 4.8, we show damages resolved spatially to the level of Core-Based Statistical Areas (CBSAs) and compared across the present-day period and the end-of-century. Coastal CBSAs with a significant exposure of assets, such as New York-Newark-Jersey City, experience a large share of the additional losses, as do Western Gulf regions. The lower panels show total damage, rather than changes over time, disaggregating this damage into surge- and wind-driven components and normalizing by the total exposed value in each CBSA. This describes the total fractional loss for each CBSA at end-of-century. As expected, surge damages are concentrated in coastal CBSAs, while wind damages reach further inland. Fractional losses are greatest in the Gulf and South Atlantic regions, with a concentration of additional surge-driven losses in the Western Gulf. This finding is consistent with the results of Marsooli et al. 2019.

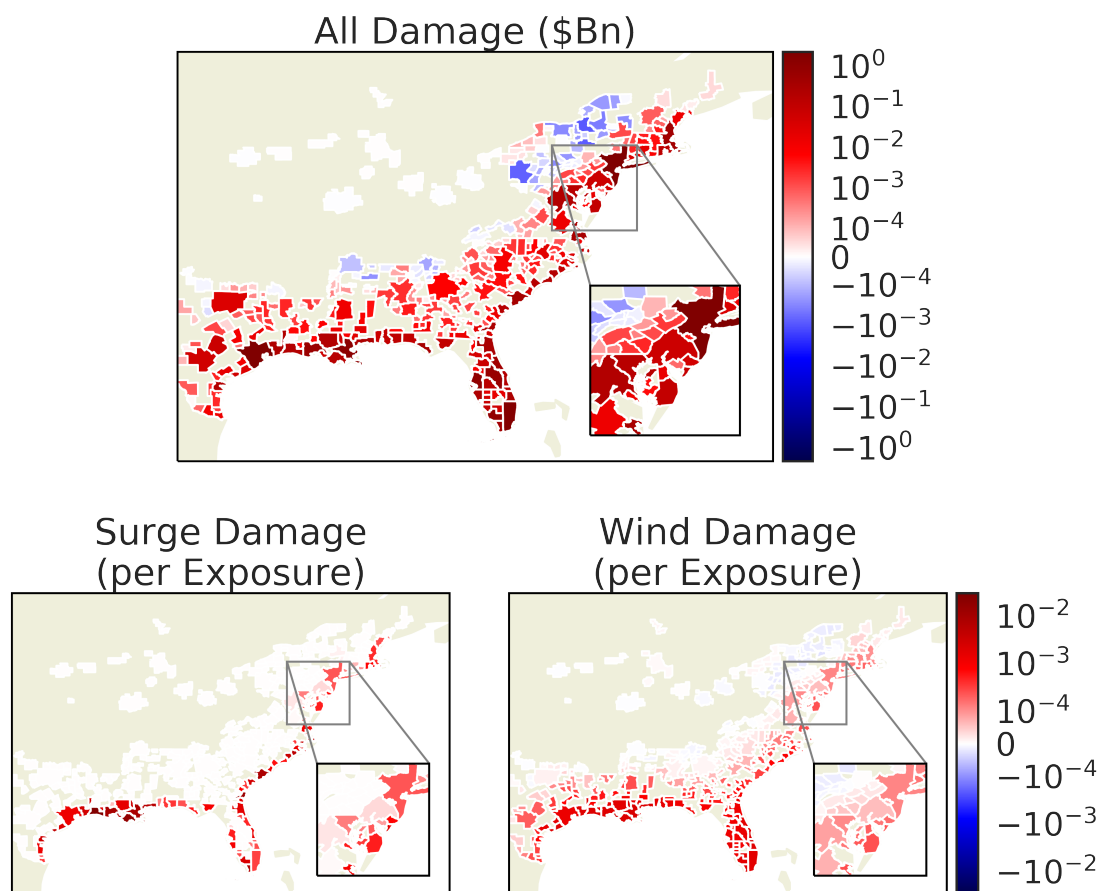


Figure 4.8: Mean estimated change in AAIL from present-day to end-of-century, by CBSA, assuming RCP 8.5. Top panel shows the multi-GCM mean change in AAIL for each CBSA. Total damages, rather than changes, are shown in the bottom row and disaggregated by wind (top left) and surge (top right). Additionally, the displayed wind and surge damages are normalized by the exposed value in each CBSA. Inset shows NY/NJ area.

4.5 Discussion and Future Work

The results of our analysis suggest that changes to hurricane patterns in the North Atlantic will bring a substantial increase in economic risk to the United States. This finding is consistent with previous studies performed at other scales (Mendelsohn et al. 2012; Neumann, K. Emanuel, et al. 2015). It provides further evidence that storm patterns must be considered along with SLR as a component of physical climate risks in coastal regions. By explicitly modeling hazards from projected storms, our results also carry novel spatial information. We demonstrate that while fractional losses are greatest in the Gulf region, areas of substantial capital in the Northeast stand to experience the greatest increase in losses absent adaptive measures. We show that increasing coastal flood damages in the later part of the century, driven largely by the superposition of storm surge and SLR, is a substantial component of climate-driven damages. This demonstrates the importance of including both surge and wind hazards in projections of future hurricane risk, as this signal would be lost had we made damages a function purely of hurricane wind speed.

While our analysis introduces explicit, spatially resolved modeling of hurricane hazards to the literature on climate-driven hurricane damages, several notable limitations of the current approach remain. We characterize several of the important limitations below and describe how they will be addressed in ongoing or future work.

First, we project only *direct* infrastructural damages. In doing this, we leverage the higher-resolution data available for these impacts but ignore the sometimes substantial societal cost of hurricane-caused mortality and long-term economic effects of a hurricane “shock” that are mediated by macroeconomic channels. Some estimates suggest such that indirect costs of climate-driven changes to global TC patterns might reach \$10 trillion (S. Hsiang and Jina 2014). This large potential price tag implies that these costs are crucial to include in a comprehensive analysis of climate change impacts. Our empirical model is flexible to a variety of outcomes other than our current measures of insured loss. Future work will incorporate mortality, following valuation approaches outlined in (Carleton et al. 2018), and long-run GDP growth rate changes, following (S. Hsiang and Jina 2014), to address these other TC impacts.

Second, we do not model the impact of tropical cyclone rainfall (TCR) and inland flooding. Numerical simulations have suggested that TCR has already been affected by climate change (Risser and Wehner 2017), and TCs are projected to deposit even more rainfall in future climates (P. Lu et al. 2018; Kossin 2018). Our results will fail to capture this climate signal, rendering our current projections conservative in our estimate of the impact of climate change. Future work will incorporate a recently developed TCR model (P. Lu et al. 2018) to properly represent this additional hazard.

Third, we explicitly project damages *in the absence of adaptation*. This serves to motivate efficient adaptation by quantifying the potential benefits from avoided hurricane losses. However, the true societal costs of climate-driven changes to hurricanes incorporate the costs and benefits of adaptive measures that will be employed. For instance, migration out of areas seeing increased economic risk might bring costs to those populations but could avoid even

greater losses as their hurricane risk rises. To measure this, one must explicitly account for changing exposure maps and dose-response functions as a consequence of adaptation, and must also incorporate the costs of these adaptive changes. Carleton et al. 2018 demonstrates a method for predicting changes in the vulnerability, or dose-response, of regions as they adapt to climate change. This approach, along with a dynamic adaptation model to represent migration due to SLR (Diaz 2016), will be incorporated in future work conducted at the global scale.

Fourth, our analysis relies on only one model that relates large-scale climate to stochastically simulated hurricanes. We attempt to account for uncertainty in future damages by including seven GCMs, two emissions scenarios, and uncertainty in local SLR. However, this uncertainty quantification would be made more complete by surveying multiple hurricane track simulators. Dinan 2017 suggests that the use of alternative representations of future hurricane intensity and frequency can have substantial effects on estimated economic risk. Until recently, the model we use from K. Emanuel, Sundararajan, and Williams 2008 was the only available environmentally forced TC simulator. A recently developed alternative model (C.-Y. Lee et al. 2018) will be incorporated in future work.

Fifth, one additional source of unquantified uncertainty comes from errors in our topography representation. We rely on the finest-resolution comprehensive DEM over the continental U.S. (*The National Map—New data delivery homepage, advanced viewer, lidar visualization* 2019); however, 95% confidence intervals surrounding the accuracy of each 1/3 arc-second pixel is 3m (Gesch, Oimoen, and Evans 2014). The bias is estimated to be much smaller, at -29 cm; however, both bias and variance contribute to additional uncertainty in the damages incurred by flooding of a given height. At a global scale, Kulp and Strauss 2019 show that biases in commonly used elevation products can change flood impacts efforts by a factor of 3. However, the lower bias in the U.S. DEM used in this analysis would limit this effect. Furthermore, our empirical approach for estimating dose-response functions can implicitly adjust to counteract systematic biases in flooding depth. Regardless, to address the contribution of elevation uncertainty to damage projections, future work will incorporate a distribution of potential elevations for each property within our broader Monte Carlo approach.

Despite these limitations, probabilistic, spatially resolved projections that account for wind, storm surge, and SLR represent a new generation of hurricane risk forecasting model. These results can inform, and have informed, adaptation decisions across a variety of stakeholders. They can guide financial institutions' investment portfolios (Blackrock Investment Institute 2019; Rhodium Group 2019), inform individuals real estate decisions (First Street Foundation 2020), and affect policy from municipal to national levels (Hess et al. 2019). Future work will apply this methodology at a global scope, using the results to help inform a data-driven estimate of the contribution of coastal damages to the Social Cost of Carbon (SCC). Additional work can leverage these spatially resolved damage forecasts to better understand the distributional impacts of changing hurricane risks.

4.6 Methods

Our goal is estimate the impact of climate change on hurricane-driven coastal economic risk. To do this, we simulate direct economic losses from a large number of synthetic hurricane seasons under several climate scenarios. Aggregating these simulated damages into a probability distribution provides a characterization of economic risk, and comparing these distributions across climate scenarios allows us to quantify the impact of climate change on this risk. Equation 4.1 defines the conceptual model used to estimate damage. Equivalently, we can refer to the “fractional loss” for building i as a function purely of hazard and the noise term.

$$D_{frac,i} := \frac{D_i}{E_i} = \gamma(h_i)^T \beta + \epsilon_i \quad (4.2)$$

Risk, R , we define as the distribution of damages:

$$R_i := CDF(D_i) \quad (4.3)$$

Our model to estimate and compare risk under different climates can broadly be divided into three stages. Each can be interpreted within this conceptual model as follows:

1. Empirical estimation of β from E , h , and D , using observed historical events (Section 4.3).
2. Simulation of D under a variety of climate conditions, using β from Step 1, E projected into the future, and h from synthetic storms (Section 4.4).
3. Statistical analysis of the distributions of D across climate scenarios to estimate R for different climates and, thus, to estimate the impact of climate change (Section 4.4).

The input data into each of these steps, along with the modeling methodology applied to those inputs, are described in detail below.

4.6.1 Dose-Response Estimation

Figure 4.2 shows the data and modeling pipeline used to estimate a dose-response function relating storm surge inundation and wind speed magnitudes to the fraction of exposed physical capital that is lost. We begin with the set of observed storm tracks in the North Atlantic and Eastern Pacific basins from the International Best Track Archive for Climate Stewardship (IBTrACS) (Knapp et al. 2010). Among other variables, this dataset contains information on the location of the storm’s eye, the maximum wind speed (V_{max}), the central pressure (P_0), the radius of maximum wind (RMW), and the “radius of the last closed isobar” ($ROCI$) – which we take to be the outer radius of the storm. For storms that are

missing radius information, we estimate RMW with the regression equation used for wind modeling in Solomon M. Hsiang 2010:

$$RMW_{km} = 63.2730 - 0.8683 \times V_{max}^{\theta} + 1.00700 \times \phi \quad (4.4)$$

where V_{max}^{θ} is the maximum circular wind speed observed at a given time, and ϕ is the latitude of the storm's eye, in degrees. V_{max}^{θ} is calculated by removing a portion of the storms translational speed component from observed maximum wind speeds (see Section 4.6.1.1). We fill in missing observations of $ROCI$ with 1000km.⁵

We subset this dataset to storms that match records in either of two sources of insurance claims data. The Property Claims Service (PCS) database (Verisk Analytics 2019) records private insurance claims from extreme weather events beginning in 1950 and is aggregated to the state level, while the National Flood Insurance Program (NFIP) reports damages at the national level from 60 damaging events beginning in 1978 (*Significant Flood Events*).⁶ Together, these represent the bulk of insured losses from hurricanes in the United States. In total, we match these claims to 118 storm tracks from 1950-2018.

For each of the 118 tracks, we simulate spatial fields of the maximum wind speed, (s_{max} , Section 4.6.1.1) and maximum total water level (TWL , Section 4.6.1.2) at 0.125° spatial resolution. We merge this with building-level estimates of elevation for 156 million properties in the United States (Section 4.6.1.3) to get estimates of inundation depth for the subset of buildings exposed to storm surge in a given storm. Using this estimated hazard and exposure data for all 118 storms, we then empirically fit dose-response functions relating observed insurance claims to depth of flooding and maximum experienced wind speed (Section 4.6.1.4). Each of these inputs are shown in Figure 4.3 for a set of three example storms.

4.6.1.1 Wind Field Estimation

To estimate wind hazard for a given storm track, we adapt the Limited Information Cyclone Reconstruction and Integration for Climate and Economics (LICRICE) model first applied in Solomon M. Hsiang 2010. This model estimates a wind field at each time step of a storm's evolution using a Modified Rankine vortex and then integrates that wind field through space. It tracks both the maximum wind speed (s_{max}) experienced by each grid cell in a domain, as well as a grid-cell-level measure of the Power Dissipation Index (PDI, K. Emanuel 2005). In our analysis, we use only s_{max} as PDI was not shown to produce additional skill in predicting damages. First, when information on RMW and $ROCI$ is available in IBTrACS, we use it rather than estimating these parameters from a latitude-based regression. Second, the model radius is extended from 250km surrounding the storm's eye to 400km to better model the tails of larger storms. Third, when separating the maximum observed wind speed

⁵This parameter affects only the outer reaches of our modeled wind fields and, given that damages are strongly nonlinear in wind speed, should have little effect on predicted wind damages. For this reason we choose 1000km as a round imputation value that lies within the support of observed radii

⁶Recently, NFIP has released disaggregated claim and policy coverage information that is currently being incorporated into a revised version of this analysis.

into translational and gradient (circular) components, a latitude-dependent fraction of the measured translational speed is subtracted from the total speed, rather than the entire translational speed:

$$v_{circular} = v_{total} - \min(1, 0.8 + .35(1 + \tanh((\phi - 35)/10)))v_{translational} \quad (4.5)$$

where $v_{translational}$ is calculated from observed storm eye positions using the mean of forward- and backward-Euler velocity calculations:

$$v_{translational,t} = \frac{dist(x_{t+1}, x_t) + dist(x_t, x_{t-1})}{2\Delta t} \quad (4.6)$$

This change better accounts for the inherent variance in gradient wind speed in a real hurricane that is not represented with a symmetric vortex model. The coefficients in Equation 4.5 are taken from analysis scripts associated with K. Emanuel 2005.

