Title
Spatial Data Science for addressing environmental challenges in the 21st century

Author
Palomino, Jenny Lizbeth

Publication Date
2018

Peer reviewed|Thesis/dissertation

Spatial Data Science for addressing environmental challenges in the 21st century

By

Jenny Lizbeth Palomino

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Science, Policy, and Management

In the

Graduate Division

Of the

University of California, Berkeley

Committee in charge:

Professor Maggi Kelly, Chair
Professor Matthew Potts
Professor Iryna Dronova

Spring 2018

Spatial Data Science for addressing environmental challenges in the 21st century

By

Jenny Lizbeth Palomino

Abstract

Spatial Data Science for addressing environmental challenges in the 21st century

By

Jenny Lizbeth Palomino

Doctor of Philosophy in Environmental Science, Policy, and Management

University of California, Berkeley

Professor Maggi Kelly, Chair

The year 2005 sparked a geographic revolution through the release of Google Maps, arguably the first geographic tool to capture public interest and act as a catalyst for neogeography (i.e. the community of non-geographers who built tools and technologies without formal training in geography). A few years later, in 2008, the scientific community witnessed another major turning point through open access to the Landsat satellite archive, which had been collecting earth observation data since 1972. These moments were critical starting points of an explosion in geographic tools and data that today remains on a rapid upward trajectory. In more recent years, new additions in data and tools have come from the Free and Open Source Software (FOSS), open and volunteered data movements, new data collection methods (such as unmanned aerial vehicles, micro-satellites, real-time sensors), and advances in computational technologies such as cloud and high performance computing (HPC). However, within the broader Data Science community, specific attention was often not given to the unique characteristics (e.g. spatial dependence) and evolutions in geospatial data (e.g. increasing temporal/spatial resolutions and extents). Beginning in 2015, researchers such as Luc Anselin as well as others who had been developing geospatial cyber-infrastructure (CyberGIS) since 2008 began to call for a *Spatial* Data Science, a field that could leverage the advances from Data Science, such as data mining, machine learning, and other statistical and visualization 'big' data techniques, for geospatial data. New challenges have emerged from this rapid expansion in data and tool options: how to scale analyses for 'big' data; deal with uncertainty and quality for data synthesis; evaluate options and choose the *right* data or tool; integrate options when only one will not suffice; and use emerging tools to effectively collaborate on increasingly more multi-disciplinary and multi-dimensional research that aims to address our current societal and environmental challenges, such as climate change, loss of biodiversity and natural areas, and wildfire management.

This dissertation addresses in part these challenges by applying emerging methods and tools in Spatial Data Science (such as cloud-computing, cluster analysis and machine learning) to develop new frameworks for evaluating geospatial tools based on collaborative potential and for evaluating and integrating competing remotely-sensed map products of vegetation change and disturbance. In Chapter One, I discuss in further detail the historical trajectory toward a Spatial

Data Science and provide a new working definition of the field that recognizes its interdisciplinary and collaborative potential and that serves as the guiding conceptual foundation of this dissertation. In Chapter Two, I identify the key components of a collaborative Spatial Data Science workflow to develop a framework for evaluating the various functional aspects of multi-user geospatial tools. Using this framework, I then score thirty-one existing tools and apply a cluster analysis to create a typology of these tools. I present this typology as the first map of the emergent ecosystem and functional niches of collaborative geospatial tools. I identify three primary clusters of tools composed of eight secondary clusters across which divergence is driven by required infrastructure and user involvement. I use my results to highlight how environmental collaborations have benefited from these tools and propose key areas of future tool development for continued support of collaborative geospatial efforts.

In Chapters Three and Four, I apply Spatial Data Science within a case study of California fire to compare the differences as well as explore the synergies between the three remotely-sensed map products of vegetation disturbance for 2001-2010: Hansen Global Forest Change (GFC); North American Forest Dynamics (NAFD); and Landscape Fire and Resource Management Planning Tools (LANDFIRE). Specifically, Chapter Three identifies the implications of the differing creation methods of these products on their representations of disturbance and fire. I identify that LANDFIRE (the traditional created product that integrates field data and public data on disturbance events with remote sensing) reported the highest amount of vegetation disturbance across all years and habitat types, as compared to GFC and NAFD, which are both produced from automated remote sensing analyses. I also find that these differences in reported disturbance are driven by differential inclusion of reference data on fire (rather than differences in environmental conditions) and identify the widest range in reported disturbance (i.e. more uncertainty) in years with more fire incidence and in scrub/shrub habitat. In Chapter Four, I use spatial agreement among the competing products as a measure of uncertainty. I identify low uncertainty in disturbance (i.e. where all products agree) across only 15% of the total area of California that was reported as disturbed by at least one product between 2001 and 2010. Specifically, I find that scrub/shrub habitat had a lower uncertainty of disturbance than forest, particularly for fire, and that uncertainty was universally high across all bioregions. I also identify that LANDFIRE was solely responsible for approximately 50% of the total area reported as disturbed and find large differences between the burned areas reported by the reference data and the areas with low uncertainty of disturbance, indicating potential overestimation of disturbance by both LANDFIRE and the reference data on fire.

Last, in Chapter Five, I conclude by highlighting how unresolved key challenges for Spatial Data Science can serve as new opportunities to guide the scaling of methods for "big" data, increased spatial-temporal integration, as well as promote new curriculum to better prepare future Spatial Data Scientists. In all, this dissertation explores the opportunities and challenges posed by Spatial Data Science and serves as a guiding reference for professionals and practitioners to successfully navigate the changing world of geospatial data and tools.

Table of Contents

List of Figures

List of Tables

Acknowledgements

To my chair, Maggi Kelly, thank you for seeing the potential of my unstructured and wild ideas and helping me to make them a reality. The early years were not an easy transition from the professional world back to academia, but you helped me to remember the reasons that I had chosen to commit to a doctorate program and encouraged me to not give up. In the end, I was able to achieve success through your advice to take it one task at a time, "bird by bird".

I am also very grateful for my dissertation committee members, Matthew Potts and Iryna Dronova, who have provided so much support and guidance throughout my time at Berkeley. A special thanks to the staff at the Geospatial Innovation Facility, both past and present, especially Kevin Koy and Nancy Thomas, who always provided friendly and encouraging ears.

To past and present Kelly Lab and Potts Lab members who supported and encouraged me on a daily basis to keep going, even when things got really hard, especially: Alice Kelly, Kelly Easterday, Hannah Salim-Doll, Matthew Luskin, Siew Chin Chua, and Sam Evans. To Oliver Muellerklein, I am grateful for such a fun and fruitful collaboration on my first publication of this dissertation. I wish you all continued success in work, and most importantly, in life.

Last, I would like to thank my wonderful husband, family, and friends who have supported and stood by me throughout these years. Brian, Mom, Dad, Derly, Joe, Julia, Samantha, Olivia, Ron, and Sherri – I am so lucky to have such a wonderful family. To the greatest friends possible, Thao, Holly, and Celia, and to my Jedi sisters (you know who you are): my sincerest thanks. To Bree Rosenblum: knowing you has changed my life in more ways than I could have ever thought possible. It has been a crazy, tumultuous, and beautiful journey, and I could not have made it through without you all.

Chapter One

Opportunities and challenges presented by Spatial Data Science

Since 2005, when Google Maps launched, the discipline of geography has experienced significant revitalization from increased public use of mapping as well as a major expansion in the availability of geospatial data that is broader in extent (from regional to global) as well as finer in spatial and temporal resolution (in real-time and daily). More recently, new data collection methods (e.g. mobile, sensor networks, unmanned aerial vehicles, micro-satellites) (Anselin, 2015; Miller & Goodchild, 2015) as well as open data (Anselin, 2015) and volunteered data movements (Goodchild, 2007) have made geospatial data "big", characterized not only by the volume, velocity, and variety that have come to characterize large and complex datasets, but also by the increased detail and complexity of the data. Intertwined with these evolutions in geospatial data collection has been an increased availability of tools to analyze, visualize and disseminate data including: scripting libraries in the widely used Python and R programming languages; the development of interactive web mapping (Haklay et al., 2008); more free and open source software (FOSS) options (Steiniger & Hunter 2013); web-based methods to seamlessly exchange data, such as application programming interfaces (APIs); and multi-user platforms that support collaborative workflows for geospatial tasks (i.e. collaborative geospatial tools). To harness this explosion in geospatial data and tools, the field of CyberGIS emerged around 2010 to create methods for improving tool interoperability and leveraging high performance computing (HPC) to overcome desktop-based limitations for analyzing of big geospatial data (Anselin, 2012; Wang et al., 2013; Wright & Wang, 2011; Yang et al., 2010). Since then, researchers in geography as well as professionals working with geospatial data have also recognized the importance of modifying and scaling analytical methods to match the size and complexity of modern geospatial data through data handling, mining, and statistical and computational techniques borrowed from Data Science and Computer Science. These efforts have been coalescing into a newly emerging discipline known as Spatial Data Science.

The first mention of Spatial Data Science in the scientific literature appeared in a 2015 online essay by Luc Anselin, who defined it as:

"*a subset of the broader data science (e.g., Schutt and O'Neil 2014), differentiated by dealing explicitly with the role of space (location, spatial arrangement, spatial interaction). In analogy to generic data science, it consists of a combination of the strengths of exploratory* spatial *data analysis,* spatial *statistics, and* spatial *econometrics from a statistical disciplinary perspective, with* spatial *data mining,* spatial *database manipulation, and machine learning from a computer science disciplinary perspective*" (Anselin, 2015, pg. 1).

Since Anselin's essay, other researchers have similarly identified that "geospatial data science is a transdisciplinary field comprising statistics, mathematics, and computer science" (Eftelioglu et

al., 2017, pg. 9), though some scholars have outlined goals and priorities for this emerging field while treating it as a known entity that needs no formal definition:

"*While data science is in general driven by data analysis and related creative inquiry, data synthesis has become increasingly important in the context of spatial data science. Two specific and interrelated challenges are fundamental to spatial data synthesis: data aggregation and integration. Data aggregation refers to the problem of bringing together various data streams at scale, while data integration concerns the processes of harmonizing diverse (in format, type, spatial scale, spatial reference, spatial unit, etc.) spatial datasets*" (Wang, 2016, pg. 967).

"Spatial or Geospatial Data Science… Primary communities…Geography, Statistics, Computer Science… Special emphases…Spatial Statistics, Spatial Big Data, Machine learning" (Yuan, 2017, pg. 429).