Fourth, because the Rankine vortex is known to overestimate winds far from the storm's eye, a hyperbolic tangent ramping function is used to gradually ramp down the gradient winds estimated by the model at large radii:

$$v^*(r) = v(r) \cdot (0.5 \cdot (1 - \tanh((r - r_{out})/100km))) \quad (4.7)$$

where r_{out} is the outer radius of the storm, chosen to be equal to $ROCI$ when reported in IBTrACS or set to 1000km when missing.

Fifth, the translational velocity is tapered to 0 close to and far from the eye. These last two changes match the wind scaling methodology used in the hydrodynamic surge model described in Section 4.6.1.2.

4.6.1.2 Surge Field Estimation

To estimate surge hazard for a given storm track, we adapt GEOCLAW (Mandli and Dawson 2014), an Adaptive Mesh Refinement (AMR)-based hydrodynamic model that has been previously used for storm surge simulation. The use of a dynamic model for surge is driven by the spatial characteristics of storm surge. As opposed to hurricane winds, which follow relatively smooth spatial patterns that can be reasonably represented by a low-dimensional parameterization, surge is affected by dynamic interaction between winds, ocean currents, and local bathymetry and topography. This leads to large differences in surge heights across space that are not easily parameterized statistically.

Our conceptual model for the storm tide experienced during an event is one of a linear superposition of components:

$$storm_{tide}(t) = SL(t) + S(t) + T(t) \quad (4.8)$$

where SL is the local sea level, S is storm surge, and T is the tidal contribution. This approach follows the linear model outlined in (Kirezci et al. 2020), with the exclusion of wave setup (WS). This exclusion means that our model predicts “storm tide” (the combination

of surge and tide) rather than TWL , which additionally incorporate wave action. Wave contributions represent a component of hurricane-driven flooding that are not simulated in most hydrodynamic models. Previous studies have coupled a model for wave setup to a storm surge model (e.g. Thomas et al. 2019) for individual storm case studies, but this is computationally infeasible at scale. While the substitution of storm tide predictions for TWL will lead to lower forecasted flood depths, the use of an empirical dose-response function allows for a systematic bias in flood hazard to be absorbed by a shift of the dose-response curve.⁷

Hydrodynamic models are often used to assess local surge risks in individual locations (Mandli and Dawson 2014; Sebastian et al. 2014; N. Lin et al. 2016), yet substantial model tuning and computational cost is typically required. This might be feasible for predicting surge from our 118 historical storm surge models across the Atlantic and Gulf coasts; however, running our 120,000 synthetic storms through a basin-scale model at sufficient resolution to recover local bathymetric effects would challenge available computational resources. For example, a single high-resolution storm surge simulation with the Advanced Circulation (ADCIRC, *ADCIRC*) model can take 2000 core-hours or longer if coupled to a wave model. Simulation of 120,000 storms is thus a 200 million core-hour problem. A recent study by Marsooli et al. 2019 utilized a basin-scale, moderate-resolution unstructured mesh simulation to facilitate surge estimation over a wide area from a large numbers of storms. The unstructured mesh approach allows for spatial heterogeneity in resolution, targeting computation in regions of interest. We choose an alternative approach by developing an automated configuration pipeline for an AMR model. The adaptive mesh allows the resolution of the model to vary not only in space but also in time by “tracking” the storm with a high-refinement region. In this way, the computational domain adapts to each individual track, and the model is resolved to a fine scale only within regions and time intervals that are relevant to that track. This further reduces computational burden, allowing for storm surge simulation from a greater number of events.

Figure 4.9 illustrates the computational benefits of AMR simulation when compared to an alternative high-resolution storm surge+wave simulation (ADCIRC+SWAN). GEOCLAW achieves better precision and comparable accuracy in predicting maximum water heights observed at tide gauges. The root-mean-squared-error (RMSE) of the GEOCLAW

⁷In a revised version of this analysis, we are fitting a simple statistical parameterization that adjusts storm tide to TWL using a linear function of maximum wind speed:

$$TWL_i = \alpha storm_tide_{i,max} + \beta s_{max,i} + \mu_i \quad (4.9)$$

$\alpha = 1.03$ and $\beta = 0.38$ are fit via linear regression where the TWL observations are obtained from a USGS High-water-mark dataset (Koenig et al. 2016), and s_{max} is obtained from the LICRICE results described in the preceeding section. Additional predictors were included in sensitivity tests, including the maximum wind speed parallel to the shoreline gradient, the elevation, slope, and direction of the shore, and interactions between these variables; however, best out-of-sample predictive performance was obtained with the simple function of s_{max} . Figure 4.10 illustrates the effect of the statistical wave runup on predicted total water levels.

predictions, -41cm, is 7cm less than that of ADCIRC+SWAN across these seven storms. ADCIRC+SWAN predictions, on the other hand, demonstrate a bias of only +2cm, while GEOCLAW predictions demonstrate a -10cm bias. This low-bias may be due to the exclusion of dynamically simulated wave action. ADCIRC+SWAN incorporates the static component of waves (wave setup), thus accounting for the component of wave action that is measured by tide gauges.⁸ We correct for this bias statistically in a post-processing step, described in the following section.

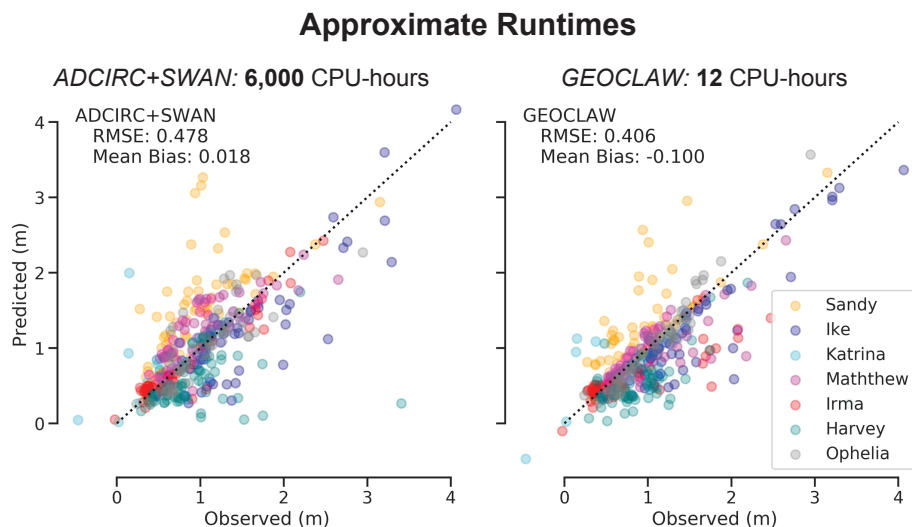


Figure 4.9: **Comparison of GEOCLAW to ADCIRC+SWAN.** This figure shows predicted versus observed maximum water heights at a selection of NOAA tide gauges near the storm track for each of seven historical hurricanes. The left panel shows simulations using ADCIRC coupled with the Simulating Waves Nearshore (SWAN) wave model (N., L.H., and R.C. 2020). The right panel shows simulations with GEOCLAW. Mean runtimes across these seven storms are displayed above the figure in CPU-hours.⁹

Our adaptation of GEOCLAW includes the introduction of pre- and post-processing steps, as well as the modification of the model itself. For the full set of additions and modifications, one should reference the source code for this manuscript (In Prep.); however, the key changes are described below.

Pre-processing First, prior to the model execution, we select an appropriate spatial and temporal computational domain based on track location and storm radius. Second, we select the initial sea level. To do this, we select all NOAA Tide Gauges (National Oceanographic and Atmospheric Administration 2009) that were active at the time of the storm and located within 5x the maximum wind radius for the storm at any point. We average observations

⁸Dynamic wave action, otherwise known as *swash* is integrated out of 6-minute tide gauge observations, and thus does not appear in either the predicted or observed values for either model

⁹ADCIRC+SWAN simulations were carried out by Celso Ferreira and Arslaan Kahlid of the Flood Hazards Research Lab at George Mason University.

for each gauge in the 7 days prior to the first storm track observation to determine the local short-term mean sea level (*LSTMSL*). Using the same NOAA API, we calculate the mean predicted astronomical tide during that same interval, which we term local short-term mean tide (*LSTMT*). We also calculate the predicted tide *during* the storm (*tide*) and set the initial sea level to the theoretical max water height that would have occurred during the simulation interval in the absence of the storm:

$$SL_{init} = \max_g(\max_t(tide) - LSTMT_g + LSTMSL_g) \quad (4.10)$$

where g indexes nearby gauges, and t indexes the time interval of the storm. GEOCLAW does not simulate tidal variability, so this step allows for the simulation of surge at a baseline water height equal to the maximum experienced tidal value. Third, we select all of the points at which we have estimated physical capital (see Section 4.6.1.3) and configure these as hourly output locations for the model.

Storm surge modeling We then run GEOCLAW, using a baseline resolution of 0.25° , with a minimum resolution of 30 arc-seconds ($\tilde{1}$ km). The resolution at any given point in space or time is determined by a variety of refinement criteria and lies somewhere between those two values. Numerical instability is a challenge for storm surge models where a noticeable amount of surge reaches the boundary. Because of this, we introduce a step into the numerical integration of GEOCLAW that checks for boundary cell velocities and restarts the model with a larger domain should these velocities exceed 5 m/s. To define bathymetry in the model, we use the SRTM15+ V2 15 arc-second global blended topography-bathymetry DEM (Tozer et al. 2019). To define topography, we use the National Map (*The National Map—New data delivery homepage, advanced viewer, lidar visualization* 2019), a $1/3$ arc-second DEM. We use the Holland 1980 wind model (Holland 1986) to define the model’s wind forcing and apply a drag law that is linearly increasing in velocity (Garratt 1977). We double the drag coefficient implied by this previously-identified relationship, a tuning step optimized to fit modeled historical storm surge return periods to observed surge in a separate analysis performed as a part of First Street Foundation 2020.

Post-processing Because we simulated surge on top of a baseline sea level equal to the maximum tidal value at any location in the domain of the storm, we next subtract the relative local tidal signal from the model outputs. To do so, we match each output grid cell to its nearest neighbor NOAA gauge and subtract that gauge’s predicted astronomical tide time series from the time series of output. We then take the maximum of this linear combination (see Equation 4.8) to get our predicted maximum storm tide:

$$\begin{aligned} storm_tide_{i,max} &= SL_{init} + \max_t(surge(t,i) + tide(t,j)) + \delta_{SL_{init},j} \\ &= \max_g(LSTMSL_g - LSTMT_g + \max_t(tide_g)) \\ &\quad + \max_t(surge(t,i) + tide(t,j)) + \delta_{SL_{init},j} \end{aligned} \quad (4.11)$$

where i indexes grid cells; j is the nearest NOAA gauge to grid cell i ; $SL_{init} + surge(t, i)$ is the output of GEOCLAW; and $tide(t, j)$ is the predicted astronomical tide for gauge j . $\delta_{SL_{init},j}$ defines the difference in pre-storm water height at gauge j from SL_{init} and is used to account for regional differences in tides and starting sea level. Formally, it is defined as the following:

$$\begin{aligned}\delta_{SL_{init},j} &= LSTMSL_j - LSTMT_j - SL_{init} \\ &= LSTMSL_j - LSTMT_j - \max_g(LSTMSL_g - LSTMT_g + \max_t(tide_g))\end{aligned}\quad (4.12)$$

Next we apply a nearest-neighbors model to interpolate theoretical water heights to locations that did not get wet within the simulation. While these “water table” heights will be lower than the simulated elevation at those locations, our elevation data is assigned to properties at a finer scale of resolution than that of our surge model (see Section 4.6.1.3). Thus, each 0.125° grid cell contains capital at a range of elevations. Using this “bathtub” model approach to interpolate water heights allows us to capture inundation of properties that may be situated at the lower end of these distributions.

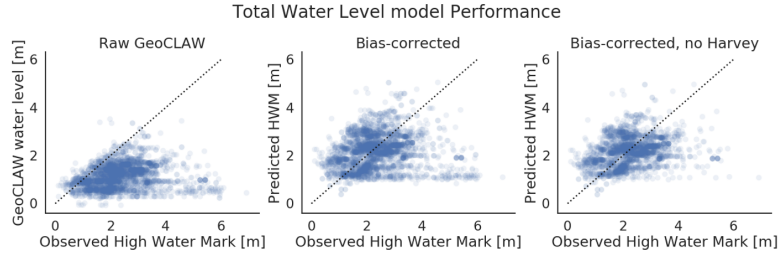


Figure 4.10: **Effect of statistical wave runup adjustment.** This figure compares observed high water marks to predicted water levels across the same seven historical hurricanes in Figure 4.9. A notable downward bias is observed prior to adjustment using 4.9. Hurricane Harvey is responsible for the majority of the under-predicted outliers due to the large rainfall-driven component to its observed values.¹⁰

4.6.1.3 Exposure Estimation

To estimate the present-day value of physical capital at each grid cell in our model, we leverage a property-level dataset containing location, estimated value, last sale price, and/or tax assessed value for each of 156 million properties nationwide (ATTOM Data Solutions 2019). We simultaneously estimate county-level yearly scaling factors that describe the change over time in building stock and building value-per-unit, using the House Price Index (HPI) (*House Price Index — Federal Housing Finance Agency*) and population data from the U.S. Census (*Population Datasets*). Because the property-level dataset only contains estimated present-day value for a subset of residential properties, we use the time series of scaling factors along with observed relationships between present-day value and sale price

or tax assessed value in a given year to fill in missing estimates. We then assign elevations to each property using the National Map. To reduce computational burden elsewhere in our analysis, we then aggregate building value to our 0.125° analysis grid, binning the results by 10cm elevation bins. We also add further bin dimensions corresponding to the nearest operating NOAA tide gauge (see Section 4.6.1.2), the nearest gauge in the SLR forecast dataset from Kopp, Horton, et al. 2014 (K14, see Section 4.6.2.3), and the ID of any levee that may protect this property (see Section 4.6.1.4). These additional dimensions facilitate operations in other components of the analysis.

Estimating the change in exposure over time

We assume that temporal change in physical capital within an arbitrarily defined area is the product of two independent changes – the average building price, P and the total building stock V . This gives us the simple equation for estimating the value within grid cell i that lies in county c , presuming we know information about P and V at the county level and that we know the capital value E at a baseline time t_0 :

$$E_{i,c,t} = E_{i,c,t_0} \frac{P_{c,t}}{P_{c,t_0}} \frac{V_{c,t}}{V_{c,t_0}} \quad (4.13)$$

As a proxy for the second term, we use HPI, which calculates scaled indices using the change in sale price across repeat home sales. Bogin, Doerner, and Larson 2019 interpolate this data to the county level from 1975-2015, which is the dataset we use. Notably, while HPI measures change in the consistent building stock, it does not capture *new* construction that is at a different value than the existing stock. It also only incorporates residential prices. Nevertheless, we believe it more plausibly resembles time trends of average building value at a local level than the national-level proxies that have been used in previous studies. Such previous approaches include a fixed assumed annual inflation value (Gettelman et al. 2018), the product of inflation and wealth metrics (Roger A. Pielke et al. 2008; Roger A. Pielke and Landsea 1998; Collins and Lowe 2001), or GDP (implicitly assumed to be per-capita) (Mendelsohn et al. 2012; W. D. Nordhaus 2010). As noted in Bogin, Doerner, and Larson 2019, the HPI dataset is able to capture differential trends including the relatively high suburban prices prior to 1990 and the relatively higher urban prices since that point. Occasionally, HPI values are missing for certain county-by-year observations. We interpolate observations assuming constant exponential growth. For smaller counties that are missing for the entire dataset, we use the state-level values. To extend the start of the dataset from 1975 to 1950 and the end of the dataset from 2015-2019, we fit a constant growth rate model at the state level. We then extrapolate the 1975 county-level HPI observations back to 1950 and 2015 county-level HPI observations up to 2019.