While these descriptions provide useful insight into the methodological gains that Spatial Data Science has provided (i.e. modern statistical and data handling methods for big data, integration of machine learning, focus on data synthesis), there has been little recognition (i.e. discussion, research) within Spatial Data Science of the computational environment or *tools* that are actually used (or needed) to conduct these analyses. Specifically, within academia, research on the computational tools used to accomplish the goals of Spatial Data Science has primarily been the domain of CyberGIS, formally defined as "geographical information science and systems (GIS) based on advanced cyberinfrastructure (CI) … [that] provides a seamless integration of advanced CI, GIS, and spatial analysis and modeling capabilities" (Wang, 2016, pg. 966). As such, CyberGIS is typically presented as a related but separate field that focuses on developing the technological infrastructure needed to conduct Spatial Data Science (i.e. interoperability between tools and HPC integration). The lack of focus on technology is a critical omission within Spatial Data Science because many widely-used geospatial tools (e.g. popular FOSS options such as QGIS, MapBox, ROpenSci, Jupyter Notebook, and PostGIS; cloud-based tools like CARTO and Google Earth Engine) are not developed by CyberGIS researchers but rather by other academic researchers (e.g. natural and physical scientists), private industry, and individuals collaborating within non-profit efforts that support the development of FOSS (e.g. OSGeo Foundation). Disregarding the integral role of technological infrastructure within Spatial Data Science ignores the need to investigate *how* tools can best be used to integrate and synthesize disparate data and methods as well as support increasingly larger and distributed teams in collaborative work.

In this dissertation, I unite these various aspects of geospatial data and tool development by providing a new, and more broad definition of Spatial Data Science: a collaborative discipline that *integrates* fundamental GIS methods and principles of spatial analysis, geoprocessing, and cartography, *with* statistical, data mining, and web-based data visualization techniques from Data Science, *within* infrastructure that leverages recent advances in computation– such as FOSS, HPC, and cloud-based computing– through the efforts of CyberGIS researchers and geospatial

professionals and practitioners. Based on this richer definition of Spatial Data Science, my research addresses several key challenges for researchers and professionals that stem from the rapid and intertwined growth in available geospatial data and analytical tools: choosing the *right* geospatial data or tool from an ever-expanding list of options; evaluating data; and examining uncertainty. As no one dataset represents a complete truth and no one tool can fulfill all needs, this dissertation provides frameworks for evaluating geospatial tools based on how they can be leveraged for *collaborative research* (Chapter Two) and for *comparing and integrating* competing or alternative geospatial datasets (Chapter Three) and *evaluating the uncertainty* in competing geospatial products (Chapter Four). Specifically, using emerging methods and tools in Spatial Data Science such as cloud-computing, cluster analysis and machine learning, I apply these frameworks in case studies to provide the first typology of collaborative geospatial tools (Chapter Two), to identify the implications of differing creation methods of remotely-sensed maps of vegetation change on the representation of disturbance and fire (Chapter Three), and to quantify spatially explicit uncertainty in disturbance (particularly fire) across landscapes (Chapter Four). Overall, I believe these frameworks provide easy and clear methodologies for researchers and practitioners to evaluate available options and to integrate multiple choices into unified solutions.

While this dissertation emphasizes the critical role of tools and technology, it also recognizes that our current societal and environmental challenges such as climate change, loss of biodiversity and natural areas, and wildfire management are challenging researchers and professionals to find solutions that require increased collaboration between multiple stakeholders and the synthesis of data across multiple scales and formats. Though Spatial Data Science cannot address all aspects of these challenges (e.g. how to best engage with public audiences or design policies that address both ecological and societal needs), the discipline can provide the theoretical and technical infrastructure for the collaboration, data synthesis, and large-scale analysis needed to explore scientific solutions to these "wicked" problems (Allen & Gould, 1986; Balint et al., 2011; Carroll et al., 2007; Temby et al., 2016), or "complex social-ecological systems" (Akamani et al., 2016) that cross multiple administrative and ecological boundaries, are part of coupled and complex spatial-temporal interactions, and do not have past precedent for solutions. In the sections that follow, I outline each chapter of this dissertation and highlight how I employ Spatial Data Science methods and technologies within case studies to address these new research questions stemming from the rapidly expansion in geospatial data and tools.

The integral role of collaboration in evaluating geospatial tools

As researchers and practitioners seek to address complex and multi-dimensional questions about environment challenges that require both geospatial data and increasingly larger teams, I are also searching for tools that can provide the functionality needed for multiple users to work together on geospatial tasks. In Chapter Two, I identify that previous typologies of geospatial tools have not explicitly considered technical functionality that tools provide for completing collaborative

tasks, and thus are not useful for understanding how geospatial tools can support collaborative geospatial research. To this end, I outline a new Spatial Data Science workflow that highlights the role of collaboration (i.e. user involvement, reproducibility) and computational infrastructure (i.e. FOSS, HPC) alongside key workflow tasks (i.e. setting up the working environment, data wrangling, analysis, visualization, publication/dissemination). I use this collaborative Spatial Data Science workflow to score thirty-one geospatial tools that aim to support multiple users working collaboratively on geospatial tasks. I then apply an unsupervised cluster analysis (using a machine learning approach) to develop the first typology of geospatial tools based on functionality for collaboration. I present this typology as a map of the emergent ecosystem of collaborative geospatial tools with three identifiable functional niches: (1) participatory data aggregators; (2) content managers; and (3) highly scalable and customizable tools. Using this new typology, I highlight the current strengths of collaborative geospatial tools (i.e. strong integration of open source technologies and interoperability), identify key areas of future development (i.e. more integration of HPC and versioning of data and workflows), and outline ongoing challenges for both collaborative tool development and Spatial Data Science (i.e. big data handling, scaling of analyses, spatial-temporal integration, data quality and synthesis).

Evaluating and integrating competing datasets to understand and quantify uncertainty

Chapters Three and Four apply Spatial Data Science to the challenge of describing and quantifying landscape change in a data-rich world, in which the existence of multiple, competing datasets result in overall uncertainty in amount and location of vegetation disturbance over time (i.e. change or reduction in vegetation due to natural or anthropogenic causes such as fire, harvesting, and mortality due to drought and pestilence). I use a case study of fire in California to compare the differences as well as explore the synergies between the three vegetation disturbance products that have complete coverage for California between 2001 and 2010: Hansen Global Forest Change (GFC); North American Forest Dynamics (NAFD); and Landscape Fire and Resource Management Planning Tools (LANDFIRE). While these products have been developed from the same data source– the Landsat Time Series (LTS) which provides satellite images at moderate spatial (30 m resolution) and temporal (approximately 2 weeks) resolutions– the products differ greatly in the methodology used to identify and map disturbance. Specifically, GFC and NAFD are both based on modern "big" data approaches that are automated and analyze the entire dataset to quantify annual changes (i.e. compare values across years to identify the change in a given year), while LANDFIRE has been developed through a traditional, year-to-year approach that manually combines public data on annual disturbance events provided by federal, state, and local agencies with semi-automated remote sensing analyses of vegetation indices derived from the LTS. Although these products use different methods to map annual vegetation disturbance, they are easily comparable due to the shared 30 m spatial resolution of the LTS. However, this spatial resolution also means that the products are "big" data, covering approximately 450 million pixels for each year of data. As such, I use the cloud-based geospatial

tool, Google Earth Engine (GEE), for its functionality for reproducibility (i.e. code-sharing and cloud data storage) as well as large data handling via cloud-based, distributed computing.

In Chapter Three, I compare the amounts of annual and total vegetation disturbance reported by the products and explore how their differing methods of creation can create uncertainty in disturbance by comparing their coverage of environmental conditions (i.e. vegetation, elevation, climate) and burn conditions from federal and state reference data on fire (i.e. fire perimeters and burn severity data). My results indicate that the products examined differ greatly in their amounts of reported disturbance, particularly in scrub/shrub habitats (where reference data on fire reported higher disturbance) and in years with high numbers of fire events. In particular, the traditionally curated product, LANDFIRE, reported much higher amounts of disturbance across all years, as compared to either of the two modern products, GFC and NAFD. I also find that these differences in reported disturbance are driven by differential inclusion of reference data on fire (e.g. the manual creation process of LANDFIRE explicitly incorporates these reference data), rather than differences in environmental conditions. These results provide key insights into how the products differentially map disturbance and the conditions under which they identify it, which are important for the research and conservation communities that use these products to study the implications of vegetation disturbance on ecosystem characteristics such as carbon dynamics and habitat fragmentation. Due to their differences, I conclude that users of these products should not rely on one product to accurately represent disturbance in their study areas, but rather need to view these products as different representations of disturbance (based on different thresholds) and seek to account for this uncertainty in their work.

In Chapter Four, I outline a simple but powerful methodology for quantifying uncertainty in disturbance based on identifying spatial agreement among competing products (i.e. where the products agree). For this methodology, I use basic raster calculations completed in GEE to identify which products overlap at each pixel, and then convert these levels of spatial agreement to measures of uncertainty based on the number of products that report disturbance at a given location (e.g. high uncertainty of disturbance due to only one product reporting disturbance versus low uncertainty where all products report disturbance). I continue the case study of fire in California to explore uncertainty among GFC, NAFD, and LANDFIRE across biogeographical divisions (i.e. habitat and bioregions) as well as burn conditions (i.e. size of fire, burn severity). I find that low spatial agreement among the disturbance products results in a low uncertainty for only 15% of the total area reported as disturbed across California between 2001 and 2010. Furthermore, I find that while scrub/shrub habitat had a lower uncertainty of disturbance than forest, particularly for fire events, uncertainty was universally high across all bioregions.

My results for Chapter Four also indicate potential over-estimation of disturbance by LANDFIRE as well as by the reference data on fire. Specifically, I identify that LANDFIRE was solely responsible for approximately 50% of the total area reported as disturbed and find large differences between the burned areas reported by the reference data and the areas with low

uncertainty of disturbance derived from spatial agreement among the disturbance products. This uncertainty in disturbance is noteworthy for two primary reasons: (1) almost 10% of California was reported as disturbed by at least one product in the same time period (but only 15% of the total area reported as disturbed had a low uncertainty); and (2) published research that quantify the impact of disturbance on ecosystem characteristics, such as aboveground carbon dynamics, do not account for this uncertainty, and instead choose only one product to represent the amount and location of disturbance in their study areas. Overall, this examination of uncertainty in disturbance in California provides both insight into the potential implications of not accounting for uncertainty in disturbance (e.g. choosing only one product to represent disturbance in a given study area or year) as well as a simple and clear methodology for all users of these products, particularly non-remote sensing experts, to account for uncertainty in disturbance in their work.

Based on the key findings of my research as well as the current developments in the field, Chapter Five concludes this dissertation with final thoughts regarding new research directions for Spatial Data Science. In the concluding chapter, I highlight how unresolved key challenges for Spatial Data Science can serve as new opportunities within this emerging field to guide the scaling of geospatial methods for "big" data, to support increased spatial-temporal integration (particularly for analysis land cover change and vegetation disturbance), as well as to promote the development of new curriculum in Spatial Data Science to better prepare future geospatial professionals and researchers. Overall, this dissertation provides a key exploration of the fundamental opportunities and challenges posed by this newly emerging field and serves as a guiding reference for researchers and practitioners to successfully navigate the changing world of data and tool options and to choose the best options for their needs.

Chapter Two

A review of the emergent ecosystem of collaborative geospatial tools for addressing environmental challenges

To solve current environmental challenges such as biodiversity loss, climate change, and rapid conversion of natural areas due to urbanization and agricultural expansion, researchers are increasingly leveraging large, multi-scale, multi-temporal, and multi-dimensional geospatial data. In response, a rapidly expanding array of collaborative geospatial tools is being developed to help collaborators share data, code, and results. Successful navigation of these tools requires users to understand their strengths, synergies, and weaknesses. In this chapter, I identify the key components of a collaborative Spatial Data Science workflow to develop a framework for evaluating the various functional aspects of collaborative geospatial tools. Using this framework, I then score thirty-one existing collaborative geospatial tools and apply a cluster analysis to create a typology of these tools. I present this typology as a map of the emergent ecosystem and functional niches of collaborative geospatial tools. I identify three primary clusters of tools composed of eight secondary clusters across which divergence is driven by required infrastructure and user involvement. Overall, my results highlight how environmental collaborations have benefited from the use of these tools and propose key areas of future tool development for continued support of collaborative geospatial efforts.

Introduction

Environmental challenges such as biodiversity loss, wildfire management, climate change, and rapid conversion of natural areas due to urbanization and agricultural expansion are recognized as "wicked problems" (Allen & Gould, 1986; Balint, Stewart, & Desai, 2011; Carroll, Blatner, Cohn, & Morgan, 2007; Temby, Sandall, Cooksey, & Hickey, 2016), or "complex social-ecological systems" (Akamani, Holzmueller, & Groninger, 2016). Many of these challenges can be described as global in scale, at the nexus of interdisciplinary approaches, and/or part of coupled processes. Research teams have also become larger, more distributed, and multi-disciplinary (Elwood, Goodchild, & Sui, 2012; MacEachren & Brewer, 2004). To address these challenges, researchers have called for collaboration not only in the environmental management and decision-making processes (Daniels & Walker, 2001; Frame, Gunton, & Day, 2004; Selin & Chevez, 1995), but also in the knowledge production process, including the sharing of data, methods and tools (Cravens, 2014;Head& Alford, 2015; Temby et al., 2016). Consequently, understanding how various technologies, including geospatial tools, can support collaborative efforts for environmental problem-solving is a critical area of ongoing research (Cravens, 2014; Cravens, 2016; MacEachren & Brewer, 2004; Wright, Duncan, & Lach, 2009).

Contemporaneous to the emergence of these complex and large scale research challenges has been a rapid expansion in the sources of geospatial data from mobile devices, environmental

sensors, and Unmanned Aerial Vehicles (Miller & Goodchild, 2015) as well as from increased public access to administrative data through cloud/web-based Application Programming Interfaces (APIs; Anselin, 2015). In addition, Volunteered Geographic Information (VGI; Goodchild, 2007) as well as data captured by citizen scientists continue to increase in volume (Dickinson, Zuckerberg, & Bonter, 2010; Dickinson et al., 2012), both complementing and challenging the anonymity and centralized nature of traditional geospatial data produced by large organizations (i.e. governments and proprietary companies). Available data are now more detailed, with changes in scale from local to global extents, from coarse spatial resolutions in 2D planimetric to fine grain sizes with 3D and 4D options, and from seasonal/monthly temporal scales to daily or real-time capture. As such, researchers working on environmental challenges are increasingly leveraging large, multi-scale, multi-temporal, and multi-dimensional geospatial data in search of solutions (Goodman, Parker, Edmonds, & Zeglin, 2014; Miller & Goodchild, 2015).

Complementing this explosion in data has been the development of diverse array of geospatial analytical tools (i.e., scripting libraries, open source and cloud/web-based mapping options) and increased functionality to support multi-user workflows (i.e. standardized working environments, code-sharing, data exchange, status updates). Through advances in Web 2.0 technologies (Haklay, Singleton, & Parker, 2008) and Free and Open Source Software for Geospatial (FOSS4G; Steiniger & Hunter, 2013), the primary use of geospatial data is evolving from proprietary desktop software and data formats used to create static cartographic products toward the leveraging of open source and cloud/web-based tools, open data format and standards, and APIs to create dynamic web visualizations shared by collaborative teams across technology, science, and the public.

These intertwined evolutions in available geospatial data and tools also highlight the ongoing discussion regarding the role of technology within collaborative projects and how to best leverage technology to support collaborative tasks. Successful collaboration is dependent on many things including dynamics of negotiation, equity in knowledge and power, inclusion and access, and trust, which have been explored by various researchers (Elwood, 2006; Sieber, 2000; Wright et al., 2009). In addition to these social dimensions, collaboration is also dependent on the technology used to complete and achieve the desired tasks and outcomes (Cravens, 2014; Cravens, 2016). In their seminal work on "geocollaboration", MacEachren and Brewer (2004) identify four "stages of group work" as "explore, analyze, synthesize, present" (pg. 7) and explain that these stages represent "collaborative tasks for knowledge construction" (pg. 19) that can be accomplished using technology, especially those for geovisualization.

MacEachren and Brewer (2004) also offer a definition of collaboration that applies well to the context of leveraging geospatial data and technology for environmental problem-solving: "a committed effort …of two or more people to use geospatial information technologies to collectively frame and address a task involving geospatial information" (pg. 2). MacEachren and Brewer (2004) categorize these multi-user collaborations into four types: same place-same time,

same place-different time, different place-same time, and different place-different time, stating that these last two (different place)were still primarily in the prototype phase at the time of their publication and were being driven by advances in database and web technology.

Since then, as these technological advances have progressed further, there has been a rise in technologies that support all of these collaborations, most notably for different place-different time collaborations. In particular, the logistics and mechanisms provided for collective work by technology in general, and geospatial ones in particular, have been identified by other researchers in varying descriptions of collaborations between scientists, non-scientists, and the general public: "collaboratories" (or collaboration laboratories; Pedersen, Kearns, & Kelly, 2007; Wulf, 1993) and "geocollaboratories" (specifically "work by geographically distributed scientists about geographic problems" MacEachren et al., 2006, pg. 201), participatory planning and management (Jankowski, 2009; Kelly, Ferranto, Lei, Ueda, & Huntsinger, 2012; Voss et al., 2004; Wright et al., 2009), citizen science efforts (Connors, Lei, & Kelly, 2012; Dickinson et al., 2010; Dickinson et al., 2012), observatory networks such as National Ecological Observatory Network (NEON; Goodman et al., 2014), virtual networks for collaboration such as Geosciences Network (GEON; Gahegan, Luo, Weaver, Pike, & Banchuen, 2009) and Human-Environment Regional Observatory (HERO; MacEachren et al., 2006) and "action ecology" (White et al., 2015). Through these collaborative efforts, researchers highlight how advances in geospatial data and tools provide technical support for collaborations through facilitation of: (i) group use and development of technology (i.e. field data collection at broad and long scales; dispersed responsibility of tasks); (ii) sharing and peer reviewing of data and results (i.e. crowdsourcing of data validation; data editing by multiple users); (iii) communication between stakeholders (i.e. ability for stakeholders to share their different representations of space and project outcomes); and (iv) integration of complementary tools (i.e. combining geospatial and communication-oriented tools; integration of big data tools and open data formats). Hence, the technical capabilities of geospatial tools can provide the practical mechanisms and infrastructure that allow people to successfully work together on tasks and goals, despite their distributions across time and space.

While it is evident that geospatial tools can support collaboration through providing the technological infrastructure needed for collaborative tasks, existing literature does not yet provide a clear framework for evaluating geospatial tools based on how well they support completion of these collaborative tasks. Furthermore, as projects can differ greatly in their requirements, there is no single tool that fulfills all needs and often, multiple tools must be integrated into workflows. As such, in addition to features that support workflows across multiple users, geospatial tools also need to support interoperability between tools (i.e. transfer of data, methods and results between tools). Consequently, successful navigation of the ever-expanding list of collaborative (i.e. multi-user) geospatial tools requires an understanding of their strengths, synergies, and weaknesses, specifically regarding functionality for collaborative tasks and capabilities for tool interoperability.

A typology of geospatial tools can provide a roadmap for these explorations by focusing on technical infrastructure for collaborative tasks such as setting up common working environments and shared data exploration, analysis, and visualization. This typology would also illustrate connections between collaborative geospatial tools as an ecosystem with identifiable niches of functionality. In this chapter, I provide such a typology of the emergent ecosystem of collaborative geospatial tools by evaluating how key multi-user tools address technical barriers to collaboration through their varying capabilities and functionality.

The three objectives of this chapter are to:

1. Select representative case studies (i.e. collaborative geospatial tools) that have been developed to support multi-user geospatial workflows;
2. Develop a quantitative and reproducible framework to evaluate the tools based on the key components of a collaborative Spatial Data Science workflow; and
3. Apply a cluster analysis to develop a typology of collaborative geospatial tools.

To provide a conceptual understanding of my evaluation framework, I first review the key factors that have led to the evolution of a collaborative Spatial Data Science workflow. Next, I describe how others have previously outlined typologies of geospatial and collaboration tools. Last, I apply my quantitative framework to score and cluster multi-user geospatial tools based on their functionality for collaborative tasks. Overall, I use this typology to present a map of the emergent ecosystem and niches of tools, highlight how environmental research collaborations have benefited from the strengths of these tools, and propose key areas of future tool development for continued support of collaborative geospatial workflows. I believe that understanding the current ecosystem of collaborative geospatial tools can highlight opportunities for expanded or new functionality, promote stronger interoperability between existing tools, and help stakeholders to leverage the best tools for their needs.

The evolution of collaborative Spatial Data Science workflow

In their fundamental work on geocollaboration, MacEachren and Brewer (2004) identify that while many geospatial projects are pursued as group efforts, most geospatial technologies at the time of their writing were developed and evaluated for individual use. To address this discrepancy, the authors propose “geocollaborative environments” that are focused on providing a shared working environment, whether or not the users are in the same physical environment or collaborating in real time. In addition to technical barriers to collaboration in working environments, Steiniger and Hunter (2013) identify barriers to open science stemming from the lack of transparency in analysis methods and programming code. These authors highlight various publications (e.g. Ince, Hatton, & Graham-Cumming, 2012; Morin et al., 2012; Rocchini & Neteler, 2012) arguing against “proprietary- ‘black box’ - programs that hinder scientific advancement and testing” (p. 147). Specifically, Rocchini and Neteler (2012) urge ecologists to embrace Stallman's (1985) “four freedoms” paradigm of FOSS to freely execute, modify, and

share programs, while also identifying the need for better mechanisms (i.e. tools) for scientists to share "the backbone of ecological software: its code" (p. 311). Seemingly responding to the call by MacEachren and Brewer (2004) for better "multi-user system interfaces" as well as calls for increased application of FOSS and open science ideas to geospatial research, there are now more multi-user options to share data and code than ever before. As such, new concerns have arisen about how to choose the right tool, especially when evaluating newer FOSS4G and cloud/web-based tools (Steiniger & Hunter, 2013).