As a proxy for the third term, we use estimates of county-level annual population from the U.S. Census Bureau. This is similar to the proxy used in Roger A. Pielke et al. 2008; Roger A. Pielke and Landsea 1998, and, implicitly thru the use of national GDP, in Mendelsohn et al. 2012. Collins and Lowe 2001 use interpolated decennial measurements of housing units, an

alternative proxy for the same concept. We interpolate and/or extrapolate missing values following the HPI approach described above.

We divide both the HPI and population scaling factors by their 2019 values to set a baseline year, t_0 . Thus, our estimate of Equation 4.13 is:

$$\hat{E}_{i,c,t} = E_{i,c,2019} \frac{\hat{HPI}_{c,t}}{\hat{HPI}_{c,2019}} \frac{\hat{POP}_{c,t}}{\hat{POP}_{c,2019}} \quad (4.14)$$

Estimating present-day property values

While the ATTOM Data Solutions 2019 dataset is geographically complete, the estimated present-day value that we need to construct an exposure dataset is present only for a subset of residential buildings. However, relevant data on previous sale prices and/or tax assessments can help us fill in the missing values. Here we take a hierarchical approach to estimating the present-day value of each property following the list below, where the highest ranked estimate that is available for a given property is chosen.

1. Where available, we use ATTOM's Automated Valuation Model (AVM) estimated present-day value.
2. For residential properties where sale price from the most recent sale is available, we scale it to 2019 using our county-level HPI panel, dropping sale prices under \$1k and cases where the assessment value is $> 100x$ the sale price. We then multiply by the mean ratio of AVM price to 2019-scaled sale price across all properties in the same county that have observations for both values.
3. When a tax assessment value is available for a residential property, we follow a similar procedure as above. However, in the final step we scale by the county-averaged ratio of AVM prices to 2019-adjusted *tax assessed values* (rather than to 2019-adjusted sale prices).
4. For other residential properties that fall within a county that has some AVM estimates, we take the mean AVM estimate from that county.
5. For all other residential properties, we assign the mean AVM value from the full dataset.
6. For non-residential properties, we follow the same procedures as (2) and (3), without the sale price-to-AVM scaling step.
7. For remaining non-residential properties, we assign the mean 2019-scaled sale price for all non-residential properties in the same county.
8. For the remaining non-residential properties, we assign the mean 2019-scaled sale price for all non-residential properties in the full dataset.

The result is an estimate of present-day value for all 156 million properties in the original dataset.

Assigning elevation and binning to reduce computational burden

We next map each geolocated property to an elevation using the 1/3 arc-second (10m) National Map DEM, flag whenever a property falls in a levee-protected region as identified by the National Levee Database (*National Levee Database*), and identify the nearest tide gauge in both the NOAA archive and the K14 SLR dataset. In theory, we could calculate the wind and flood hazard at each individual property when estimating damages for each of our 120,000 simulated storms. However, our hazard models are only run at 0.1° so there is no additional information gained from the increased horizontal resolution provided by calculating property-level exposure. We greatly reduce computational complexity by aggregating the property-level dataset to a $0.1^\circ \times 0.1^\circ \times 10\text{cm}$ grid. Here we make the assumption that 10cm is within the range of the combined error introduced by imperfect surge modeling and imperfect DEMs, and thus binning at this level provides equally precise estimates while minimizing computational cost. In addition to horizontal and vertical dimensions, we further bin properties by ownership type, nearest NOAA and K14 gauges as well as levee status. Having binned across all of these dimensions, all properties within a given bin will receive the same modeled magnitude of hazard in our models of both historical and synthetic storms. Because our conceptual damage model (Equation 4.1) is linear in exposure, we can sum the property value in each bin in the following way:

$$\hat{D}_j := \hat{D}(h_j, E_j) = \gamma(h_j)^T \beta \sum_{i \in j} E_{i,t} \quad (4.15)$$

where i is a building in bin j .¹¹

Resulting exposure panel dataset

The result of the above pipeline is the first 4-dimensional dataset of physical capital in the continental United States, aggregated to 0.125° horizontal resolution, 10cm vertical resolution, and yearly temporal resolution. Additional dimensions bin capital along further categories.

4.6.1.4 Dose-Response Function Estimation

To estimate a dose-response function relating both wind speed and inundation depth to fractional losses, we rely on two datasets that together comprise the bulk of insurance claims from historical storms. The first, PCS from Verisk Analytics, contains private insurance claims matched to events from 1950 to present. In the U.S. most, but not all, private insurance does not cover flood damage. Thus, we expect this dataset to predominantly express

¹¹In an ongoing revision of this analysis, we allow for β to vary across building types (residential, commercial, and other).

insured wind-related damage. The second, from the National Flood Insurance Program, contains losses from storms beginning in 1978. While the vast majority of damage in this dataset is flood driven, it can sometimes be hard to distinguish when a building receives both wind- and flood-driven damage. Thus, it is possible that some flood damage exists in the PCS dataset and that some wind damage exists in the NFIP dataset.

One challenge of utilizing these claims datasets is the spatial mismatch between our independent variables of hazard and exposure, modeled at a 0.1° resolution, and the dependent variable of claim value, observed at state level. To address this, we follow an approach formalized in S. Hsiang and Jina 2014, in which the product of exposure and nonlinear features of hazard is calculated at the pixel level and then summed to the aggregation level of the outcome. A dose-response function is then fit at this aggregate level. Formally, this requires the assumptions that (a) damages are linear in those pixel-level products and (b) damages in different grid cells are independent:

$$f_t(h_j) := \gamma(h_j)^T \beta_t \quad (4.16)$$

$$D'_{st} := \sum_{j \in st} D'_j \quad (4.17)$$

where st indexes a given state, γ is a vector-valued function in $\mathbb{R}^2 \rightarrow \mathbb{R}^m$ creating nonlinear features of inundation depth and maximum wind speed, and β_t is a vector of coefficients in $\mathbb{R}^m$. D'_j is the (unobserved) damages at each grid cell j .

The first assumption must be true in the limit $m \rightarrow \infty$ as γ approaches a universal function approximator. With a limited dataset, our ability to fit parameters to a large number of features is limited, and thus this assumption may not hold in practice. However, given the uncertainty in the existing literature surrounding the functional form of depth-damage and wind-damage relationships, we argue that restricting ourselves to the set of functions linear in $\gamma(h_j)$ – even when m is small – is reasonable. All dose-response functions previously proposed in the literature may be represented this way. The second assumption may not hold exactly. For example, increasing wind damage to one building may cause shrapnel to fly into another building. However, at the larger resolution scales at which we are forecasting damages, we argue that these cases are less likely to occur.

For convenience, we refer to the scalar-vector product of hazard features and exposure, $\gamma(h_j)E_j$, as the “dose”, and the full inner product $f_t(h_j) = \gamma(h_j)E_j\beta$ as the “dose-response function”.

Substituting Equation 4.16 into Equation 4.1 and then into Equation 4.17 gives:

$$\begin{aligned}
 D'_{st} &= \sum_{j \in st} [(\gamma(h_j)^T \beta + \epsilon_j) E_j] \\
 &= \sum_{j \in st} [\gamma(h_j)^T E_{j,t}] \beta + \sum_{j \in st} \epsilon_j E_j \\
 &= \sum_{j \in st} [\gamma(h_j)^T E_{j,t}] \beta + \mu_{st}
 \end{aligned} \tag{4.18}$$

where $\mu_{st} = \sum_{j \in st} \epsilon_j$ is a residual term that follows a distribution equal to the weighted sum of the distributions $\epsilon_j(h_t)$ across all j in st , weighted by E_j .

This summation yields a state-level dataset where each row is a state-by-storm observation and where the columns include the dependent variable of damages and the concatenated “dose” vectors.

We estimate our dose-response coefficients, $\hat{\beta}$, using Ordinary Least Squares (OLS). We choose a low-order polynomial (quadratic in inundation depth and cubic in wind speed) that is consistent with energetic wind relationships (Bell et al. 2000; K. Emanuel 2005) and previously used depth curves (U.S. Army Corps of Engineers 2000; U.S. Army Corps of Engineers 2003). For consistency, we utilize the same time period for both the PCS and NFIP datasets. The regression on NFIP data is performed as described above, using a national aggregation level rather than a state level. Additionally, we drop Hurricane Katrina from the analysis as we are unable to reliably model the dynamics of failing levees. The resulting curves are shown in Figure 4.4.¹²

We validate our model by predicting annual historical losses from storms in the PCS and NFIP datasets using a 5-fold cross-validation approach. Results for predictions of PCS losses are shown in Figure 4.4b, indicating strong correlation of observed losses with our predicted values.

Another challenge in estimating total economic losses from a storm, where $loss_{economic} := loss_{insured} + loss_{uninsured}$, lies in estimating uninsured losses. There is little validated historical data on uninsured losses, and indeed the National Hurricane Center has often employed a simple “2x” multiplier when estimating economic loss from insured loss (Roger A. Pielke et al. 2008). We adopt this multiplier in our analysis, acknowledging that recently released NFIP claims and coverage data (U.S. Department of Homeland Security 2019) may provide more spatially granular information on coverage rates.¹³

¹²In ongoing work, we replace this OLS model with a cross-validated Generalized Linear Model (GLM), in which we explore alternative (Tweedie) error distributions that may more accurately represent observed residuals.

¹³In an ongoing revision of this analysis multiplier in our analysis, we incorporate this claim-level NFIP data, along with hazard-specific uninsured loss rates from the U.S. Congressional Budget Office (Congressional Budget Office 2019) to adjust the claims data prior to regression and thus improve our approach to scaling from insured to economic loss.

4.6.2 Projection of Future Damages

4.6.2.1 Generation of Synthetic Hurricane Tracks

To represent the link between climate and hurricane activity, we use the synthetic track simulator from K. Emanuel, Sundararajan, and Williams 2008. This physically based model uses observed relationships between large-scale environmental conditions, such as sea surface temperature and vertical wind shear, to parameterize a 1-dimensional physically based tropical cyclone model. While many analyses of the impacts of climate change through other mechanisms might rely on the outputs of GCMs to understand changes, two properties of hurricanes make the direct use of GCM output inappropriate. First, the dynamics that form hurricanes are of small enough scale that they are not adequately resolved in most GCMs used to forecast climate over century timescales; thus, their frequency and intensity are not well represented. Second, hurricanes are stochastic events, with damages driven by the most extreme storms. To accurately represent the tails of the damage distribution, it can be necessary to simulate many realizations of hurricane seasons.

A particular GCM in the Coupled Model Intercomparison Project 5 (CMIP5, Taylor, Stouffer, and Meehl 2012) archive will contain only a handful of realizations of future climates. The K. Emanuel, Sundararajan, and Williams 2008 model, on the other hand, can produce thousands of plausible storms from a single GCM and can estimate the frequency and intensity of these storms based on historical relationships between these parameters and large-scale environmental conditions. It is computationally efficient and has been used in numerous studies (K. Emanuel, Sundararajan, and Williams 2008; Marsooli et al. 2019; Mendelsohn et al. 2012; S. Hsiang and Jina 2014).

Here, we employ it to generate 200 storm tracks per emissions scenario, per year, and per GCM. In total, we examine two emissions scenarios: A moderate-emissions scenario (RCP 4.5) and a high-emissions scenario (RCP 8.5). We utilize seven GCMs: CCSM4 (Community Climate System Model, the University Corporation for Atmospheric Research); GFDL5 (Geophysical Fluid Dynamics Laboratory Climate Model, USA); HadGEM5 (Hadley Centre Global Environment Model, U.K. Meteorological Office); IPSL5 (Institut Pierre Simon Laplace Climate Model, France); MIROC (Model for Interdisciplinary Research on Climate, University of Tokyo, National Institute for Environmental Studies, Japan, and Japan Agency for Marine-Earth Science and Technology Frontier Research Center for Global Change); MPI5 (Max Planck Institute for Meteorology, Germany); and MRI5 (Meteorological Research Institute, Japan). And we examine four 11-year time periods: Present (2015-2025), near-term (2035-2045), medium-term (2055-2065), and end-of-century (2085-2095).

In order to bias-correct these GCMs and to examine historical climate-driven trends in hurricane risk, we also generate storm tracks for current and past climates using the NCEP/NCAR Reanalysis (Kalnay et al. 1996). We generate 400 storms per year for the periods 1979-1989 and 2008-2018. Additionally, we extend our earliest GCM time period to overlap with the 2008-2018 period, including the years 2008-2014 for each of the 14 combinations of GCM and RCP scenario.

In total, we generate 118,600 synthetic tracks that are then used as input into the same storm surge and wind models used for historical storms.

4.6.2.2 Hazard Estimation and Bias Correction

Our approach to estimating hazard from synthetic storms is largely identical to that used for historical storms, with one exception. In setting SL_{init} (Equation 4.10), we use median projections of local SLR from K14. Specifically, we use the same procedure as in Section 4.6.1.2 to identify nearby NOAA tide gauges for a given storm track. For tide gauges that are not included in the K14 dataset, we assign them the K14 SLR values from the nearest K14 gauge. We take the median SLR projection for the year in which the synthetic storm occurs for each gauge, and then add a time series of astronomical tides on top of this SLR projection. To estimate tides for future periods we use the `pytides` python package (*maritimeplanning/pytides: Tide prediction and analysis in python*). We then take the maximum tidal value at each gauge, add it to the SLR projection, and take the maximum across gauges as in Equation 4.10. Formally, this takes the form:

$$SL_{init,s,y} = \max_g(K14_{g,50,y} + \max_t(tide_{g,s})) \quad (4.19)$$

where s is a synthetic storm occurring in year y , $K14_{g,50,y}$ is a median K14 SLR projection at gauge g in year y , and $tide_{g,s}$ is the time series of astronomical tide predicted by `pytides`. The construction of $K14_{g,50,y}$ from the data reported in Kopp, DeConto, et al. 2017 is described in the following section.

Following the modeling of storm surge, at each grid cell we subtract the tidal signal and the local deviance in $K14_{g,50,y}$ for the nearest gauge, following the pattern we use for the tidal signal and $\delta_{SL_{init}}$ in Equation 4.11. The result of the above steps is a dataset of hazard magnitudes from synthetic storms. It can be represented as an N-dimensional array that is indexed by latitude, longitude, GCM, emissions scenario, year, storm number, and hazard type (surge or wind).