The proliferation of cloud/web-based and FOSS4G tools also highlights the progression from the traditional desktop model of Geographic Information Science (Goodchild, 1992; abbreviated to GI Science per Hall, 2014) to an advancing geospatial cyberinfrastructure, or CyberGIS (Anselin, 2012; Wang et al., 2013; Wright &Wang, 2011; Yang, Raskin, Goodchild, & Gahegan, 2010). In particular, the CyberGIS community has promoted the integration of existing GI Science and spatial analysis tools with cyberinfrastructure tools that harness cloud and high performance computing technologies (i.e. distributed, parallel, clustered) for scalable geospatial data research. Another strong focus of CyberGIS has been on tool interoperability in order to promote the sharing of data and methods as well as reduce the plethora of narrowly customized tools and "non-sharable stove-piped data systems" (Yang et al., 2010, pg. 272). In the quest to transform the technological infrastructure available for geospatial research, CyberGIS has also recognized the importance of support for shared problem-solving, distribution of geospatial data in flexible and secure ways, and community-driven solutions for wrangling and analyzing large and complex datasets (Wang et al., 2013; Wright & Wang, 2011; Yang et al., 2010).

Supported by CyberGIS technical frameworks for tool integration and interoperability, a complementary field of Spatial Data Science is emerging as an interdisciplinary approach to leveraging the spatial data explosion provided by sensor networks, VGI and mobile technologies, and by the open data and science movements (Anselin, 2015; Jiang & Thill, 2015; Wang, 2016; Yuan, 2016). Identifying Spatial Data Science as a branch of the broader Data Science field, Anselin (2015) describes it as "a combination of… exploratory spatial data analysis, spatial statistics, and spatial econometrics from a statistical disciplinary perspective, with spatial data mining, spatial database manipulation, and machine learning from a computer science disciplinary perspective" (p. 1). Maintaining the connection to GI Science, Yuan (2016) similarly describes Spatial Data Science as the domain of "Geography, Statistics, Computer Science" communities that focus on "Spatial Statistics, Spatial Big Data, Machine learning" (p. 5).While exploring synergies between CyberGIS and Spatial Data Science, Wang (2016) identifies key aims of Spatial Data Science as "scalable spatial data access, analysis and synthesis" (p. 3), with key challenges to these goals being data aggregation and data integration. Similarly, Anselin (2015) identifies related challenges as issues of "scale …endogeneity…[and] computational efficiency to deal with large amounts of very fine grained geographical data in near real-time" (p. 2).

While not explicitly mentioned in these definitions of Spatial Data Science, data wrangling (i.e. harnessing, cleaning, transforming) "often constitutes the most tedious and time-consuming aspect of analysis" (Kandel et al., 2011, pg. 271). As such, effective data wrangling plays a key role within modern geospatial workflows and is integral for overcoming the identified challenges of data aggregation, integration and scalability. Similarly, in these descriptions of Spatial Data Science, little emphasis is placed on data representation and visualization, as compared to analysis and synthesis. In particular, cartographic principles remain central to the display and representation of geospatial data, especially for the web. This is evident through the focus on color palettes (i.e. tools like ColorBrewer), vector line simplification (i.e. algorithm-based tools like MapShaper and Simplify), typological representations (i.e. data formats like TopoJSON), and efficient rendering of basemaps and large datasets (i.e. data formats like Vector Tiles). Similarly, visualization has also been identified as a key component of data analysis, as it is "particularly essential for analysing phenomena and processes unfolding in geographical space" (Andrienko & Andrienko, 2013, pg. 3). For example, visual analytics provides methods and tools for analyses of large spatial datasets through interactive visualization of iteratively mined data and has proven particularly important for movement data such as mobile and VGI (Andrienko & Andrienko, 2013; Beecham, Wood, & Bowerman, 2014; Stange, Liebig, Hecker, Andrienko, & Andrienko, 2011).

In light of these descriptions, Spatial Data Science can be seen as standing at the intersection of the three fields of GI Science, Data Science and CyberGIS (Figure 2-1). Through this intersection, Spatial Data Science unites the statistical, data mining, and web-enabled data visualization techniques of Data Science with fundamental GI Science methods and principles of spatial analysis, geoprocessing, and cartography within the computational infrastructure and interoperability potential provided by CyberGIS. With an emphasis on standardized and repeatable workflows, Spatial Data Science promotes the compilation and integration of disparate data from multiple sources, the use of open source and cloud/web-based technologies for robust data analysis, and the leveraging of an expanding suite of data visualization and publication tools to support communication between project collaborators, the public, and other stakeholders. Due to the increasing overload of geospatial data available, the harnessing of tools that assist users in data wrangling, management, analysis, visualization, and publication is critical for collaborative geospatial research. To this end, Spatial Data Science provides a path (i.e. workflow) for navigating the rapidly expanding field of data, methods, and multi-user tools for working with and analyzing large and complex geospatial data.

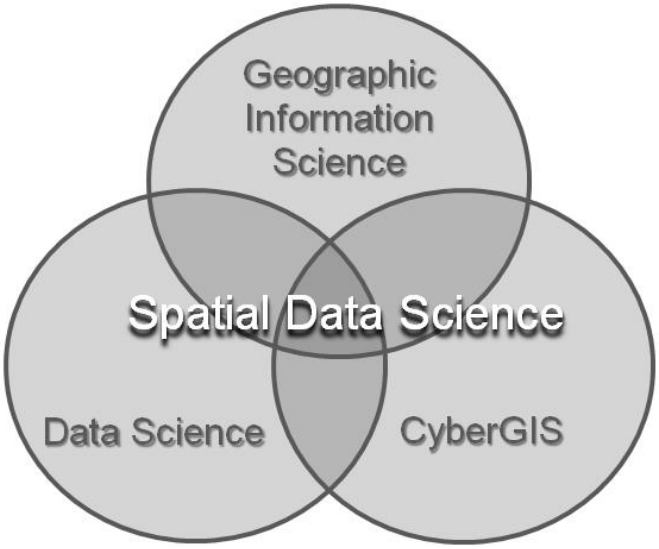


Figure 2-1. Spatial Data Science at the Intersection of GI Science, Data Science and CyberGIS

At the heart of Spatial Data Science is a common workflow (Figure 2-2) that leverages cloud/web-based and open source geospatial tools to address technical impediments to collaboration such as non-standardized working environments, siloing of data, unreproducible analyses, and static map visuals. While there are many possible routes available when navigating a collaborative Spatial Data Science workflow, these routes generally consist of four key primary tasks through which collaboration can not only be fostered, but are actually fundamental to geospatial problem-solving in the 21st century: (1) setting up the working environment; (2) data wrangling (i.e. harnessing, cleaning, transforming); (3) data analysis; (4) data visualization and publication. Both data management and visualization are deeply embedded within all tasks, particularly data wrangling and analysis. Data visualization is highlighted specifically with publication (Figure 2-2) to emphasize its important role in facilitating the dissemination of results and knowledge gained. Facilitating this workflow are (5) the integration and support of FOSS4G and (6) user involvement by the public, scientists, technologists, practitioners, and governments. Given the iterative nature of collaboration, this Spatial Data Science workflow is adaptive; the tools chosen for each task can be modified or replaced as the needs of the projects are further refined or new tools become available. By addressing technical challenges at each step, this collaborative Spatial Data Science workflow allows researchers and stakeholders to more easily share research ideas, analyses, code, results, and conclusions to work toward the integration and synthesis of knowledge.

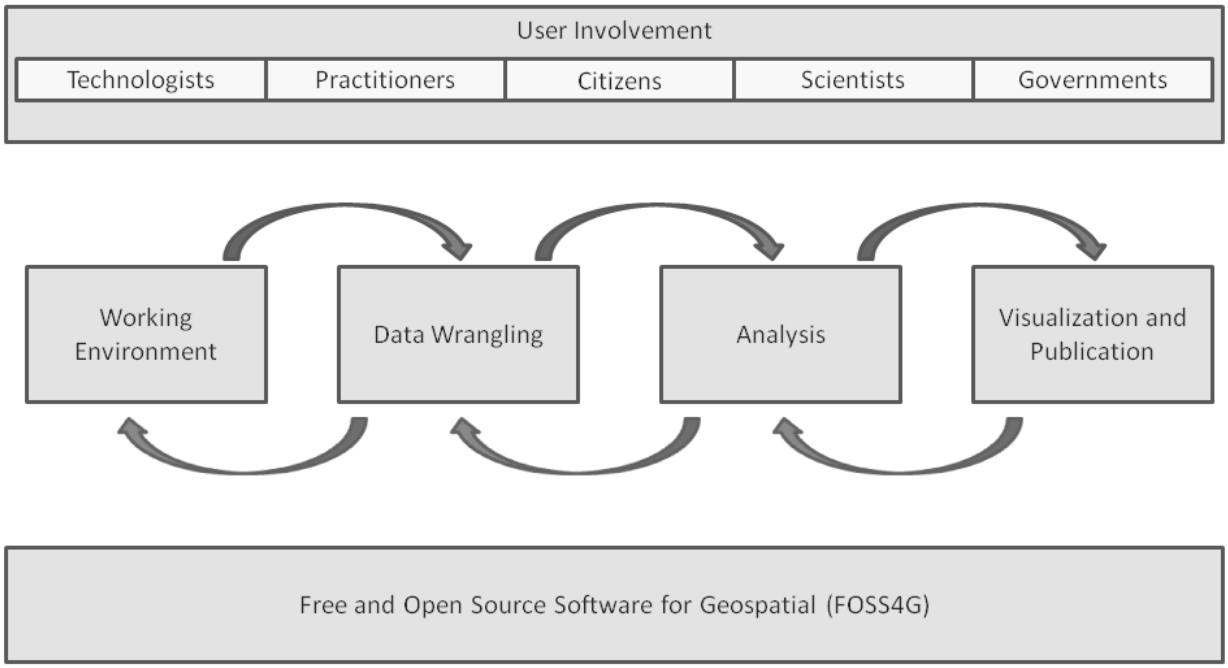


Figure 2-2. Collaborative Spatial Data Science Workflow

Typologies of geospatial tools and collaboration

Existing typologies (or classifications based on general type) of geospatial tools have been created through qualitative categorization and comparison of tool capabilities. To date, reviews of tool applications have been conducted to identify technical approaches to building tools, domain-specific evolutions in tools, and capabilities of tools for specific applications (Table 2-1). Although these qualitative typologies provide fundamental understanding of the evolution and landscape of geospatial tools, none provide a framework for evaluating how tools address technical barriers to collaborative tasks.

Table 2-1. Existing Typologies of Geospatial Tools

Technologies Reviewed	Publication
Web-based spatial decision support tools	Rinner (2003)
"Trends and developments in GIS-based multi-criteria decision analysis" tools	Malczewski (2006), p. 1
"Visually-enabled geocollaboration" tools	MacEachren and Brewer (2004), p. 1
GI Systems for public participation	Sieber (2006)
Historical evolution in web tools for geospatial applications	Haklay (2008)
Historical evolution of Participatory GIS	Jankowski (2009)
FOSS4G tools for landscape ecology	Steiniger and Hay (2009)

Marine Geospatial Ecology Tools	Roberts et al. (2010)
"Enabling technologies" for CyberGIS	Yang et al. (2010), p. 266
Evolution of software for spatial analysis	Anselin (2012)
"Domains of VGI"	Elwood et al. (2012), p. 573
"Capabilities and interfaces of existing tools" for GIS and spatial analysis tools integrated within CyberGIS	Wang et al. (2013), p. 2026
FOSS4G landscape in 2012	Steiniger and Hunter (2013)
"Major classes of technology tools and needs they might meet", among them being GI Systems, decision support tools, visualization tools, and "distance collaboration" tools	Cravens (2014), p. 23
"Map-based web tools supporting climate change adaptation"	Neset et al. (2016), p. 1

Though specifically highlighting only FOSS4G options, Steiniger and Hunter (2013) have provided the most comprehensive qualitative typology of GI Systems (GIS) software to date, expanding to nine categories of software from the seven original types identified in Steiniger and Weibel (2009): (1) desktop GIS; (2) spatial database management systems; (3) server GIS; (4) mobile GIS; (5) exploratory spatial data analysis tools; (6) remote sensing software; (7) GIS libraries (i.e. projection and geometry libraries); (8) GIS extensions, plug-ins, and APIs; and (9) Web Mapping Servers and Development Frameworks (p. 136 and 139). In addition to this fundamental qualitative typology, the authors also identify benefits of FOSS4G, key factors to consider for evaluations of options, and the primary barriers to FOSS4G adoption (referencing others such as Cruz, Wieland & Ziegler, 2006 and Nagy, Yassin & Bhattacherjee, 2010).Not identified as key functionality, the potential to support collaboration is not addressed in the criteria for evaluation or adoption.

The literature focusing on categorizations of tools with an explicit focus on collaborative work have been broader in scope and not specific to geospatial options. In support of geocollaboration tools, MacEachren and Brewer (2004) provided a summary of Computer-Supported Cooperative Work tools, or "CSCW technologies, often called groupware…characterized as information technology that allows people to work together… with an emphasis on sharing tasks and decision-making" (p. 10). In the listed categories of CSCW tools, multi-criteria evaluation tools integrated with GIS are the only geospatially enabled options. Similarly, in discussing the evolution of web mapping technologies, Haklay et al. (2008) presented a general "series of 'technologies of collaboration'" from Saveri, Rheingold, and Vian (2005), including "Self-organising mesh networks…Community computing grids…Peer production networks…Social mobile computing…Group-forming networks…Social software…Social accounting tools…Knowledge collectives" (p. 2025–2026).

While some geospatial tools can be embedded within these broader typologies of collaborative technologies (i.e. OpenStreetMap and other FOSS4G tools are products of peer production networks), I expand on these works by providing a new typology of geospatial tools that is specifically centered on collaboration. I ask the following specific questions: what are the common types of collaborative geospatial tools, and what functional niches do they fill? To answer these questions, I develop a quantitative and reproducible framework to evaluate multi-user geospatial tools based on their functionality for supporting common tasks in collaborative projects (i.e. wrangling, analyzing, visualizing, and publishing geospatial data) and present a typological map of the emergent ecosystem of collaborative geospatial tools.

Methods

Selection of Tools

I evaluated thirty-one multi-user geospatial tools based on their functionality to support collaborative tasks (Table 2-2). The tools represent a variety of platform types: cloud-based (i.e. hosted on the cloud by the tool provider), web-based (i.e. hosted by user on a web server), local installation (i.e. installed locally on an individual computer or cluster of computers), and mobile (i.e. application installed on mobile device). The tools also vary in their FOSS status and in the industry type of their primary creators and contributors. Specifically, the included tools express a mission of supporting collaboration and/or offer functionality for supporting collaboration (e.g. sharing of data and code, asynchronous tasks, status updates). This requirement excludes tools focused on big data processing such as distributed computation engines (e.g. Spark) or scenario modeling such as ecosystem valuation tools (e.g. Integrated Valuation of Ecosystem Services and Tradeoffs, or InVEST). The included tools also provide a set of analytical and/or data collection functionality within a multi-user environment (e.g. beyond basic online data providers or interactive web maps such as Cal-Adapt, WorldMap, etc.). This requirement also excludes tools focused primarily on workflows by individuals (e.g. desktop GIS tools such as GeoDa). In addition, included tools provide an out-of-the-box user interface and do not require the creation of a custom user interface or the use of a third party user interface. This definition excludes tool extensions such as widgets and plug-ins, which are not considered to be distinct from the platform onto which they are installed. This requirement also excludes geostack components whether open source or not (e.g. ArcGIS Server, OpenLayers, PostGIS, Leaflet). Tools currently in Beta mode were also excluded (e.g. GeoGig, a promising versioning tool). Finally, multi-user tools not exclusively limited to geospatial tasks were also included, if the stated criteria were met and the tool was able to integrate geospatial tasks (e.g. Jupyter Hub, RShiny). Though not an exhaustive list, the thirty-one tools evaluated in this chapter are representative of the wide range of available platforms that support multi-user workflows for geospatial data.

Table 2-2. List of Multi-user Geospatial Tools Included In Analysis

Label	Name	Platform Type	Creators/Contributors	FOSS status*
T1	CARTO	Cloud-based	CARTO (private sector)	Limited free, not open source
T2	MapGuide	Web-based	OSGeo (non-profit)	FOSS
T3	XchangeCore	Web-based	National Institute for Hometown Security (non-profit)	Free (restricted access), not open source
T4	Jupyter Hub	Web-based	NumFOCUS Foundation (non-profit)	FOSS
T5	NASA NEX sandbox	Local install	NASA (public sector)	Free (restricted access), not open source
T6	OS Geo Live	Local install	OSGeo (non-profit)	FOSS
T7	ROpenSci	Local install	Project of the NumFOCUS Foundation (non-profit)	FOSS
T8	Rshiny	Local install or cloud-based	RStudio (private sector)	Limited free, limited open source
T9	Global Forest Watch	Cloud-based	World Resources Institute (non-profit)	FOSS
T10	NextGIS	Local install or cloud-based	NextGIS (private sector)	Limited free, limited open source
T11	QGIS Cloud	Local install or cloud-based	Sourcepole (private sector)	Limited free, limited open source
T12	FME	Local install or web-based	Safe Software (private sector)	Neither free nor open source
T13	Google Earth Engine	Cloud-based	Google (private sector)	Free, not open source
T14	Madrona	Local install or web-based	Ecotrust (non-profit)	FOSS
T15	MapBox Studio	Cloud-based	MapBox (private sector)	Limited free, limited open source
T16	Field Papers	Cloud-based	Stamen Design (private sector)	FOSS
T17	iNaturalist	Mobile	California Academy of Sciences (non-profit)	FOSS
T18	OpenDataKit_ GeoODK	Mobile	University of Washington, Seattle (academia)	FOSS
T19	OpenStreetMap	Cloud-based	OpenStreetMap Foundation (non-profit)	FOSS

T20	eBird	Cloud-based	Partnership between Audubon (non-profit) and Cornell University (academia)	Free, not open source
T21	GeoLocate	Local install or cloud-based	Tulane University (academia)	Free, not open source
T22	HOLOS	Local install or cloud-based	University of California, Berkeley (academia)	FOSS
T23	Data Basin	Cloud-based	Conservation Biology Institute (non-profit)	Free, not open source
T24	ESRI Collector for ArcGIS	Mobile	ESRI (private sector)	Limited free, not open source
T25	Geopaparazzi	Mobile	HydroloGIS (private sector)	FOSS
T26	Locus Map	Mobile	Asamm Software (private sector)	Limited free, not open source
T27	Orux Maps	Mobile	OruxMaps (private individuals)	Free, not open source
T28	ArcGIS Online	Cloud-based	ESRI (private sector)	Limited free, not open source
T29	Seasketch	Web-based	University of California, Santa Barbara (academia)	Neither free nor open source
T30	AmigoCloud	Cloud-based	AmigoCloud (private sector)	Limited free, not open source
T31	ArcGIS Open Data	Cloud-based	ESRI (private sector)	Limited free, not open source

* see Appendix 2-1 and 2-2 for more information on FOSS status

A workflow-based evaluation of functionality

The included tools were evaluated on twenty-nine different features that support multi-user workflows for geospatial data (Appendix 2-1). Based on a collaborative Spatial Data Science workflow (Figure 2-1), these features represent functionality provided to address traditional technical impediments to collaboration and are organized into groups that represent the key components of the workflow: (1) setting up the working environment; (2) data wrangling; (3) data analysis; (4) data visualization and publication; (5) the integration and support of FOSS4G; and (6) user involvement. A standardized scoring rubric was used to assign a value of 1–3 for each feature, with 1 indicating little to no application of that feature within the tool to 3 indicating that the feature is critical to the functionality of the tool (Appendix 2-1). As no single tool can provide functionality for all features, I chose to treat the features as categorical variables referred to as factors, instead of as continuous variables. As such, tools with scores of 1 for particular features are not automatically clustered more closely with tools scoring 2, than with tools scoring 3; each score is simply considered to represent a different level of functionality. For example, all tools were scored on their reliance on cloud/web-based functionality. Tools that rely primarily on local installations (i.e. Desktop, Server, and Mobile) were given a score of 1, while tools that are completely cloud/web-based (i.e. no local installations of any kind) were given a score of 3. Tools that have both local and cloud/web-based components were given a score of 2. As factor variables, this scoring does not promote a score of 3 as more desirable for collaboration than a score of 1 or 2; these scores simply indicate different functionality based on the level of cloud/web-based integration. Last, the scores for all features were based on the mission statement or stated capacity provided on the tool website as well as my professional experience. All of the included tools that are available for download or online access were tested by the authors; the exceptions being Seasketch, NASA NEX, and XchangeCore, as these tools require granted permissions to download or access. When a feature could not easily be scored using online references or professional experience, questions regarding those features were sent to the tool provider, with a 100% response rate. The individual tool scores were also used to calculate an average score across all tools (i.e. an average tool score) for each of the twenty-nine features (Appendix 2-2).

Cluster Analysis

Next, I applied a statistical clustering method on the individual tool scores to determine the common typologies among tools and identify existing niches. My clustering method uses a custom R package called Threshold Smoothing Ensemble Clustering (TSEC or TSEClustering, developed by Oliver C. Muellerklein),which incorporates the Weighted K-Means Clustering method from the R package wskm (Zhao, Salloum, Cai, & Huang, 2015). The TSEC model clusters the observations (i.e. tools) and variables (i.e. features) based on an ensemble of co-occurrences across pre-defined subsets of the features using a smoothing threshold function. Conceptually, the final cluster memberships are the result of a threshold approximated ensemble

of similarities between the tools across subsets of the features (see Appendix 2-1 for list of features and subsets). This method is targeted at datasets with low sample size yet a high number of variables, building an ensemble of similarity measurements used to assess intra-cluster and inter-cluster variance for optimized information gain. Further, cluster membership weights are generated for observations at local (i.e. for subsets of features) and global (i.e. complete list of features) levels of variable importance. In sum, this unsupervised approach generates similarities among the observations based on various subsets of the features (i.e. subsets of predictor space) to assign clusters, concluding with a final ensemble of all cluster assignments.