GCMs are known to display systematic biases relative to observations. A common approach is to “bias-correct” the GCMs by applying a correction function based on differences between the GCM predictions and reference observations during an overlapping time period. In our case, we treat the distribution of hazards across storms in the NCEP-based synthetic track set as “truth” because they provide a more fully characterized distribution than the short historical record. The data generating process, which relies on the hurricane simulator model, is also consistent with that used for the GCM projections. We then follow a process to bias correct both the frequency of predicted storms and the resulting spatial distributions of hazard, following the two-stage approach used in N. Lin et al. 2016.

Using the overlapping time period of 2008-2018, we first calculate a GCM-specific offset in the yearly frequency parameters estimated for the synthetic track sets driven by NCEP and by each of the GCMS. This defines a “frequency” bias correction. The hazard distributions resulting for this frequency correction step are shown in Figures 4.11 and 4.12 for the reference

period (2008-2018) for nine example cities. Differences exist across locations and across GCMs because the frequency correction is employed at the basin scale. Nevertheless, we see general agreement between the NCEP-derived curves and the bias-corrected GCM-derived curves. This indicates that the spatial structure of GCM-derived synthetic tracks largely mirrors that of NCEP-derived synthetic tracks over the reference period. Certain locations, including Fort Myers, FL and Norfolk, VA, show a greater variance across the GCMs and a greater deviance from the NCEP-derived return period curves. This suggests that these are areas in which the variance in large-scale environment present across GCMs has a greater impact on hurricane risk.

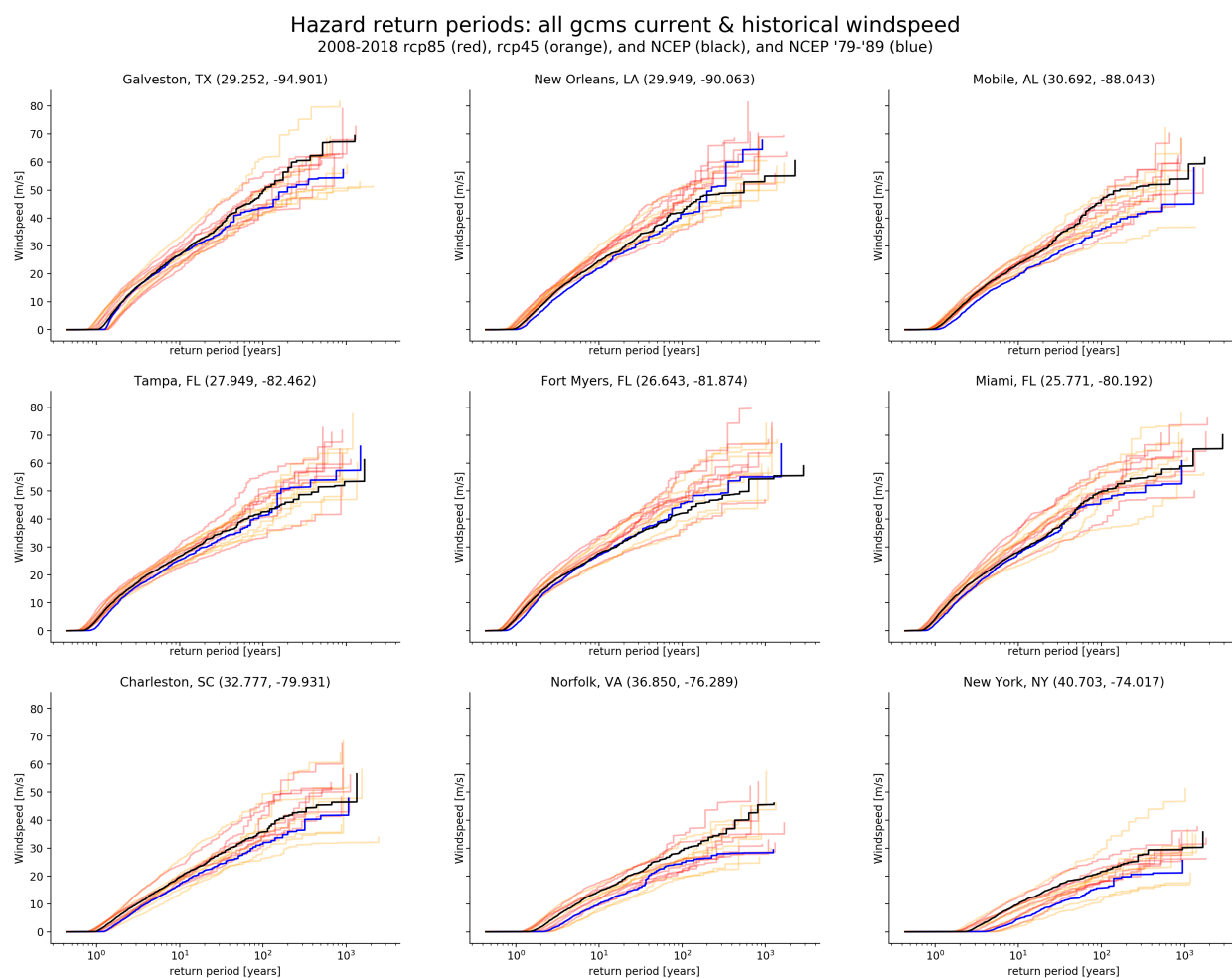


Figure 4.11: **Estimated s_{max} return period curves for 1979-1989 and 2008-2018 for nine example locations.** Each of the nine panels shows estimates of the return periods of maximum sustained wind speed at the centroid of a given city. There are seven red lines and seven orange lines, defining the curves as estimated for each of the 7 GCMs for RCP 8.5 and RCP 4.5, respectively. The black curve defines the return periods as estimated using the 2008-2018 NCEP-derived climatology and the blue curve defines the return periods as estimated using the 1979-1989 NCEP-derived climatology.

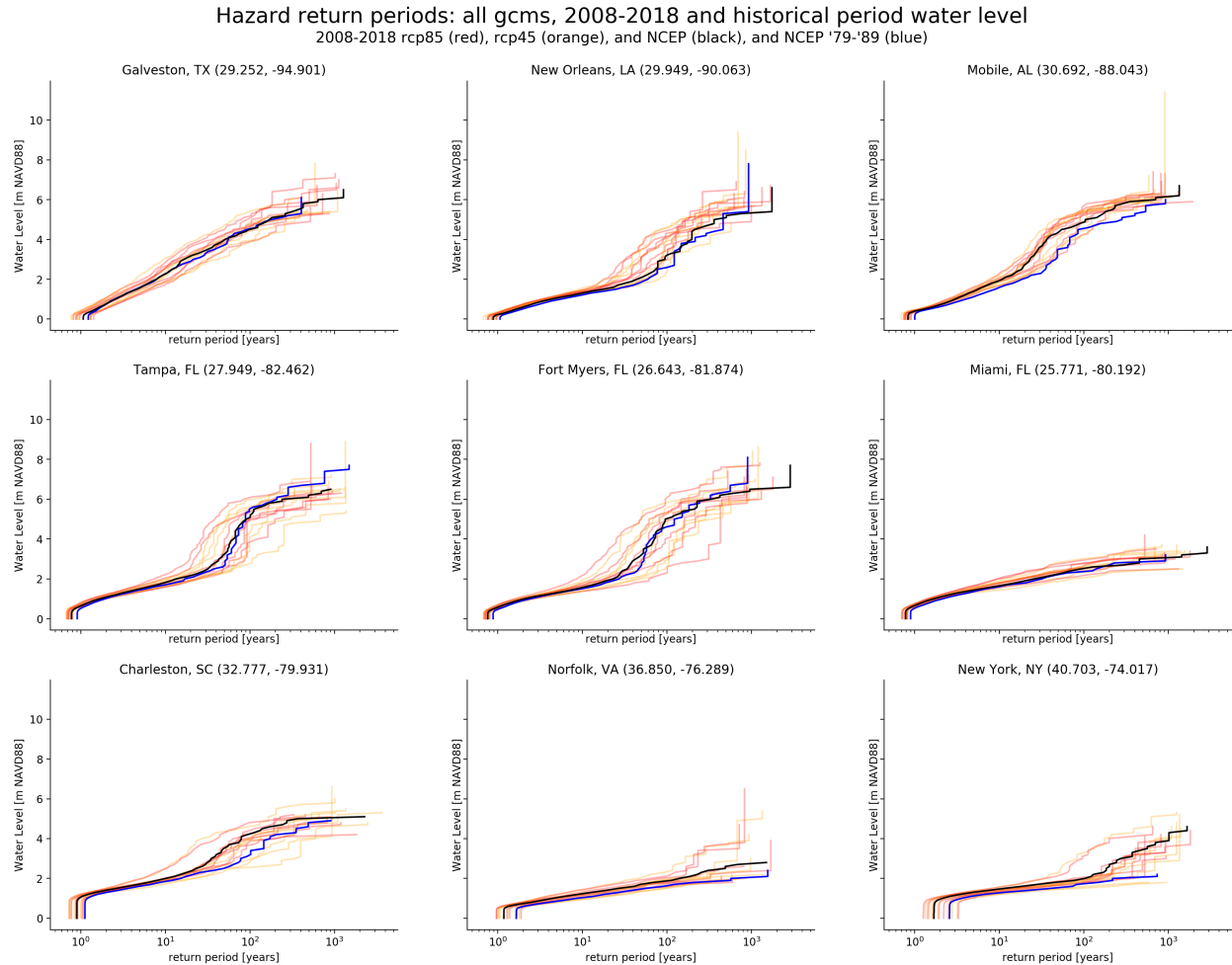


Figure 4.12: **Estimated TWL_{max} return period curves for 1979-1989 and 2008-2018 for nine example locations.** Each of the nine panels shows estimates of the return periods of hurricane-driven flood heights at the centroid of a given city. There are seven red lines and seven orange lines, defining the curves as estimated for each of the 7 GCMs for RCP 8.5 and RCP 4.5, respectively. The black curve defines the return periods as estimated using the 2008-2018 NCEP-derived climatology and the blue curve defines the return periods as estimated using the 1979-1989 NCEP-derived climatology.

We next employ a quantile-mapping approach to obtain a “hazard intensity” bias correction. Using the 4400 NCEP-derived tracks and the 2200 GCM-derived tracks for that 11 year period, we match hazards (TWL_{max} and s_{max}) at the pixel level by quantile. Taking the difference yields adjustment factors that we then propagate through to estimated hazards from the rest of the modeled years. For both frequency and hazard intensity corrections, we perform this adjustment independently for RCP 8.5 and RCP 4.5. This approach assumes that systematic biases in predicted hazards present in the overlapping time period persist as a linearly additive error in future periods. The result is to collapse the bias and variance across models seen in Figures 4.11 and 4.12 such that they match the NCEP-derived curves

at each location.

4.6.2.3 Incorporation of Sea-Level Rise Uncertainty

Substantial uncertainty surrounds future projections of both global and localized SLR (Kopp, Horton, et al. 2014; Kopp, DeConto, et al. 2017), and this uncertainty would affect the magnitude of coastal flooding experienced from a given hurricane. To incorporate this uncertainty, we generate estimates for each of 100 evenly-spaced quantiles from the gauge-level local K14 SLR projections. These are then re-centered to their median projections at the gauge-by-year level and added to the TWL estimates from the surge modeling described in the preceding section. In this way, we generate 100 equally likely estimates of future flood hazard from each of the synthetic storm tracks.

Several steps must be performed to utilize the K14 data. First, they must be brought into a consistent datum. K14 projections are relative to a 2000 CE baseline, whereas our estimates of building elevation are in NAVD88. We use a temporal Gaussian Process Model to fit a conversion between the MSL_{2000} datum of K14 and the MSL_{1982} datum of the National Tidal Datum Epoch (NTDE) and then use VDATUM (White et al. 2016) to convert from MSL_{1982} to NAVD88. Second, we must interpolate between the quantiles reported in the paper (0.5th, 5th, 17th, 50th, 83rd, 95th, 99.5th, and 99.9th). This is performed via linear interpolation.¹⁴

For the NCEP-based synthetic tracks, historical linear trends at each NOAA tide gauge in the K14 dataset were used to estimate historical MSL in a consistent manner (*Sea Level Trends - NOAA Tides & Currents*). No uncertainty was assumed for these historical local sea level values.

4.6.2.4 Fractional Loss Estimation

To translate hazard values into a distribution of fractional losses for each GCM, we first apply the dose-response function from Section 4.6.1 to our hazard array from the synthetic track set. Next, we sample a large number of fictitious hurricane “seasons” (100) from the available tracks within each GCM/scenario/year group.

The second step is accomplished by using the estimated “number of storms per year” parameter that emerges from the K. Emanuel, Sundararajan, and Williams 2008 model. Following Mendelsohn et al. 2012, Marsooli et al. 2019, and others that have used this stochastic hurricane model, we draw from a Poisson distribution with mean parameter equal to this model-generated value. We do this 100 times, providing a value for the number of storms by GCM, scenario, year, and sample. For each sample, we then draw that number of

¹⁴This approach implicitly assumes that all Atlantic and Gulf coast locations experience the same quantile of SLR. This may be valid at national scales; however, in an ongoing analysis of global TC damages, we directly ingest Monte Carlo draws from Kopp, DeConto, et al. 2017 rather than interpolating from the summary statistics reported in the paper. Such an approach can better represent the spatial correlation structure between gauges.

storms from the fractional loss array and sum the fractional losses across all of the storms that occur within each season. The output of this step is 100 realizations of fractional loss for each GCM/scenario/year/SLR quantile. Assuming each year within an 11-year time period corresponds to independent draws from the same climatology, we then aggregate across years to give 1100 realizations of hurricane seasons for each GCM/scenario/period/SLR quantile.

4.6.3 Risk Analysis and Climate Impact Assessment

We combine these estimated fractional losses with modeled exposure to generate estimates of economic risk in future climates. To do so, we assume a constant and spatially homogeneous exposure growth rate of 0.9% per year, following Neumann, K. Emanuel, et al. 2015. We apply this homogeneous growth to our exposure map from Section 4.6.1.3.¹⁵

This approach provides an estimated exposure map for each year of the 21st century. We then take the product of these exposure maps and the fractional loss estimates from the preceding section. We compare differences in the AAL and the median loss across emissions scenarios, aggregating losses to CBSAs. As a baseline scenario, we assume that the NCEP 2008-2018 fractional losses persist through the century as exposure evolves. To calculate the costs of anthropogenic climate change on hurricane risk, we then difference the national AAL of each scenario from that of the NCEP losses, interpolate between the centerpoint of the analysis periods, and apply a 3% discount rate to that time series.¹⁶ Accumulating the discounted losses yields an estimate of the present discounted cost of changing hurricane activity under both RCP 4.5 and 8.5 scenarios. Together, this approach takes the following form:

$$NPC_{s,m} = \sum_y \left[\sum_i \left[\left(\overline{\gamma(h_{i,y,s,m})} - \overline{\gamma(h_{i,2008-2018,NCEP})} \right) E_i e^{(0.009-0.03)y} \right] \right]^T \beta \quad (4.20)$$

where $\overline{\gamma(h_{i,y,scen,model})}^T \beta$ is the mean fractional loss estimated across the 1100 simulated seasons for grid cell i , year y , emissions scenario s , and GCM m . This is interpolated between the analysis interval midpoints of 2020, 2040, 2060, and 2090.