The cluster analysis workflow is shown in Figure 2-3. The workflow begins with (a) input data of n × p (i.e. number of observations x number of features, or predictors). From the input data, six subsets of predictor space (i.e. gi, predefined subsets of the twenty-nine features) are created manually by splitting the features into groups that represent one of the six components of the collaborative Spatial Data Science workflow (i.e. from the first subgroup describing the working environment to the sixth subgroup outlining user involvement, see Appendix 2-1). Six additional subsets of predictor space are created by iteratively grouping all subsets except the initial gi subset (i.e. for all gx not gi, the leave one-out method). A final subset of predictor space is created by grouping all gi subsets (i.e. the complete list of twenty-nine features).

Intra-observational similarity matrices (b) are generated through correlation matrices for each of the thirteen subsets of predictor space (i.e. gi). Then, using entropy-weighted K-means clustering (c), cluster assignments are produced for each of the thirteen similarity matrices. The Elbow Method is used to obtain the optimal number of clusters by calculating the relative percentage of variance captured by the clusters versus the total number of clusters (Tibshirani, Walther, & Hastie, 2001). Next, observational co-occurrences (d) are generated for each of the thirteen K-means cluster memberships to obtain an ensemble of probabilistic assignments (e) (i.e. a correlational matrix of co-assignment among observations). These co-assignments are displayed as a dendrogram, a visualization of the partitions from hierarchical clustering (Figure 2-4). Threshold approximation is used to dropout low pairwise observational assignments (i.e. force to zero). Finally, a last K-means clustering (f) is run on the resulting Smooth Ensemble, an n × n correlational matrix of observational assignments, to generate the final cluster memberships. These final cluster memberships are visualized in a bivariate cluster plot, which uses Principal Components to make a two dimensional representation of the clusters (Figure 2-5). The final cluster assignment for each of the thirty-one tools is listed in Table 2-3.

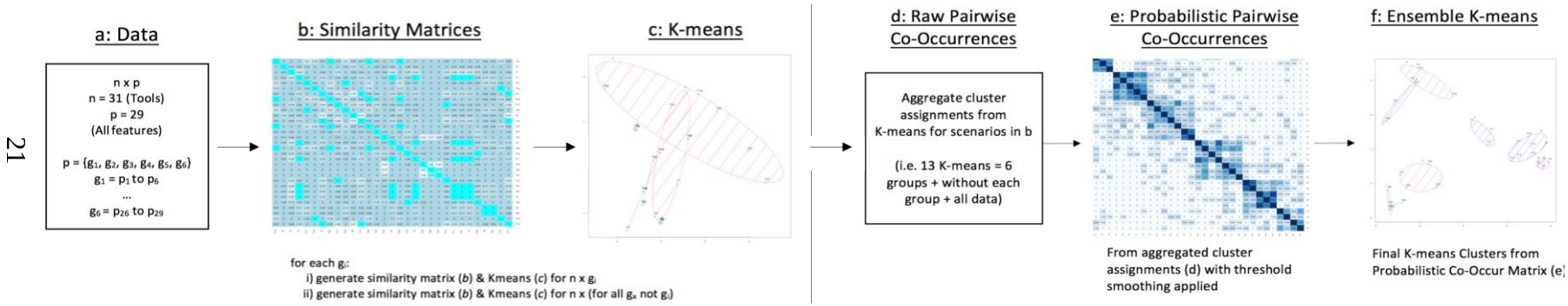


Figure 2-3. Cluster Analysis Workflow. Steps a-c generate a similarity matrix and K-means cluster assignment for thirteen subsets of features (i.e. each of the six g_i, iterative grouping of all g_x except the intial g_i, and for all g_i as a complete dataset). The resulting thirteen similarity matrices and K-means cluster assignments are aggregated in steps d-f with a threshold smoothing function to produce a final ensemble K-means cluster assignment.

Results

Three primary clusters composed of eight secondary clusters were revealed from the cluster analysis (Table 2-3). The divergences between the clusters are visualized in complementary ways in a dendrogram (Figure 2-4) and a map of the K-means Two-Dimensional Space (Figure 2-5). The first primary cluster A, composed of subclusters 1–3, contains tools that are highly scalable and customizable, allowing for the highest integration of advanced spatial analysis and visualization techniques as well as interoperability with other tools. The second primary cluster B, composed of subclusters 4–5, demarcates participatory data aggregators that have inherently large project scopes (i.e. functionality is optimally leveraged with high numbers of public users). The third primary cluster C, composed of subclusters 6–8, identifies content managers, or project-based tools focused on managing access to data and tasks for a predefined set of users.

Table 2-3. Summary of Cluster Results

Label	Name	Primary Cluster	Secondary Cluster
T1	CARTO	A	1
T2	MapGuide	A	1
T3	XchangeCore	A	1
T4	Jupyter Hub	A	2
T5	NASA NEX sandbox	A	2
T6	OS Geo Live	A	2
T7	ROpenSci	A	2
T8	Rshiny	A	2
T9	Global Forest Watch	A	3
T10	NextGIS	A	3
T11	QGIS Cloud	A	3
T12	FME	A	3
T13	Google Earth Engine	A	3
T14	Madrona	A	3
T15	MapBox Studio	A	3
T16	Field Papers	B	4
T17	iNaturalist	B	4

T18	OpenDataKit_GeoODK	B	4
T19	OpenStreetMap	B	4
T20	eBird	B	5
T21	GeoLocate	B	5
T22	HOLOS	B	5
T23	Data Basin	C	6
T24	ESRI Collector for ArcGIS	C	6
T25	Geopaparazzi	C	7
T26	Locus Map	C	7
T27	Orux Maps	C	7
T28	ArcGIS Online	C	8
T29	Seasketch	C	8
T30	AmigoCloud	C	8
T31	ArcGIS Open Data	C	8

Primary drivers of divergence

The first bifurcation between the clusters emerges from the required level of infrastructure needed to best leverage the tool (i.e. required user setup, use of high performance computing and cloud services, support for multi-tier users, user knowledge needed to extend functionality). This bifurcation is reflected along Component 1 of the K-means Two-Dimensional Space (Figure 2-5), wherein tools with the heaviest infrastructure needs are clustered on the right-hand side (primary cluster A of highly scalable and customizable tools), while tools with lighter infrastructure requirements are clustered on the left-hand side (primary cluster B of the participatory data aggregators and primary cluster C of the content managers).

A second key divergence between the clusters is driven by user involvement, a key determinant of project scope (i.e. optimal number of users and public accessibility). The K-means Two Dimensional Space reflects this divergence along Component 2 (Figure 2-5). Tools with inherently larger scopes are clustered toward the top (i.e. primary cluster B of the participatory data aggregators). The functionality of these tools is best leveraged with high number of public users engaging in data collection. Tools with smaller scopes due to a focus on managing user access to data and tasks are clustered toward the bottom (i.e. primary cluster C of the content managers). The functionality of these tools does not vary with a change in the number of users, and access to these data is controlled by a project manager. Along the center of Component 2 are

the highly scalable and customizable tools of primary cluster A. For these tools, an increase in the number of users leads to a leveraging of expandable functionality that is not necessary for small user groups, such as differential access to datasets and workflows facilitated by custom web visualizations and APIs (i.e. multi-tier versions of the tools with differing functionality and access based on the user type).

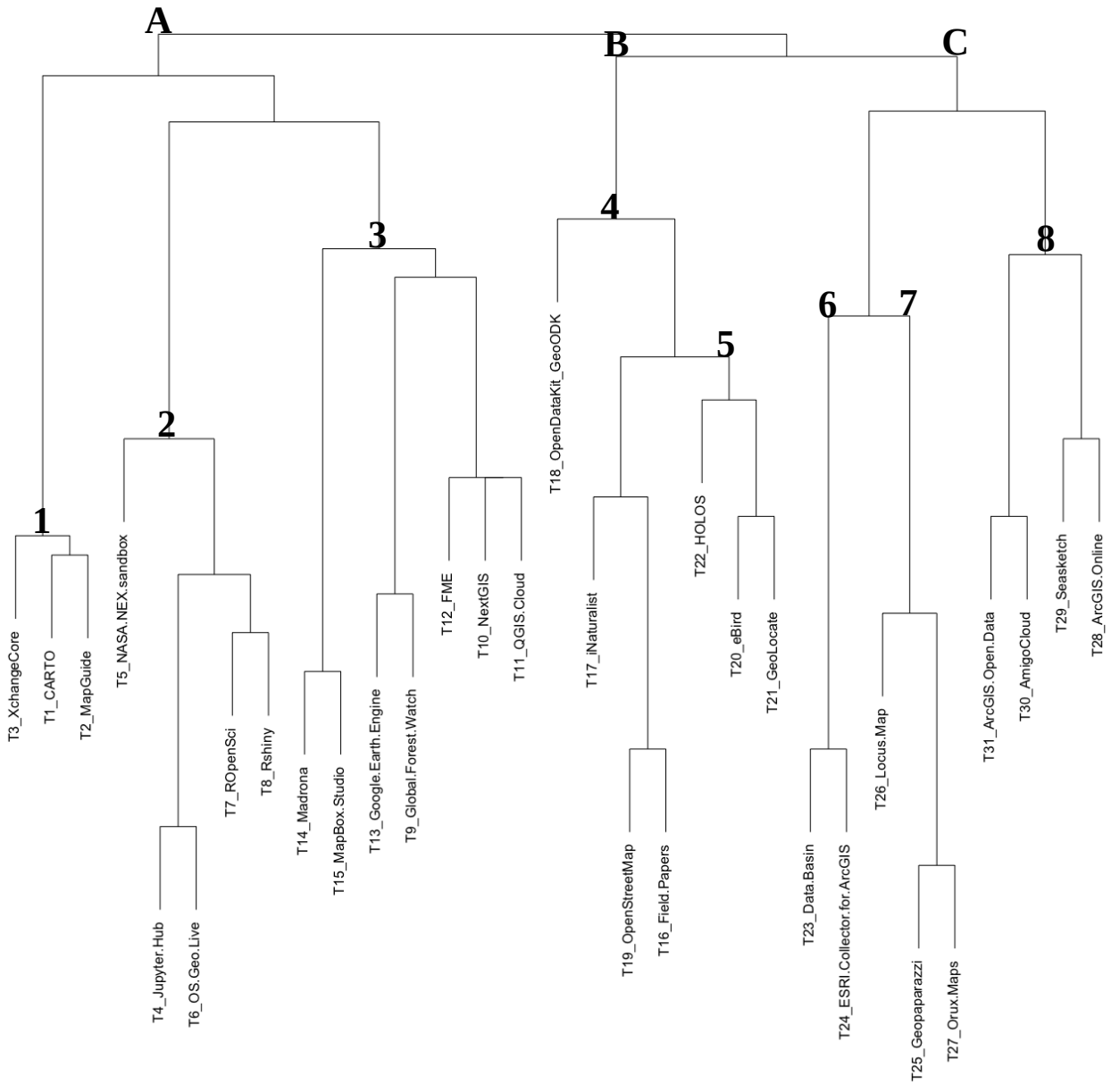


Figure 2-4. Dendrogram of Collaborative Geospatial Tools. Primary clusters are designated as A-C, with secondary clusters labeled as 1-8.

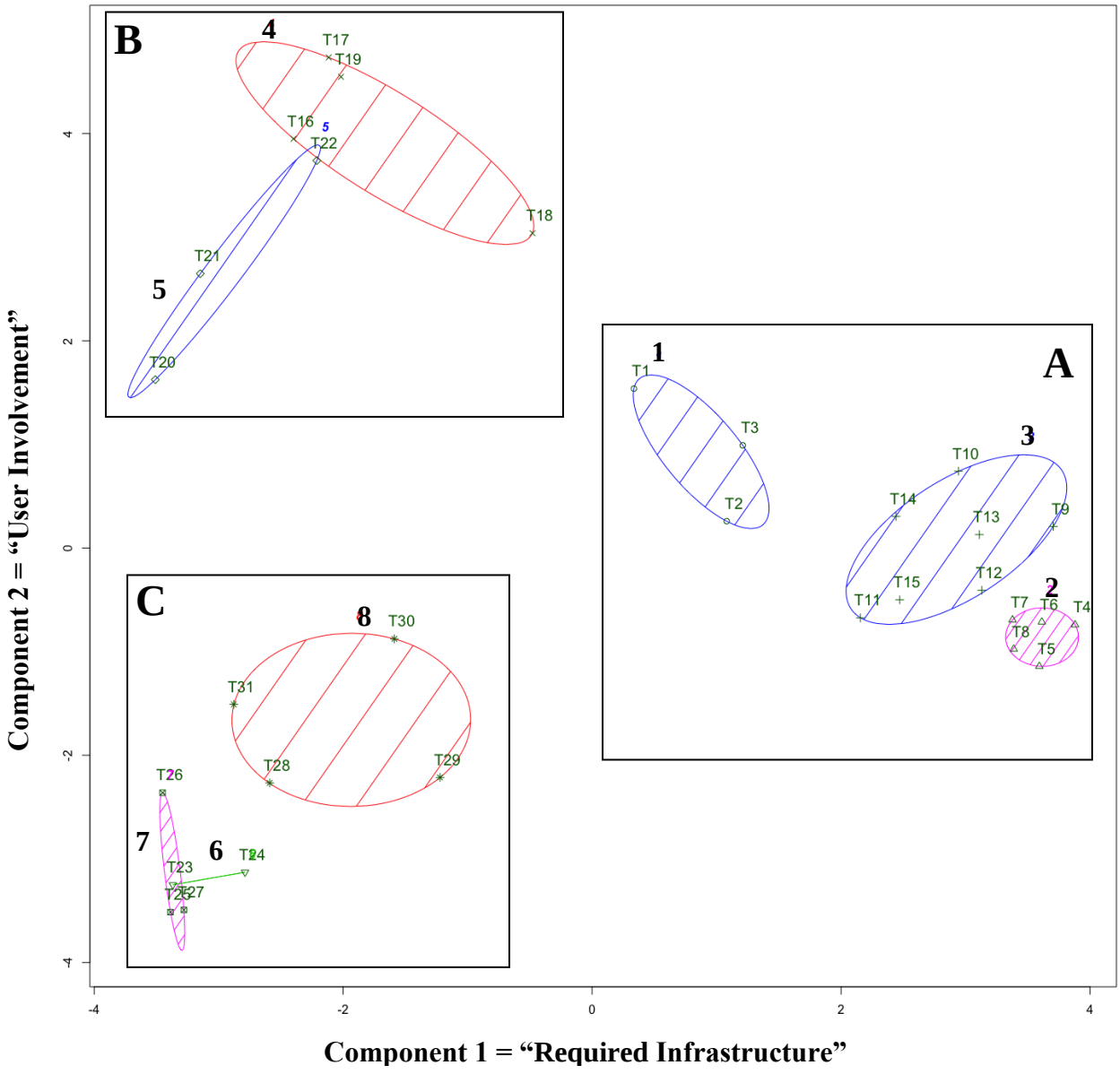


Figure 2-5. Typological Map of Collaborative Geospatial Tools, based on the K-means Two-Dimensional Space. Bifurcation driven by required infrastructure is reflected along Component 1, from low (left) to high (right). Divergence driven by user involvement (i.e. optimal number of users and public accessibility) is reflected along Component 2, from project-based content managers (bottom) to participatory data aggregators (top). The identified clusters account for 76.3% of the total variance.

Highly scalable and customizable tools

Primary cluster A (subclusters 1–3) is characterized by tools that are the most extendable for integrating advanced analysis and data visualization and for supporting reproducible code and interoperability through APIs and tool integration. Due to this flexibility, these highly scalable and customizable tools also require the most infrastructure to best leverage the full range of functionality offered. The primary products of these tools are datasets and workflows resulting from advanced analytical, data visualization, or querying methods. As the user base increases, the functionality provided (particularly by APIs and tool integration) can be leveraged to create multiple versions of the tools based on user needs (i.e. from low to high interactivity and privileges). The divergences between subclusters 1–3 are driven by differences in out-of-the-box functionality. Subcluster 1 highlights cloud/web-based tools (with APIs available) that have more built-in functionality for data exploration, visualization, and publication than for spatial analysis and querying (CARTO, MapGuide, XchangeCore). Subcluster 2 identifies tools with a stronger focus on facilitating reproducible workflows or standardized working environments (Jupyter Hub, NASA NEX, OS Geo Live, ROpenSci, RShiny). These tools require either server or desktop installed components and are best leveraged by integration with additional tools and packages for spatial data analysis and visualization. Subcluster 3 differentiates tools with the highest built-in functionality for geoprocessing or spatial analysis (Global Forest Watch, NextGIS, QGIS Cloud, FME, Google Earth Engine, Madrona, MapBox Studio). These tools provide APIs or expose open source capabilities to users, and thus, easily integrate custom scripting or can be expanded to build new tools in a multi-tier user environment.

Participatory data aggregators

Primary cluster B (subclusters 4 and 5) represents crowdsourcing tools that have inherently large and public scopes. The functionality of these tools is optimally leveraged with a high number of users. Often with a specifically defined focus (i.e. crowdsourcing data for a particular ecological phenomenon), the primary products of these tools are aggregated datasets compiled from many public contributors. The divergence between subclusters 4 and 5 is primarily driven by the differing roles of citizen scientists. Subcluster 4 delineates FOSS4G tools focused on crowdsourced data collection by the public either in real-time or asynchronously from the field or based on lived experience (Field Papers, iNaturalist, OpenDataKit/GeoODK, OpenStreetMap). Subcluster 5 represents research-driven participatory tools that are more focused on expert data curation by scientists (i.e. no mobile applications or syncing of field data) and provide APIs to engage and exchange data with the public (eBird, GeoLocate, HOLOS).

Content managers

Primary cluster C (subclusters 6–8) delineates project-based tools that focus on the management of users and their access to data and tasks. The primary products of these tools are content management systems (some with supporting APIs) controlled by a project manager. The

divergences between subclusters 6–8 are driven by differences in data management functionality and infrastructure. Subcluster 6 characterizes tools with functionality for managing projects and tasks (i.e. organization of workspaces, group communication tools, assignment of tasks), in addition to managing user access to data (Data Basin, ESRI Collector for ArcGIS). Subcluster 7 designates Android-based mobile data collectors that provide functionality for navigation and surveying (Geopaparazzi, Locus Map, Orux Maps). These tools allow a predefined set of users to collect geospatial data asynchronously and are not restricted to cloud-based databases. Finally, subcluster 8 identifies tools with light spatial analysis or querying capabilities that rely on "live" databases (i.e. cloud/web-based database services) for managing the exchange of data (ArcGIS Online, Seasketch, ArcGIS OpenData, AmigoCloud).

Discussion

Based on previously cited calls for more collaboration in geospatial research and technologies, it is clear that evaluation of geospatial tools must also include how they facilitate collaboration in the wrangling, analysis, visualization, and publication of geospatial data. Previous typologies have qualitatively categorized and compared geospatial tools without explicit consideration of the functionality provided to support collaborative tasks (see Table 2-1). By providing an essential assessment of geospatial tools specifically centered on functionality for collaboration, this typology can help geospatial researchers and stakeholders of collaborative geospatial projects evaluate and choose the best tools for their needs. By following the Spatial Data Science tenet of standardized and reproducible workflows, this typological map can evolve and expand over time, as more collaborative geospatial tools continue to be developed and adopted. In this paper, I use this typology to highlight the strengths of existing collaborative tools, identify key areas of future technical development, and elucidate ongoing challenges for collaborative geospatial tool development.

Strengths of collaborative geospatial tools

Even as the ecosystem of collaborative geospatial tools continues to expand, research focused on global environmental change are already benefiting from the existing technical strengths of these tools. Key benefits stem from increased integration of open source technologies as well as from an increased focus on interoperability through APIs and integration across tools. Users benefit not only from the cost-effectiveness of additional functionality from open source integration and interoperability with other tools, but also from being able modify and expand on these built-in open source capabilities and APIs.

Overall, the scored tools range from a moderate to high level of open source integration on the backend, whether a mix of open and closed source technologies to completely built on open source technologies (average tool score=2.61, along a gradient in which 1.0 indicates no integration of open source technology and 3.0 indicates completely open source). Similarly, the

scored tools support a range of moderate to high level of interoperability, from being able to push or pull data from other tools to providing a fully open API (average tool score = 2.68).

These high scores for open source integration and interoperability (as compared to all other scored features) provide a clear understanding of the role of cost and accessibility, as the resulting clusters do not represent groupings based on the price of the tools. Each cluster is a mix of free and/or completely open access tools to "freemium" (i.e. cost applied to access higher levels of functionality) and/or restricted access tools (i.e. domain specific applications that are free once access is granted, such as XchangeCore and NASA NEX from primary cluster A). Focusing solely on the cost of the tools as the key barrier to collaborative projects would result in a very different typology of collaborative geospatial tools, one that would not fully account for the niches of technical functionality that tools provide for collaborative tasks.

In my analysis, the individual clusters reflect differences in the level of open source integration, the level of modification allowed by users, and overall interoperability (see Appendix 2-2 for individual tool scores). Primary cluster A of the highly scalable and customizable tools demarcate tools that are primarily built on open source technologies, provide strong support for users to modify built-in open source capabilities, and generally, are highly interoperable either through the ability to integrate with other tools or through APIs. On the other hand, primary cluster C composed of content managers represent tools that are primarily a mix of open and closed source technologies, provide less support for users to modify the built-in open source capabilities, and as such, are generally less interoperable (with a few exceptions of tools that provide API access such as AmigoCloud, ArcOpenData and Locus Map). This is perhaps unsurprising as the strength of content managers are built-in functionalities that do not require much modification or technical knowledge by users (i.e. support for asynchronous tasks - such as offline capture of data, task assignment, and status updates - and user/content management - such as access/workspace control and group definitions). Primary cluster B of the participatory data aggregators is more evenly split; about half of the tools are mixed open and closed source that provide less support for modification by users (similar to primary cluster C), while the other half are primarily built on open source technologies and provide mechanisms for users to modify open source functionality (similar to primary cluster A). As compared to primary clusters A and C, the distinguishing characteristic of primary cluster B is that all of its tools scored the highest value (3.0) for interoperability (i.e. primary clusters A and B had wider ranges of scores). In fact, all participatory data aggregators included in this analysis provide access to APIs and/or Software Development Kits (SDKs).