¹⁵Future work will extend the use of growth projections to incorporate other scenarios, including the Shared Socioeconomic Pathways (O'Neill et al. 2017). It will also incorporate a dynamic adaptation model from Diaz 2016

¹⁶Ongoing analysis examines the effect of assuming a range of plausible discount rates

Chapter 5

Implications and Future Work

In general, these studies demonstrate how to augment analytical tools that leverage historical relationships to understand the impact of potential actions on global welfare. The actions they target have widespread implications, from limiting the negative consequences of climate change to saving lives in an unprecedented pandemic. They address challenges for which appropriate historical data does not exist, creating measurements through alternative sources or leveraging our understanding of fundamental processes to supplant direct observation. In the two application areas studied, the results describe substantially larger costs of inaction or, equivalently, substantially larger benefits of action, than were previously known. In Chapter 3 we see this borne out in the number of infections averted by anti-contagion policies, while in Chapter 4 we see this borne out in the magnitude of coastal climate damages from changing hurricane patterns.

While two observations does not describe a robust pattern, it is likely that these results are not isolated findings among such challenging estimation environments. If true, this indicates that by focusing on the problems that can more directly be solved by traditional methods, we may underestimate the true consequences of inaction. This is a classic example of the streetlight effect, in which we look for a pair of lost keys only underneath a streetlight, because that's the one place we are able to search. The goal of this dissertation has been to expand the scope of our search by leveraging new sources of light, or new interdisciplinary approaches, such that we are better able to account for the universe of impacts, positive and negative, of large-scale policies.

In any thesis developing new data sources and modeling methodologies, it is worth recognizing that improved quantification of benefits and costs is not an inherently positive contribution in isolation. The same technologies that can be used to better target subnational and international aid can also be used to optimize oppression and marginalization. My hope is that the approaches developed here can provide evidence for the marginalized groups where fewer resources are expended to monitor societal health and welfare. However, this contribution will only be realized through continued effort to ensure that these tools are further developed and applied in appropriate contexts. For this reason, I am furthering the reach of each project described here in ongoing work. Many of these future efforts are

described within the chapters themselves, but I provide concise summaries below.

5.1 Socioenvironmental prediction with satellite imagery and machine learning

The work described in this thesis represents a proof of concept of a tool for expanding the usability of powerful machine learning techniques. It shows that global predictions of arbitrary metrics can be generated through the use of satellite imagery and simple algorithms. It also shows that the scalability and speed of these algorithms facilitates more thorough “stress testing” of predictive skill than is possible for more complex neural network approaches. For broad indicators of societal welfare, such as those used to assess progress toward the Sustainable Development Goals (SDGs), a comprehensive global dataset of these metrics would create an invaluable resource for improving the impact of targeted aid. Numerous steps must be taken to reach a point where the approaches developed in this dissertation could be used in such a context, but ongoing work will begin to address these challenges. First, our team will evaluate the performance of our model in assessing temporal change – a fundamentally different problem than the cross-sectional prediction task assessed in this dissertation. Second, we will further develop a suite of standardized tests that can assess our confidence in the ability of the approach to predict a given metric. Third, we will introduce new regression models that may allow for better predictions under different distributions of data. Together, these approaches will allow for the scope of the model’s application to be expanded while maintaining confidence in the robustness of the results. Parallel efforts are also seeking to develop a set of best practices for using satellite-predicted data in causal inference settings. New bias and covariance structures can be introduced when treating satellite-based predictions as truth, and we must be careful to account for this possibility in experimental designs.

5.2 Econometric-epidemiological prediction

The paper presented in this dissertation was a first step in exploring the use of econometric approaches in contagious disease modeling, but a far tighter coupling to more sophisticated epidemiological projections could allow for a new generation of rapid response models. This form of model would leverage statistical techniques for early assessment of anti-contagion policy impacts, allowing for more efficient and early action to control pandemic growth. Yin, Z. Liu, and Kammen 2020 has developed an age-specific model that provides a platform for understanding differential transmission rates across different population structures. Integrating empirically identified policy impacts into a more representative transmission model would help reduce the uncertainty that has thus far hampered efforts to coordinate national and global policy. Additional efforts are underway to extend the modeling methodology of Chapter 3 to the global scale using a novel dataset of non-pharmaceutical interventions

across 186 countries. It is unclear how long this disease will continue to affect the world, but the longer it lasts the more crucial it becomes to understand exactly which containment policies offer the most health benefits while enacting the least societal costs.

5.3 Impacts of climate change on extreme event damages

While Chapter 4 explores the impact of climate-driven hurricane trends in the context of the United States, a substantial burden of tropical cyclones exists along coastlines in many other ocean basins. Ongoing work extends the methodology to leverage globally available datasets and seeks to use the research outputs to contribute to a quantification of the Social Cost of Carbon (SCC). This quantity, which defines the marginal societal cost of a ton of emitted carbon, contributes to benefit-cost analyses in formal rule-making contexts. A large body of work exists to estimate this value, yet little accounts for the societal impact of climate on extreme event distributions in any physically representative way. If the theme of this dissertation's findings persists – namely that under-measured impacts may correlate with under-predicted impacts – a better accounting for extreme events may substantially improve our SCC estimation.

Another effort will seek to leverage the advancements brought about by the methods described in this dissertation to understand the unequal impacts of changing hurricane patterns across varying populations. The results presented in Chapter 4 demonstrate that projected future hurricane risks exhibit spatial patterns across the Atlantic and Gulf coasts. Knowing how these impacts are distributed along socioeconomic and demographic dimensions can further aid the generation of effective adaptation policy by targeting resources to communities that are both disadvantaged and substantially affected by these changing risks. A third research trajectory will pair the methods of Chapter 4 with alternative indirect outcomes, improving our understanding of the impacts of projected hurricane changes on health burden and long-term economic growth.

Simultaneous to these new research endeavors, a number of ongoing partnerships will continue to incorporate the U.S. forecasts in public and private decision-making contexts. First Street Foundation 2020 ingests the coastal flood forecasts of Chapter 4 into a publicly available dataset of property-level flood risk. Hess et al. 2019 applies the methodology of Chapter 4 to analyze local hurricane risks for New Jersey state policymakers. And Blackrock Investment Institute 2019 uses the risk information from Chapter 4 to assess risk to broad investment portfolios. Each of these stakeholders, from the public to the government to large financial institutions, represents a platform through which quantitative impact analyses can incentivize better climate change mitigation and adaptation. By better identifying the contribution of extreme events to this impact, we can better align these incentives and provide evidence for effective and efficient climate action.

Bibliography

ADCIRC. URL: <https://adcirc.org/>.

Agarwal, Alekh et al. (2014).

“Least squares revisited: Scalable approaches for multi-class prediction.”

In: *International Conference on Machine Learning*, pp. 541–549.

Agence Régionale de Santé. *Agence Régionale de Santé: Agir pour la santé de tous*.

URL: <https://www.ars.sante.fr/>.

Alber, Maximilian et al. (2017).

“An empirical study on the properties of random bases for kernel methods.”

In: *Neural Information Processing Systems*.

URL: <https://papers.nips.cc/paper/6869-an-empirical-study-on-the-properties-of-random-bases-for-kernel-methods.pdf>.

Alkama, R. and A. Cescatti (Feb. 2016).

“Biophysical climate impacts of recent changes in global forest cover.”

In: *Science* 351.6273, pp. 600–604. ISSN: 0036-8075. DOI: 10.1126/science.aac8083.

URL: <http://www.ncbi.nlm.nih.gov/pubmed/26912702><http://www.sciencemag.org/cgi/doi/10.1126/science.aac8083>.

Amazon Web Services (2018). *Terrain Tiles*.

URL: <https://registry.opendata.aws/terrain-tiles/>.

Anderson, Roy M and Robert M May (1992).

Infectious diseases of humans: dynamics and control. Oxford university press.

Angrist, Joshua D and Jörn-Steffen Pischke (2008).

Mostly {H}armless {E}conometrics: {A}n {E}mpiricist’s {C}ompanion.

Princeton University Press.

Athey, Susan (Feb. 2017). “Beyond prediction: Using big data for policy problems.”

In: *Science* 355.6324, 483 LP –485. DOI: 10.1126/science.aal4321.

URL: <http://science.sciencemag.org/content/355/6324/483.abstract>.

ATTOM Data Solutions (2019). *Real Estate Property Data*.

URL: <https://www.attomdata.com/>.

Bedi, Tara, Aline Coudouel, and Kenneth Simler, eds. (2007). *More than a pretty picture:*

Using poverty maps to design better policies and interventions.

Washington, DC: The World Bank. ISBN: 0-8213-6932-6.

DOI: 10.1596/978-0-8213-6931-9.

- URL: <https://openknowledge.worldbank.org/bitstream/handle/10986/6800/414470PAPER0Pr1010official0Use00only1.pdf?sequence=1&isAllowed=y>.
- Bell, Gerald D. et al. (June 2000). "Climate Assessment for 1999."
In: *Bulletin of the American Meteorological Society* 81.6, S1–S50. ISSN: 0003-0007.
DOI: 10.1175/1520-0477(2000)81[S1:CAF]2.0.CO;2.
- Bhatia, Kieran et al. (2018). "Projected Response of Tropical Cyclone Intensity and Intensification in a Global Climate Model." In: *Journal of Climate* 31, pp. 8281–8303.
- Blackrock Investment Institute (2019).
Getting physical: assessing climate risks — BlackRock. Tech. rep.
New York, NY: Blackrock.
URL: <https://www.blackrock.com/us/individual/insights/blackrock-investment-institute/physical-climate-risks>.
- Blumenstock, Joshua (Sept. 2018).
Don't forget people in the use of big data for development.
DOI: 10.1038/d41586-018-06215-5.
URL: <http://www.nature.com/articles/d41586-018-06215-5>.
- Bogin, Alexander, William Doerner, and William Larson (June 2019).
"Local House Price Dynamics: New Indices and Stylized Facts."
In: *Real Estate Economics* 47.2, pp. 365–398. ISSN: 15406229.
DOI: 10.1111/1540-6229.12233.
- Bolliger, Ian et al. (Oct. 2017). "Ground Control to Major Tom: the importance of field surveys in remotely sensed data analysis." In: *Data For Good Exchange*.
URL: <https://arxiv.org/abs/1710.09342>.
- Bootsma, Martin C J and Neil M Ferguson (May 2007). "The effect of public health measures on the 1918 influenza pandemic in {U}. {S}. cities."
In: *Proceedings of the National Academy of Sciences* 104.18, pp. 7588–7593.
ISSN: 0027-8424. DOI: 10.1073/pnas.0611071104.
- Burke, Marshall, Solomon M Hsiang, and Edward Miguel (2015).
"Global non-linear effect of temperature on economic production."
In: *Nature* 527.7577, pp. 235–239.
- Camargo, Suzana J (Dec. 2013).
"Global and Regional Aspects of Tropical Cyclone Activity in the CMIP5 Models."
In: *Journal of Climate* 26.24, pp. 9880–9902. ISSN: 0894-8755.
DOI: 10.1175/JCLI-D-12-00549.1.
URL: <https://doi.org/10.1175/JCLI-D-12-00549.1>.
- Carleton, Tamma et al. (Aug. 2018). "Valuing the Global Mortality Consequences of Climate Change Accounting for Adaptation Costs and Benefits."
In: *NBER Working Paper*. DOI: 10.2139/ssrn.3224365.
URL: <https://www.ssrn.com/abstract=3224365>.
- Carlson, Kimberly M et al. (Jan. 2018).
"Effect of oil palm sustainability certification on deforestation and fire in Indonesia."
In: *Proceedings of the National Academy of Sciences of the United States of America*

- 115.1, pp. 121–126. ISSN: 1091-6490. DOI: 10.1073/pnas.1704728114.
 URL: <http://www.ncbi.nlm.nih.gov/pubmed/29229857><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5776786>.
- Center for International Earth Science Information Network (CIESIN) (2016).
Gridded Population of the World, Version 4.
 DOI: <http://dx.doi.org/10.7927/H4NP22DQ>.
- Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.
COVID-19 Data Repository. URL: <https://github.com/CSSEGISandData/COVID-19>.
- Cheng, Gong, Junwei Han, and Xiaoqiang Lu (Oct. 2017).
 “Remote sensing image scene classification: Benchmark and state of the art.”
 In: *Proceedings of the IEEE* 105.10, pp. 1865–1883. ISSN: 0018-9219.
 DOI: 10.1109/JPROC.2017.2675998.
 URL: <http://ieeexplore.ieee.org/document/7891544/>.
- Cheriyadat, Anil M. (2014). “Unsupervised feature learning for aerial scene classification.”
 In: *IEEE Transactions on Geoscience and Remote Sensing* 52.1, pp. 439–451.
 ISSN: 01962892. DOI: 10.1109/TGRS.2013.2241444.
- Chinazzi, Matteo et al. (2020). “The effect of travel restrictions on the spread of the 2019 novel coronavirus ({COVID}-19) outbreak.” In: *Science*. ISSN: 0036-8075.
 DOI: 10.1126/science.aba9757.
- Chowell, Gerardo et al. (2016).
 “Mathematical models to characterize early epidemic growth: A review.”
 In: *Physics of Life Reviews* 18, pp. 66–97.
- Coates, Adam, Honglak Lee, and Andrew Y Ng (2011).
 “An Analysis of Single-Layer Networks in Unsupervised Feature Learning.”
 In: *14th International Conference on Artificial Intelligence and Statistics*. Vol. 15.
 Fort Lauderdale, FL.
 URL: <http://proceedings.mlr.press/v15/coates11a/coates11a.pdf>.
- Coates, Adam and Andrew Y. Ng (2012).
 “Learning Feature Representations with K-means.”
 In: *Neural networks: Tricks of the trade*.
 Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller.
 Springer, Berlin, Heidelberg. Chap. 22, pp. 561–580.
 DOI: 10.1007/978-3-642-35289-8_{_}30.
 URL: http://link.springer.com/10.1007/978-3-642-35289-8_30https://www-cs.stanford.edu/~acoates/papers/coatesng_nntot2012.pdf.
- Collins, Douglas J. and Stephen P. Lowe (2001).
 “A macro validation dataset for U.S. hurricane models.”
 In: *Casualty Actuarial Society Forum*. Arlington, VA.
- Congressional Budget Office (2019).
Expected Costs of Damage From Hurricane Winds and Storm-Related Flooding.
 Tech. rep. United States Congress. URL: www.cbo.gov/publication/55019.

COVID-19 pandemic in France (2020).

URL: https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_France&oldid=956505489.

COVID-19 pandemic in Iran (2020).

URL: https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_Iran&oldid=956402285.

COVID-19 pandemic lockdown in Hubei (2020).

URL: https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_lockdown_in_Hubei&oldid=955933271.

COVID-19 pandemic lockdown in Italy (2020).

URL: https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_lockdown_in_Italy&oldid=956053371.

Cox, Sam. *maritimeplanning/pytides: Tide prediction and analysis in python*.

URL: <https://github.com/maritimeplanning/pytides>.

Daniely, Amit, Roy Frostig, and Yoram Singer (2016). "Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity."

In: *Advances in Neural Information Processing Systems* Nips, pp. 2261–2269.

ISSN: 10495258.

De Sherbinin, A. M. et al. (2017). "Improved Mapping of Human Population and Settlements through Integration of Remote Sensing and Socioeconomic Data."

In: *AGUFM* 2017, IN51H-06.

Desmet, Klaus et al. (Aug. 2018). "Evaluating the Economic Cost of Coastal Flooding."

URL: <http://www.nber.org/papers/w24918>.

Diaz, Delavane B. (July 2016). "Estimating global damages from sea level rise with the Coastal Impact and Adaptation Model (CIAM)."

In: *Climatic Change* 137.1-2, pp. 143–156. ISSN: 0165-0009.

DOI: 10.1007/s10584-016-1675-4.

URL: <http://link.springer.com/10.1007/s10584-016-1675-4>.

Dinan, Terry (Aug. 2017). "Projected Increases in Hurricane Damage in the United States: The Role of Climate Change and Coastal Development."

In: *Ecological Economics* 138, pp. 186–198. ISSN: 0921-8009.

DOI: 10.1016/J.ECOLECON.2017.03.034.

URL: <https://www.sciencedirect.com/science/article/pii/S0921800916309752>.

Emanuel, Kerry (Aug. 2005).

"Increasing destructiveness of tropical cyclones over the past 30 years."

In: *Nature* 436.7051, pp. 686–8. ISSN: 1476-4687. DOI: 10.1038/nature03906.

URL: <http://www.ncbi.nlm.nih.gov/pubmed/16056221>.

— *Anthropogenic Effects on Tropical Cyclone Activity*.

URL: <http://wind.mit.edu/~emanuel/anthro.html>.

Emanuel, Kerry, Raghoth Sundararajan, and John Williams (Mar. 2008).

"Hurricanes and Global Warming: Results from Downscaling IPCC AR4 Simulations."

In: *Bulletin of the American Meteorological Society* 89.3, pp. 347–368. ISSN: 0003-0007.

- DOI: 10.1175/BAMS-89-3-347.
 URL: <http://journals.ametsoc.org/doi/10.1175/BAMS-89-3-347>.
- Ferguson, Neil M et al. (Mar. 2020). "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand." London, UK. URL: <https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf>.
- Fink, Larry (2020). *A Fundamental Reshaping of Finance*. Tech. rep. Blackrock, Inc. URL: <https://www.blackrock.com/corporate/investor-relations/larry-fink-ceo-letter?cid=ppc:CEOLetter:PMS:US:NA>.
- Firat, Orhan, Gulcan Can, and Fatos T.Yarman Yarman Vural (2014). "Representation learning for contextual object and region detection in remote sensing." In: *Proceedings - International Conference on Pattern Recognition*, pp. 3708–3713. ISSN: 10514651. DOI: 10.1109/ICPR.2014.637.
- First Street Foundation (2020). *The First National Flood Risk Assessment: Defining America's Growing Risk*. Tech. rep. URL: https://assets.firststreet.org/uploads/2020/06/first_street_foundation_first_national_flood_risk_assessment.pdf.
- Fisman, David, Edwin Khoo, and Ashleigh Tuite (2014). "Early epidemic dynamics of the West African 2014 Ebola outbreak: estimates derived with a simple two-parameter model." In: *PLoS currents* 6.
- Flaxman, Seth et al. (2020). *Report 13: Estimating the number of infections and the impact of non-pharmaceutical interventions on {COVID}-19 in 11 {E}uropean countries*. Tech. rep. Imperial College London.
- Garratt, J. R. (July 1977). *Review of Drag Coefficients over Oceans and Continents*. Tech. rep. 7, pp. 915–929. DOI: 10.1175/1520-0493(1977)105<0915:RODC00>2.0.CO;2. URL: <http://journals.ametsoc.org/mwr/article-pdf/105/7/915/4165781/1520-0493>.
- Gechter, Michael et al. (2018). "The Welfare Consequences of Formalizing Developing Country Cities: Evidence from the Mumbai Mills Redevelopment." In: *Working paper*.
- Gesch, Dean B, Michael J Oimoen, and Gayla A Evans (2014). *Accuracy Assessment of the U.S. Geological Survey National Elevation Dataset, and Comparison with Other Large-Area Elevation Datasets-SRTM and ASTER*. Tech. rep. Reston, Virginia: United States Geological Survey. DOI: 10.3133/ofr20141008. URL: <http://dx.doi.org/10.3133/ofr20141008>.
- Gettelman, A. et al. (2018). "Projections of future tropical cyclone damage with a high-resolution global climate model." In: *Climatic Change* 146.3-4, pp. 575–585. ISSN: 0165-0009. DOI: 10.1007/s10584-017-1902-7. URL: <http://link.springer.com/10.1007/s10584-017-1902-7>.
- Gindelsky, Marina, Jeremy Moulton, and Scott Wentland (2019). "Valuing Housing Services in the Era of Big Data: A User Cost Approach Leveraging Zillow Microdata." In: *Big Data for 21st Century Economic Statistics*. Ed. by NBER. University of Chicago Press.

- Glenn, Ezra Haber (2019). *acs: Download, Manipulate, and Present American Community Survey and Decennial Data from the US Census*.
- Google Developers. *Maps Static API*.
URL: <https://developers.google.com/maps/documentation/maps-static/intro>.
- Greene, William H (2003). *Econometric {A}nalysis*. Prentice Hall.
- Haack, Barry and Robert Ryerson (Mar. 2016). "Improving remote sensing research and education in developing countries: Approaches and recommendations." In: *International Journal of Applied Earth Observation and Geoinformation* 45, pp. 77–83.
ISSN: 0303-2434. DOI: 10.1016/J.JAG.2015.11.003.
URL: <https://www.sciencedirect.com/science/article/pii/S0303243415300477>.
- Hallegatte, Stéphane (Nov. 2007). "The use of synthetic hurricane tracks in risk analysis and climate change damage assessment."
In: *Journal of Applied Meteorology and Climatology* 46.11, pp. 1956–1966.
ISSN: 15588424. DOI: 10.1175/2007JAMC1532.1. URL: <http://www.nhc.noaa.gov/>.
- Hansen, M C et al. (Nov. 2013).
"High-resolution global maps of 21st-century forest cover change."
In: *Science (New York, N.Y.)* 342.6160, pp. 850–3. ISSN: 1095-9203.
DOI: 10.1126/science.1244693.
URL: <http://www.ncbi.nlm.nih.gov/pubmed/24233722>.
- Hatchett, Richard J, Carter E Mecher, and Marc Lipsitch (May 2007). "Public health interventions and epidemic intensity during the 1918 influenza pandemic."
In: *Proceedings of the National Academy of Sciences* 104.18, pp. 7582–7587.
ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0610941104.
- He, Kaiming et al. (2016). "Deep residual learning for image recognition."
In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.
- Head, Andrew et al. (2017).
"Can human development be measured with satellite imagery?" In: *ICTD*, pp. 1–8.
- Hess, Hannah et al. (2019). *New Jersey's Rising Coastal Risk*. Tech. rep.
Oakland, CA: Climate Impact Lab. URL: <https://njcoastalrisk.com>.
- Holland, Paul W. (Dec. 1986). "Statistics and Causal Inference."
In: *Journal of the American Statistical Association* 81.396, p. 945. ISSN: 01621459.
DOI: 10.2307/2289064.
URL: <https://www.jstor.org/stable/2289064?origin=crossref>.
- House Price Index — Federal Housing Finance Agency. URL:
<https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx>.
- HOUSER, TREVOR et al. (Aug. 2015). *Economic Risks of Climate Change*.
Columbia University Press. DOI: 10.7312/hous17456.
URL: <http://www.jstor.org/stable/10.7312/hous17456>.
- Hsiang, Solomon (Oct. 2016). "Climate Econometrics."
In: *Annual Review of Resource Economics* 8.1, pp. 43–75. ISSN: 1941-1340.
DOI: 10.1146/annurev-resource-100815-095343.

- URL: <http://dx.doi.org/10.1146/annurev-resource-100815-095343>
<http://www.annualreviews.org/doi/pdf/10.1146/annurev-resource-100815-095343>
<http://www.annualreviews.org/doi/10.1146/annurev-resource-100815-095343>.
- Hsiang, Solomon M. (Aug. 2010). "Temperatures and cyclones strongly associated with economic production in the Caribbean and Central America." In: *Proceedings of the National Academy of Sciences of the United States of America* 107.35, pp. 15367–15372. ISSN: 00278424. DOI: 10.1073/pnas.1009510107.
 URL: www.pnas.org/cgi/doi/10.1073/pnas.1009510107.
- Hsiang, Solomon, Daniel Allen, et al. (June 2020).
 "The effect of large-scale anti-contagion policies on the COVID-19 pandemic."
 In: *Nature*, pp. 1–9. ISSN: 14764687. DOI: 10.1038/s41586-020-2404-8.
 URL: <https://www.nature.com/articles/s41586-020-2404-8>.
- Hsiang, Solomon and Amir Jina (July 2014). "The Causal Effect of Environmental Catastrophe on Long-Run Economic Growth: Evidence From 6,700 Cyclones."
 In: *NBER Working Paper*. DOI: 10.3386/w20352.
 URL: <http://www.nber.org/papers/w20352.pdf>.
- Hu, Wenjie et al. (2019). "Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery."
 In: *Conference on Artificial Intelligence, Ethics, and Society*. Honolulu, HI. URL: http://www.aies-conference.com/wp-content/papers/main/AIES-19_paper_157.pdf.
- IHME COVID-19 health service utilization forecasting team and Christopher JL L Murray (2020). "Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months."
 In: *medRxiv* 114, p. 2020.03.27.20043752. DOI: 10.1101/2020.03.27.20043752. URL: <http://medrxiv.org/content/early/2020/03/30/2020.03.27.20043752.abstract>.
- Inglada, Jordi et al. (2017). "Operational high resolution land cover map production at the country scale using satellite image time series." In: *Remote Sensing* 9.1, p. 95. ISSN: 2072-4292. DOI: 10.3390/rs9010095.
- Jean, Neal, Marshall Burke, et al. (Aug. 2016).
 "Combining satellite imagery and machine learning to predict poverty." en.
 In: *Science* 353.6301, pp. 790–794. ISSN: 10959203. DOI: 10.1126/science.aaf7894.
 URL: <http://www.ncbi.nlm.nih.gov/pubmed/27540167>
<http://science.sciencemag.org/content/353/6301/790>
<http://science.sciencemag.org/content/sci/353/6301/790.full.pdf>.
- Jean, Neal, Sherrie Wang, et al. (2019).
 "Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data."
 In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 3967–3974. DOI: 10.1609/aaai.v33i01.33013967.
- Jonas, Eric et al. (2018). "Flare Prediction Using Photospheric and Coronal Image Data."
 In: *Solar Physics* 293.3, pp. 1–22. ISSN: 1573093X. DOI: 10.1007/s11207-018-1258-9.
 URL: <http://dx.doi.org/10.1007/s11207-018-1258-9>.

- Kalnay, E et al. (Mar. 1996). "The NCEP/NCAR 40-Year Reanalysis Project."
In: *Bulletin of the American Meteorological Society* 77.3, pp. 437–472. ISSN: 0003-0007.
DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.
URL: [https://doi.org/10.1175/1520-0477\(1996\)077%3C0437:TNYRP%3E2.0.CO%20http://0.0.0.2](https://doi.org/10.1175/1520-0477(1996)077%3C0437:TNYRP%3E2.0.CO%20http://0.0.0.2).
- Kandula, Sasikiran et al. (2018).
"Evaluation of mechanistic and statistical methods in forecasting influenza-like illness."
In: *Journal of The Royal Society Interface* 15.144, p. 20180174.
- Kantis, Caroline, Samantha Kiernan, and Jason Socrates Bardi (July 2020).
UPDATED: Timeline of the Coronavirus. URL:
<https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus>.
- Kermack, William Ogilvy and Anderson G McKendrick (1927).
"A contribution to the mathematical theory of epidemics."
In: *Proceedings of the royal society of london. Series A, Containing papers of a mathematical\ and physical character* 115.772, pp. 700–721.
- Kirezci, Ebru et al. (2020). "Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st Century." In: *Scientific Reports* 10.1, p. 11629.
ISSN: 2045-2322. DOI: 10.1038/s41598-020-67736-6.
URL: <https://doi.org/10.1038/s41598-020-67736-6>.
- Klotzbach, Philip J. et al. (July 2018). "Continental U.S. Hurricane Landfall Frequency and Associated Damage: Observations and Future Risks."
In: *Bulletin of the American Meteorological Society* 99.7, pp. 1359–1376.
ISSN: 0003-0007. DOI: 10.1175/BAMS-D-17-0184.1.
URL: <http://journals.ametsoc.org/doi/10.1175/BAMS-D-17-0184.1>.
- Knapp, Kenneth R et al. (Mar. 2010). "The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data."
In: *Bulletin of the American Meteorological Society* 91.3, pp. 363–376. ISSN: 0003-0007.
DOI: 10.1175/2009BAMS2755.1. URL: <https://doi.org/10.1175/2009BAMS2755.1>.
- Knutson, Thomas R. et al. (Mar. 2010). "Tropical cyclones and climate change."
In: *Nature Geoscience* 3.3, pp. 157–163. ISSN: 1752-0894. DOI: 10.1038/ngeo779.
URL: <http://www.nature.com/articles/ngeo779>.
- Koenig, Todd A et al. (2016). *Identifying and preserving high-water mark data*. Tech. rep., p. 47. DOI: 10.3133/tm3A24. URL: <https://pubs.er.usgs.gov/publication/tm3A24>.
- Kopp, Robert E., Robert M. DeConto, et al. (Dec. 2017). "Evolving Understanding of Antarctic Ice-Sheet Physics and Ambiguity in Probabilistic Sea-Level Projections."
In: *Earth's Future* 5.12, pp. 1217–1233. ISSN: 23284277. DOI: 10.1002/2017EF000663.
URL: <https://agupubs.pericles-prod.literatumonline.com/doi/full/10.1002/2017EF000663%20https://agupubs.pericles-prod.literatumonline.com/doi/abs/10.1002/2017EF000663%20https://agupubs.onlinelibrary.wiley.com/doi/10.1002/2017EF000663>.

- Kopp, Robert E., Radley M. Horton, et al. (Aug. 2014). “Probabilistic 21st and 22nd century sea-level projections at a global network of tide-gauge sites.”
In: *Earth’s Future* 2.8, pp. 383–406. ISSN: 23284277. DOI: 10.1002/2014EF000239.
URL: <http://doi.wiley.com/10.1002/2014EF000239>.
- Kossin, James P. (June 2018). “A global slowdown of tropical-cyclone translation speed.”
In: *Nature* 558.7708, pp. 104–107. ISSN: 0028-0836. DOI: 10.1038/s41586-018-0158-3.
URL: <http://www.nature.com/articles/s41586-018-0158-3>.
- Kossin, James P., Kerry A. Emanuel, and Gabriel A. Vecchi (May 2014).
“The poleward migration of the location of tropical cyclone maximum intensity.”
In: *Nature* 509.7500, pp. 349–352. ISSN: 0028-0836. DOI: 10.1038/nature13278.
URL: <http://www.nature.com/articles/nature13278>.
- Kraemer, Moritz U G et al. (2020). “The effect of human mobility and control measures on the {COVID}-19 epidemic in {C}hina.” In: *Science*. ISSN: 0036-8075.
DOI: 10.1126/science.abb4218.
- Krizhevsky, Alex and Geoffrey Hinton (2009).
Learning multiple layers of features from tiny images. Tech. rep. Citeseer.
URL: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012).
“Imagenet classification with deep convolutional neural networks.”
In: *Advances in neural information processing systems*, pp. 1097–1105.
- Kucharski, Adam J et al. (2020). “Early dynamics of transmission and control of {COVID}-19: a mathematical modelling study.” In: *The Lancet Infectious Diseases*.
- Kulp, Scott A and Benjamin H Strauss (2019). “New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding.”
In: *Nature Communications* 10.1, p. 4844. ISSN: 2041-1723.
DOI: 10.1038/s41467-019-12808-z.
URL: <https://doi.org/10.1038/s41467-019-12808-z>.
- Lee, Chia-Ying et al. (Jan. 2018).
“An Environmentally Forced Tropical Cyclone Hazard Model.”
In: *Journal of Advances in Modeling Earth Systems* 10.1, pp. 223–241. ISSN: 19422466.
DOI: 10.1002/2017MS001186. URL: <http://doi.wiley.com/10.1002/2017MS001186>.
- Lehman, Will and R Hasanzadeh Nafari (2016).
“An Empirical, Functional approach to Depth Damages.”
In: *FLOODrisk 2016 - 3rd European Conference on Flood Risk Management*.
DOI: 10.1051/05002.
- Li, Qun et al. (2020). “Early transmission dynamics in {W}uhan, {C}hina, of novel coronavirus-infected pneumonia.” In: *New England Journal of Medicine*.
- Li, Ruiyun et al. (2020). “Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus ({SARS}-{C}o{V}2).” In: *Science*.
- Lin, Jiabao. *COVID-19/2019-nCoV Time Series Infection Data Warehouse*.
URL: <https://github.com/BlankerL/DXY-COVID-19-Data>.

Lin, Ning et al. (Oct. 2016).

“Hurricane Sandy’s flood frequency increasing from year 1800 to 2100.”

In: *Proceedings of the National Academy of Sciences*. ISSN: 0027-8424.

DOI: 10.1073/PNAS.1604386113.

URL: <https://www.pnas.org/content/early/2016/10/06/1604386113>.

Littlepage, Jay (2017). *DigitalGlobe moves to the cloud with AWS Snowmobile*.

URL: <http://blog.digitalglobe.com/industry/digitalglobe-moves-to-the-cloud-with-aws-snowmobile/>.

Lorie, Mark et al. (Jan. 2020).

“Modeling coastal flood risk and adaptation response under future climate conditions.”

In: *Climate Risk Management* 29, p. 100233. ISSN: 22120963.

DOI: 10.1016/j.crm.2020.100233.

Lourenço, José et al. (2020).

“Fundamental principles of epidemic spread highlight the immediate need for large-scale serological surveys to assess the stage of the {SARS}-{C}o{V}-2 epidemic.”

In: *medRxiv*.

Lu, Ping et al. (July 2018). “Assessing Hurricane Rainfall Mechanisms Using a Physics-Based Model: Hurricanes Isabel (2003) and Irene (2011).”

In: *Journal of the Atmospheric Sciences* 75.7, pp. 2337–2358. ISSN: 0022-4928.

DOI: 10.1175/JAS-D-17-0264.1.

URL: <http://journals.ametsoc.org/doi/10.1175/JAS-D-17-0264.1>.

Ma, Junling (2020).

“Estimating Epidemic Exponential Growth Rate And Basic Reproduction Number.”

In: *Infectious Disease Modelling*.

Maggiori, Emmanuel et al. (Feb. 2017).

“Convolutional Neural Networks for Large-Scale Remote-Sensing Image Classification.”

In: *IEEE Transactions on Geoscience and Remote Sensing* 55.2, pp. 645–657.

ISSN: 0196-2892. DOI: 10.1109/TGRS.2016.2612821.

URL: <http://ieeexplore.ieee.org/document/7592858/>.

Maier, Benjamin F and Dirk Brockmann (2020). “Effective containment explains

subexponential growth in recent confirmed {COVID}-19 cases in {C}hina.” In: *Science*.

Mandli, Kyle T. and Clint N. Dawson (Mar. 2014).

“Adaptive mesh refinement for storm surge.” In: *Ocean Modelling* 75, pp. 36–50.

ISSN: 1463-5003. DOI: 10.1016/J.OCEMOD.2014.01.002. URL: <https://www.sciencedirect.com/science/article/pii/S1463500314000031?via%3Dihub>.

Marsooli, Reza et al. (Dec. 2019). “Climate change exacerbates hurricane flood hazards

along US Atlantic and Gulf Coasts in spatially varying patterns.”

In: *Nature Communications* 10.1, p. 3785. ISSN: 2041-1723.

DOI: 10.1038/s41467-019-11755-z.

URL: <http://www.nature.com/articles/s41467-019-11755-z>.

Mendelsohn, Robert et al. (Jan. 2012).

“The impact of climate change on global tropical cyclone damage.”

- In: *Nature Climate Change* 2.3, pp. 205–209. ISSN: 1758-678X.
 DOI: 10.1038/nclimate1357.
 URL: <http://www.nature.com/doifinder/10.1038/nclimate1357>.
- Meyerowitz-Katz, Gideon and Lea Merone (2020). “A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates.” In: *medRxiv*.
 URL: <https://www.medrxiv.org/content/10.1101/2020.05.03.20089854v1>.
- Mills, Christina E, James M Robins, and Marc Lipsitch (2004).
 “Transmissibility of 1918 pandemic influenza.” In: *Nature* 432.7019, pp. 904–906.
- Ministère des Solidarités et de la Santé (2020).
Préparation Au Risque Épidémique Covid-19. Tech. rep. URL: https://solidarites-sante.gouv.fr/IMG/pdf/guide_methodologique_covid-19-2.pdf.
- Morrow, Alyssa et al. (2017).
 “Convolutional Kitchen Sinks for Transcription Factor Binding Site Prediction.”
 In: *arXiv preprint* *Id*. URL: <http://arxiv.org/abs/1706.00125>.
- Moulton, Jeremy and Scott Wentland (2018). “Monetary Policy and the Housing Market.”
 In: *Annual Meeting of the American Economic Association*. Philadelphia, PA.
 URL: <https://www.aeaweb.org/conference/2018/preliminary/paper/HTnsAQrn>.
- Munich Re Group (2002).
Winter storms in Europe (II): Analysis of 1999 Losses and Loss Potentials. Tech. rep.
 Munich, Germany.
 URL: http://www.planat.ch/fileadmin/PLANAT/planat_pdf/alle_2012/2001-2005/Munich_Re_Group_2002_-_Winter_storms_in_Europe_II.pdf.
- Muniz-Rodriguez, K et al. (2020).
 “Doubling Time of the {COVID}-19 Epidemic by Province, {C}hina.”
 In: *Emerging Infectious Diseases* 26.8.
- N., Booij, Holthuijsen L.H., and Ris R.C. (July 2020).
 “The ”Swan” Wave Model for Shallow Water.” In: *Coastal Engineering 1996*.
 Proceedings, pp. 668–676. ISBN: 9780784402429.
 DOI: doi:10.1061/9780784402429.053.
 URL: <https://doi.org/10.1061/9780784402429.053>.
- National Oceanic and Atmospheric Administration.
Sea Level Trends - NOAA Tides & Currents.
 URL: <https://tidesandcurrents.noaa.gov/sltrends/sltrends.html>.
- National Oceanographic and Atmospheric Administration (2009).
NOAA Tidal Current Data for the Coastal United States.
 URL: <https://catalog.data.gov/dataset/noaa-tidal-current-data-for-the-coastal-united-states>.
- Neumann, James E., Kerry Emanuel, et al. (Mar. 2015).
 “Joint effects of storm surge and sea-level rise on US Coasts: new economic estimates of impacts, adaptation, and benefits of mitigation policy.”
 In: *Climatic Change* 129.1-2, pp. 337–349. ISSN: 0165-0009.

- DOI: 10.1007/s10584-014-1304-z.
 URL: <http://link.springer.com/10.1007/s10584-014-1304-z>.
- Neumann, James E., Daniel E. Hudgens, et al. (July 2010). "Assessing Sea-Level Rise Impacts: A GIS-Based Framework and Application to Coastal New Jersey."
 In: *Coastal Management* 38.4, pp. 433–455. ISSN: 0892-0753.
 DOI: 10.1080/08920753.2010.496105.
 URL: <http://www.tandfonline.com/doi/abs/10.1080/08920753.2010.496105>.
- Nishiura, Hiroshi et al. (2010). "Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza A ($\{H\}_1\{N\}_1$) 2009."
 In: *Theoretical Biology and Medical Modelling* 7.1, p. 1.
- NOAA National Centers for Environmental Information (2019).
Version 1 VIIRS Day/Night Band Nighttime Lights. DOI: 10.1117/12.2023107.
 URL: https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html#NTL_2015.
- Nordhaus, William (Dec. 2006). *The Economics of Hurricanes in the United States*.
 Tech. rep. Cambridge, MA: National Bureau of Economic Research.
 DOI: 10.3386/w12813. URL: <http://www.nber.org/papers/w12813.pdf>.
- Nordhaus, William D. (May 2010).
 "The Economics of Hurricanes and Implications of Global Warming."
 In: *Climate Change Economics* 1.1, pp. 1–20. ISSN: 20100086.
 DOI: 10.1142/S2010007810000054. URL: <http://www.aoml.noaa.gov/hrd/hurdat/>.
- O'Neill, Brian C. et al. (Jan. 2017). "The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century."
 In: *Global Environmental Change* 42, pp. 169–180. ISSN: 09593780.
 DOI: 10.1016/j.gloenvcha.2015.01.004.
- Pan, Sinno Jialin and Qiang Yang (2010). "A survey on transfer learning."
 In: *IEEE Transactions on Knowledge and Data Engineering* 22.10, pp. 1345–1359.
 ISSN: 10414347. DOI: 10.1109/TKDE.2009.191.
- Patricola, Christina M. and Michael F. Wehner (Nov. 2018).
 "Anthropogenic influences on major tropical cyclone events."
 In: *Nature* 563.7731, pp. 339–346. ISSN: 14764687. DOI: 10.1038/s41586-018-0673-2.
 URL: <https://doi.org/10.1038/s41586-018-0673-2>.
- Penatti, Otávio A B et al. (2015). "Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?" In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 44–51.
 URL: <http://www.torch.ch..>
- Pérez-Suay, Adrián et al. (Dec. 2017).
 "Randomized kernels for large scale Earth observation applications."
 In: *Remote Sensing of Environment* 202, pp. 54–63. ISSN: 00344257.
 DOI: 10.1016/j.rse.2017.02.009.
 URL: <https://www.sciencedirect.com/science/article/abs/pii/S0034425717300615>20<https://doi.org/10.1016/j.rse.2017.02.009>.

Perez, Anthony et al. (2017).

“Poverty prediction with public Landsat 7 satellite imagery and machine learning.”

In: *NIPS 2017 Workshop on Machine Learning for the Developing World*.

URL: <http://arxiv.org/abs/1711.03654>.

Pielke, Roger A (Nov. 2007). “Future economic damage from tropical cyclones: sensitivities to societal and climate changes.” In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365.1860, pp. 2717–2729.

ISSN: 1364-503X. DOI: 10.1098/rsta.2007.2086.

URL: <https://royalsocietypublishing.org/doi/10.1098/rsta.2007.2086>.

Pielke, Roger A. et al. (Feb. 2008).

“Normalized Hurricane Damage in the United States: 1900–2005.”

In: *Natural Hazards Review* 9.1, pp. 29–42. ISSN: 1527-6988.

DOI: 10.1061/(ASCE)1527-6988(2008)9:1(29). URL: <http://ascelibrary.org/doi/10.1061/%28ASCE%291527-6988%282008%299%3A1%2829%29>.

Pielke, Roger A and Christopher W Landsea (Sept. 1998).

Normalized Hurricane Damages in the United States: 1925-95. Tech. rep. 3,

pp. 621–631. DOI: 10.1175/1520-0434(1998)013<0621:NHDITU>2.0.CO;2. URL:

<http://journals.ametsoc.org/waf/article-pdf/13/3/621/4627010/1520-0434>.

Prahl, B F et al. (2015). “Comparison of storm damage functions and their performance.”

In: *Hazards Earth Syst. Sci* 15, pp. 769–788. DOI: 10.5194/nhess-15-769-2015.

URL: www.nat-hazards-earth-syst-sci.net/15/769/2015/.

Presidenza del Consiglio dei Ministri. *COVID-19*.

URL: <https://github.com/pcm-dpc/COVID-19>.

Protezione Civile. *Coronavirus Emergency*.

URL: <http://www.protezionecivile.it/web/guest/home>.

Prudential Regulation Authority (2019).

General Insurance Stress Test 2019: Scenario Specification, Guidelines and Instructions.

Tech. rep. Bank of England.

Rahimi, Ali and Ben Recht (2007). “Random Features for Large-Scale Kernel Machines.”

In: *Advances in Neural Information Processing Systems*.

URL: <https://people.eecs.berkeley.edu/~brecht/papers/07.rah.rec.nips.pdf>.

Rahimi, Ali and Benjamin Recht (Sept. 2008a).

“Uniform approximation of functions with random bases.”

In: *46th Annual Allerton Conference on Communication, Control, and Computing*.

IEEE, pp. 555–561. ISBN: 9781424429264. DOI: 10.1109/ALLERTON.2008.4797607.

URL: <http://ieeexplore.ieee.org/document/4797607/>.

— (2008b). “Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning.”

In: *Advances in neural information processing ...* 1.1, pp. 1313–1320.

URL: <http://papers.nips.cc/paper/3495-weighted-sums-of-random-kitchen-sinks-replacing-minimization-with-randomization-in-learning>.

- Recht, Benjamin et al. (2019). “Do ImageNet Classifiers Generalize to ImageNet?”
In: *International Conference on Machine Learning*, pp. 5389–5400.
- Reed, Fennis et al. (Sept. 2018).
“Gridded Population Maps Informed by Different Built Settlement Products.”
In: *Data* 3.3, p. 33. DOI: 10.3390/data3030033.
URL: <http://www.mdpi.com/2306-5729/3/3/33>.
- Rhodium Group (2019).
Clear, Present and Underpriced: The Physical Risks of Climate Change. Tech. rep.
Oakland, CA. URL: https://rhg.com/wp-content/uploads/2019/03/RHG_PhysicalClimateRisk_Report_April_Final.pdf.
- Risser, Mark D. and Michael F. Wehner (Dec. 2017).
“Attributable Human-Induced Changes in the Likelihood and Magnitude of the
Observed Extreme Precipitation during Hurricane Harvey.”
In: *Geophysical Research Letters* 44.24, pp. 457–12. ISSN: 00948276.
DOI: 10.1002/2017GL075888. URL: <http://doi.wiley.com/10.1002/2017GL075888>.
- Robinson, Caleb, Fred Hohman, and Bistra Dilkina (2017).
“A Deep Learning Approach for Population Estimation from Satellite Imagery.”
In: *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities -
GeoHumanities 2017*. New York, New York, USA: ACM Press, pp. 47–54.
ISBN: 9781450354967. DOI: 10.1145/3149858.3149863.
URL: <http://dl.acm.org/citation.cfm?doid=3149858.3149863>.
- Rolf, Esther et al. (2020).
“Generalizing Global Observation With Satellite Imagery and Machine Learning.”
In: *Under Review*.
- Romer, Christina D and David H Romer (2010). “The macroeconomic effects of tax
changes: estimates based on a new measure of fiscal shocks.”
In: *American Economic Review* 100.3, pp. 763–801.
- Romero, Adriana, Carlo Gatta, and Gustau Camps-Valls (2016).
“Unsupervised deep feature extraction for remote sensing image classification.”
In: *IEEE Transactions on Geoscience and Remote Sensing* 54.3, pp. 1349–1362.
ISSN: 01962892. DOI: 10.1109/TGRS.2015.2478379.
- Roussel, Olivier (2020). *Fr-SARS-CoV-2*.
URL: <https://www.data.gouv.fr/en/datasets/fr-sars-cov-2/>.
- Ruder, Sebastian (June 2017).
“An Overview of Multi-Task Learning in Deep Neural Networks.” In: *arXiv*.
URL: <http://arxiv.org/abs/1706.05098>.
- Russell, Timothy W et al. (2020).
Using a delay-adjusted case fatality ratio to estimate under-reporting. Tech. rep.
Centre for the Mathematical Modelling of Infectious Diseases, London School of
Hygiene & Tropical Medicine. URL: https://cmmid.github.io/topics/covid19/severity/global_cfr_estimates.html.

Santé Publique France. *Santé Publique France*.

URL: <https://www.santepubliquefrance.fr/>.

Schmidt, Silvio, Claudia Kemfert, and Peter Höppe (Nov. 2009).

“Tropical cyclone losses in the USA and the impact of climate change — A trend analysis based on data from a new approach to adjusting storm losses.”

In: *Environmental Impact Assessment Review* 29.6, pp. 359–369. ISSN: 0195-9255.

DOI: 10.1016/J.EIAR.2009.03.003. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0195925509000493>:

<https://www.sciencedirect.com/science/article/pii/S0195925509000493?via%3Dihub>.

— (Mar. 2010). “The impact of socio-economics and climate change on tropical cyclone losses in the USA.” In: *Regional Environmental Change* 10.1, pp. 13–26.

ISSN: 1436-3798. DOI: 10.1007/s10113-008-0082-4.

URL: <http://link.springer.com/10.1007/s10113-008-0082-4>.

Sebastian, Antonia et al. (June 2014). “Characterizing hurricane storm surge behavior in Galveston Bay using the SWAN + ADCIRC model.”

In: *Coastal Engineering* 88, pp. 171–181. ISSN: 0378-3839.

DOI: 10.1016/J.COASTALENG.2014.03.002.

URL: <https://www.sciencedirect.com/science/article/pii/S0378383914000556>.

Simonyan, Karen and Andrew Zisserman (Sept. 2015).

“Very deep convolutional networks for large-scale image recognition.”

In: *International Conference on Learning Representations*. San Diego, CA.

URL: <http://arxiv.org/abs/1409.1556>.

Sobel, Adam H et al. (July 2016). “Human influence on tropical cyclone intensity.”

In: *Science (New York, N.Y.)* 353.6296, pp. 242–6. ISSN: 1095-9203.

DOI: 10.1126/science.aaf6574.

URL: <http://www.ncbi.nlm.nih.gov/pubmed/27418502>.

Sriver, Ryan L et al. (Feb. 2018).

“Characterizing uncertain sea-level rise projections to support investment decisions.”

In: *PLOS ONE* 13.2, e0190641.

URL: <https://doi.org/10.1371/journal.pone.0190641>.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019).

“Energy and Policy Considerations for Deep Learning in NLP.” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3654–3650.

Tamura, Makoto et al. (Mar. 2019). “Global assessment of the effectiveness of adaptation in coastal areas based on RCP/SSP scenarios.”

In: *Climatic Change* 152.3-4, pp. 363–377. ISSN: 15731480.

DOI: 10.1007/s10584-018-2356-2.

URL: <https://doi.org/10.1007/s10584-018-2356-2>.

Tang, Biao et al. (2020). “Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions.”

In: *Journal of Clinical Medicine* 9.2, p. 462.

- Taylor, Karl E, Ronald J Stouffer, and Gerald A Meehl (Apr. 2012).
 “An Overview of CMIP5 and the Experiment Design.”
 In: *Bulletin of the American Meteorological Society* 93.4, pp. 485–498. ISSN: 0003-0007.
 DOI: 10.1175/BAMS-D-11-00094.1.
 URL: <https://doi.org/10.1175/BAMS-D-11-00094.1>.
- The National Map—New data delivery homepage, advanced viewer, lidar visualization*
 (2019). English. Tech. rep. Reston, VA, p. 2. DOI: 10.3133/fs20193032.
 URL: <http://pubs.er.usgs.gov/publication/fs20193032>.
- Thomas, Ajimon et al. (2019). “Influence of storm timing and forward speed on tides and storm surge during Hurricane Matthew.” In: DOI: 10.1016/j.ocemod.2019.03.004.
 URL: <https://doi.org/10.1016/j.ocemod.2019.03.004>.
- Tian, Huaiyu et al. (2020). “An investigation of transmission control measures during the first 50 days of the {COVID}-19 epidemic in {C}hina.” In: *Science*.
- Tiggeloven, Timothy et al. (Apr. 2020). “Global-scale benefit-cost analysis of coastal flood adaptation to different flood risk drivers using structural measures.”
 In: *Natural Hazards and Earth System Sciences* 20.4, pp. 1025–1025. ISSN: 15618633.
 URL: <https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=15618633&v=2.1&it=r&id=GALE%7CA621126931&sid=googleScholar&linkaccess=fulltext%20https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=15618633&v=2.1&it=r&id=GALE%7CA621126931&sid=googleScholar&linkaccess=abs>.
- Tozer, B et al. (Aug. 2019).
 “Global Bathymetry and Topography at 15 Arc Sec: SRTM15+.”
 In: *Earth and Space Science* 6, pp. 1–18. DOI: 10.1029/2019EA000658.
 URL: <https://doi.org/10.1029/>.
- Tsagkatakis, Grigorios et al. (2019).
 “Survey of deep-learning approaches for remote sensing observation enhancement.”
 In: *Sensors (Switzerland)* 19.18, pp. 1–39. ISSN: 14248220. DOI: 10.3390/s19183929.
- Tsang, Tim K et al. (2020). “Impact of changing case definitions for {COVID}-19 on the epidemic curve and transmission parameters in mainland {C}hina.” In: *medRxiv*.
- U.S. Army Corps of Engineers (2000).
Economic Guidance Memorandum (EGM) 01-03, Generic Depth-Damage Relationships.
 Tech. rep.
 URL: <https://planning.erdc.dren.mil/toolbox/library/EGMs/egm01-03.pdf>.
- (2003). *Economic Guidance Memorandum (EGM) 04-01, Generic Depth-Damage Relationships for Residential Structures with Basements*. Tech. rep.
 U.S. Army Corps of Engineers.
 URL: <https://planning.erdc.dren.mil/toolbox/library/EGMs/egm04-01.pdf>.
- U.S. Census Bureau (2015).
2015 American Community Survey 5-Year Estimates, Table B19013.
 URL: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_15_5YR_B19013&prodType=table.

- U.S. Census Bureau (2016). *TIGER/Line Geodatabases*.
 URL: <https://www.census.gov/geo/maps-data/data/tiger-line.html>.
- (2021). *Budget Estimates, Fiscal Year 2021*.
 URL: <https://www2.census.gov/about/budget/census-fiscal-year-21-presidents-budget.pdf>.
- U.S. Department of Homeland Security (2019). *FIMA NFIP Redacted Claims Data Set*.
 URL: <https://www.fema.gov/media-library/assets/documents/180374>.
- U.S. Federal Emergency Management Agency. *Significant Flood Events*.
- UNEP (2019). *Emissions Gap Report 2019*. Tech. rep.
 Nairobi: United Nations Environment Programme.
 URL: <http://www.unenvironment.org/emissionsgap>.
- Union of Concerned Scientists (2018). *Underwater: Rising Seas, Chronic Floods, and the Implications for US Coastal Real Estate*. Tech. rep. Union of Concerned Scientists.
 URL: <http://www.zillow.com/ztrax..>
- (Jan. 2019). *UCS Satellite Database*. URL:
<https://www.ucsusa.org/nuclear-weapons/space-weapons/satellite-database>.
- United States Census Bureau. *Population Datasets*.
 URL: <https://www2.census.gov/programs-surveys/popest/datasets/>.
- United States Global Change Research Program (2018).
Fourth National Climate Assessment. Tech. rep., pp. 1–470.
 URL: <https://nca2018.globalchange.gov/>.
- US Army Corps of Engineers. *National Levee Database*.
 URL: <https://levees.sec.usace.army.mil/#/>.
- US Census Bureau (2017). *American Community Survey (ACS)*.
 URL: <https://census.gov/programs-surveys/acs/>.
- USA Facts. *Coronavirus Locations: COVID-19 Map by County and State*. URL:
<https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/>.
- Vecchi, Gabriel A. and Thomas R. Knutson (Mar. 2011).
 “Estimating Annual Numbers of Atlantic Hurricanes Missing from the HURDAT Database (1878–1965) Using Ship Track Density.”
 In: *Journal of Climate* 24.6, pp. 1736–1746. ISSN: 0894-8755.
 DOI: 10.1175/2010JCLI3810.1.
 URL: <http://journals.ametsoc.org/doi/abs/10.1175/2010JCLI3810.1>.
- Vecchi, Gabriel A, Kyle L Swanson, and Brian J Soden (Oct. 2008).
 “Whither Hurricane Activity?” In: *Science* 322.5902, 687 LP –689.
 DOI: 10.1126/science.1164396.
 URL: <http://science.sciencemag.org/content/322/5902/687.abstract>.
- Verisk Analytics (2019). *PCS Overview*. URL: <https://www.verisk.com/insurance/products/property-claim-services/overview/#>.
- Vincent, Pascal et al. (2010). “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.”
 In: *Journal of machine learning research* 11.Dec, pp. 3371–3408.

- Volpi, Michele and Devis Tuia (2017). “Dense semantic labeling of subdecimeter resolution images with convolutional neural networks.”
In: *IEEE Transactions on Geoscience and Remote Sensing* 55.2, pp. 881–893.
ISSN: 01962892. DOI: 10.1109/TGRS.2016.2616585.
- Vuuren, Detlef P van et al. (2011).
“The representative concentration pathways: an overview.”
In: *Climatic Change* 109.1, p. 5. ISSN: 1573-1480. DOI: 10.1007/s10584-011-0148-z.
URL: <https://doi.org/10.1007/s10584-011-0148-z>.
- White, Stephen A. et al. (Oct. 2016). *An Assessment of the Revised VDATUM for Eastern Florida, Georgia, South Carolina, and North Carolina*. Tech. rep.
National Oceanic and Atmospheric Administration.
- WHO Ebola Response Team (2014). “Ebola virus disease in West Africa—the first 9 months of the epidemic and forward projections.”
In: *New England Journal of Medicine* 371.16, pp. 1481–1495.
- Wing, Oliver E.J. et al. (Dec. 2020).
“New insights into US flood vulnerability revealed from flood insurance big data.”
In: *Nature Communications* 11.1, pp. 1–10. ISSN: 20411723.
DOI: 10.1038/s41467-020-15264-2.
URL: <https://doi.org/10.1038/s41467-020-15264-2>.
- World Health Organization (2020). *WHO Coronavirus Disease (COVID-19) Dashboard*.
URL: <https://covid19.who.int/>.
- Wu, Fan et al. (2020).
“A new coronavirus associated with human respiratory disease in {C}hina.”
In: *Nature* 579.7798, pp. 265–269.
- Wu, Joseph T et al. (2020). “Estimating clinical severity of {COVID}-19 from the transmission dynamics in {W}uhan, {C}hina.” In: *Nature Medicine*, pp. 1–5.
Wuhan pneumonia: 30 days from outbreak to out of control (Jan. 2020).
URL: <https://www.bbc.com/zhongwen/simp/chinese-news-51290945>.
- Xie, Michael et al. (Sept. 2016).
“Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping.”
In: *Thirtieth AAAI Conference on Artificial Intelligence*.
URL: <http://arxiv.org/abs/1510.00098>
<http://www.arxiv.org/pdf/1510.00098.pdf>
<https://arxiv.org/abs/1510.00098>.
- Yang, Kun, Vladimir Paramygin, and Y. Peter Sheng (Nov. 2019). “An objective and efficient method for estimating probabilistic coastal inundation hazards.”
In: *Natural Hazards* 99.2, pp. 1105–1130. ISSN: 15730840.
DOI: 10.1007/s11069-019-03807-w.
URL: <https://doi.org/10.1007/s11069-019-03807-w>.
- Yin, Hao, Zhu Liu, and Daniel M Kammen (Jan. 2020).
“Impacts of Early Interventions on the Age-Specific Incidence of COVID-19 in New York, Los Angeles, Daegu and Nairobi.” In: *medRxiv*, p. 2020.04.19.20071803.

- DOI: 10.1101/2020.04.19.20071803. URL:
<http://medrxiv.org/content/early/2020/04/22/2020.04.19.20071803.abstract>.
- Yu, Le et al. (July 2014).
“Meta-discoveries from a synthesis of satellite-based land-cover mapping research.”
In: *International Journal of Remote Sensing* 35.13, pp. 4573–4588. ISSN: 0143-1161.
DOI: 10.1080/01431161.2014.930206.
URL: <https://www.tandfonline.com/doi/full/10.1080/01431161.2014.930206>.
- Zhang, Gan et al. (Apr. 2020). “Tropical cyclone motion in a changing climate.”
In: *Science Advances* 6.17, eaaz7610. DOI: 10.1126/sciadv.aaz7610.
URL: <http://advances.sciencemag.org/content/6/17/eaaz7610.abstract>.
- Zhong, Yanfei et al. (Feb. 2017). “SatCNN: satellite image dataset classification using agile convolutional neural networks.” In: *Remote Sensing Letters* 8.2, pp. 136–145.
ISSN: 21507058. DOI: 10.1080/2150704X.2016.1235299. URL: <https://www.tandfonline.com/doi/full/10.1080/2150704X.2016.1235299>
<http://dx.doi.org/10.1080/2150704X.2016.1235299>.
- Zillow (2017). *ZTRAX: Zillow Transaction and Assessor Dataset*.
URL: <http://www.zillow.com/ztrax/>.
- Zillow Research. *zillow-research/ztrax*.
URL: <https://github.com/zillow-research/ztrax>.
- Zoph, Barret and Quoc V. Le (Nov. 2016).
“Neural Architecture Search with Reinforcement Learning.”
URL: <http://arxiv.org/abs/1611.01578>.