For research centered on global environmental change, increased open source integration and support for interoperability in collaborative geospatial tools are allowing for unprecedented cross-disciplinary integration of data and methods, beyond simply powerful data processing or spatial analysis capabilities. Citizens with varying levels of technological skillsets (from non-scientists to practitioners) are commonly leveraging these strengths through access to source

code on Github and public API access to data and analytical methods. On the data side, citizen science projects that leverage participatory data collection tools are becoming more cost effective due to the availability of low-to-no cost, easy-to-launch tools that require little infrastructure investment or technical knowledge by users. Through these citizen science efforts, researchers are granted a mechanism for "dovetailing research with conservation and management" (Dickinson et al., 2012, p. 294). On the methods side, repeatability of complex workflows is facilitated by increased availability of APIs that allow for seamless exchange of geospatial data and by the ability to integrate functionality from other specialized tools.

A key example of open source integration benefitting citizen science efforts is the application of the mobile geospatial data collector iNaturalist from primary cluster B of the participatory data aggregators. This lightweight mobile application is increasingly being used in BioBlitzes, which are short-duration field collection efforts to inventory biodiversity or to monitor a particular species within a specified area, typically parks and protected areas (Dickinson et al., 2012; Francis, Easterday, Scheckel, & Beissinger, 2017). Citizen scientists simply download the free mobile application and capture photos and notes that automatically sync to the iNaturalist database. All data collected with iNaturalist are available for public exploration and use through their web mapping application and API and are also shared with free and open access scientific databases such as the Global Biodiversity Information Facility. In a unique global collaboration, National Geographic, the iNaturalist team, and citizens in 100 countries participated in The Great Nature Project between 2013 and 2015 to collect "over half a million images of over 20,000 different species of plants, animals, and fungi" (Francis et al., 2017; National Geographic, 2016).

Key examples of the benefits from increased interoperability are Global Forest Watch (of primary cluster A of the highly scalable and customizable tools) and Seasketch (of primary cluster C of the content managers). Global Forest Watch leverages the Google Earth Engine API and the CARTO platform (both tools also in primary cluster A of the highly scalable and customizable tools) to create interactive web maps that can analyze forest change on-the-fly for an area of interest. Building off of the Google Earth Engine API, Global Forest Watch freely provides its own customized APIs as well as templates for ArcGIS Online (in primary cluster C of the content managers) to facilitate additional tool building and data sharing by others. Leveraging the benefits of tool integration, Seasketch is a key example of a collaborative environmental planning tool that has benefited from integration with widely used spatial decision support tools such as Marxan and InVEST, as well as from integration with ArcGIS Online for content management. Focused on marine area protection, Seasketch is currently being used "around the globe by 4441 users in 229 active projects" to provide stakeholders with the capability to explore scenarios and propose their own plans for new marine protected areas (Seasketch, 2016). For example, through collaboration between Parque Nacional Galapagos, Conservation International, and World Wildlife fund, user-sketched plans from Seasketch are being integrated with the InVEST toolkit to allow public stakeholders to evaluate habitat risk and

explore outcomes of proposed zoning scenarios for marine protection around the Galapagos Islands (Seasketch, 2016).

Key areas of future technical development

In addition to highlighting the strengths of collaborative geospatial tools, my typology can help identify key areas of future technical development. One such area of needed development is the continued integration of cloud and high performance computing (average tool score = 1.7, along a gradient in which 1.0 indicates no integration of cloud and high performance computing and 3.0 indicates full integration). For collaborations centered around global environmental change, the leveraging of cloud and high performance computing can shift the cost-benefit structure, such that research questions that previously would have been very difficult or even possible to answer (due to computing time and resources) can now be addressed. Ongoing support of CyberGIS research as well as collaborations between scientists and technologists are key for continued integration of cloud and high performance computing into geospatial tools.

My analysis indicates that primary cluster A of the highly scalable and customizable tools has the highest overall application or potential for cloud and high performance computing (i.e. all tools scored at least 2.0, with the majority scoring 3.0). An exemplar of this cluster is Google Earth Engine, which was successfully leveraged to create the Hansen Global Forest Change dataset by a team consisting of fifteen collaborators, including technologists from Google, scientists from the USGS and Woods Hole Research Center, and researchers from the University of Maryland-College Park, SUNY-Syracuse, and South Dakota State University. Hansen et al. (2013) applied the distributed computing power of Google Earth Engine to map global forest loss and gain for 2000–2014 at the finest combined spatial and temporal resolution of any global product to date (yearly data at a 30mpixel resolution). A BBC News article quoted lead author Matt Hansen: "This is the first map of forest change that is globally consistent and locally relevant. What would have taken a single computer 15 years to perform was completed in a matter of days using Google Earth Engine computing" (BBC, November 14, 2013).

Another key area of future development is increased support for non-traditional raster and vector data formats (i.e. open data options, cloud-based tile services). Although these non-traditional data formats are becoming critical for environmental collaborations investigating questions of larger extents and finer resolutions, existing functionality to support these formats varies greatly depending on the data type and the task. For example, across all features scored, the highest average tool score is for data download of non-traditional vector formats (average tool score = 2.81), while the lowest average tool score is for data editing of non-traditional raster formats (average tool score = 1.39). Overall, average tool scores for non-traditional vector formats are higher than for non-traditional raster formats across all data tasks (i.e. creation, editing, upload, download), and for both non-traditional vector and raster formats, average tool scores for data uploads and downloads are higher than for data creation and editing.

Regarding discrepancies between non-traditional vector and raster formats, there are two primary contributing factors. First, increased integration of open source technologies and development of APIs have both lead to and been reinforced by stronger support and wider use of non-traditional vector formats such as GeoJSON, Vector Tiles, and MBTiles (average tool score = 2.55 for data uploads to 2.81 for data downloads). This dual reinforcement is not as strongly reflected within collaborative geospatial tools for non-traditional raster formats such as HDF5 and Tile Mapping Services or for older raster formats that are seeing a resurgence such as NetCDF (average tool score=2.32 for data uploads and 1.68 for data downloads). This could be driven by the fact that non-traditional raster formats are increasingly being used to cover larger extents and/or finer resolutions, resulting in larger datasets and storage needs, which are ongoing challenges for geospatial tools in general.

Second, the typical process for creating and editing raster data often differs greatly from that of vector data. Most raster data are still expert curated in single user environments, and unlike editing of individual features in vector data, editing of raster data typically involves global re-calculations of pixels for which GUI-based editing tools are not as useful. These differences in the curation of vector and raster data are reflected in the average tool scores. For both editing and creation, the average tool scores for non-traditional vector data are higher than for non-traditional raster (for editing, average tool score= 2.13 for vector compared to 1.39 for raster; for creation, average tool score = 2.42 for vector compared to 1.52 for raster).

Regarding the higher average tool scores for data uploads and downloads of both non-traditional vector and raster formats (as compared to editing and creation), these scores are reflective of the unique strengths of each primary cluster which focus on a different aspect of data management. For example, the highly scalable and customizable tools of primary cluster A provide flexibility and expandability for data integration, while the participatory data aggregators of primary cluster B provide infrastructure for data aggregation. Similarly, the content managers of primary cluster C provide strong built-in functionality for managing access by users to data, projects, and workflows.

Other key areas of needed tool development include the wider adoption of functionality to support reproducibility of workflows (i.e. sharing of code or steps of workflow, average tool score=1.9), custom scripting for analysis (average tool score=2.03) and for data visualizations (average tool score= 1.9), and the integration of time (average tool score =2.16). For tools that best support these options at present (i.e. primary cluster A of the highly scalable and customizable tools), users are able to modify open source capabilities or harness APIs to create tailored analyses and applications for a second tier of users. However, an intermediate to high level skillset in programming is often needed for leveraging these functionalities. In addition, stronger support for integrating time into analyses is an outstanding need in Spatial Data Science beyond that of collaborative geospatial tools, particularly for visualization and analysis across continuous timelines (i.e. dynamic modeling approaches). While of all geospatial technologies, remote sensing analytical tools have most successfully addressed time, these same tools are not

structured to provide multi-user support (with few exceptions such as Google Earth Engine), and typically function within discrete timelines. Similarly, support for reproducibility of workflows and results is a key component not only for collaborative geospatial workflows, but for Spatial Data Science as a field of study focused on repeatability and transparency of workflows.

A final key area of needed functionality is user controlled versioning of data and workflows; this feature was not scored in my evaluation, as so few of the representative tools offer this functionality. While many collaborative geospatial tools provide some light versioning capabilities (i.e. revision history of code in Google Earth Engine, the ability to create and compare different runs of a model in Seasketch, user contribution history for participatory data aggregators), what is not yet available is true distributed versioning of data, workflows and code that allows users to track changes at the object level (i.e. a data attribute or function), to reconcile conflicts that arise in competing edits (i.e. multiuser versioning), and to roll back changes as needed (i.e. adaptive management of data and workflows). In a fully versioned environment, all of these tasks are documented and available for review. For environmental planning and management projects, these kinds of native versioning capabilities would provide stakeholders with a structured and transparent mechanism for examining the trajectory of data and models and for actively contributing to their construction. Stakeholders could move from being primarily users of scenario exploration tools to active developers of them, as constructors of alternative stories and models beyond just offering their version of a controlled output map.

Of the key areas of future tool development presented in this paper, support for distributed versioning in collaborative geospatial tools is clearly in the earliest phase of its evolution. In general, collaborative geospatial tools have focused on other asynchronous tasks (i.e. off-line capture of data, task assignment, status updates; average tool score = 2.26) and user and content management (i.e. access and workspace control, group definitions; average tool score=2.32). As the integration of "live" databases through web and cloud-based data services continues to become more standard in collaborative geospatial tools, particularly for multi-user collecting of data (average tool score = 2.39), versioning capabilities can also continue to be expanded. Mechanisms for supporting further integration of versioning can be adopted from existing spatial database engines (i.e. ESRI ArcSDE, PostGIS) which offer versioning of geospatial data or allow it to be programmed, and from existing version control frameworks such as Git/Github, which has played a key role in the FOSS4G movement, allowing any user to contribute to and modify the source code to fix bugs and extend functionality. Tools such as GeoGig, a Git-like versioning tool for geospatial data (currently in Beta testing), can serve as a preview of collaborative geospatial tool functionality that will likely become standard in the near future.

Challenges and Future Directions

The technical challenges for continued development of collaborative geospatial tools parallel existing research areas within Spatial Data Science centered around issues inherent to large and complex geospatial datasets. While data mining techniques integrated from Data Science have

provided ways to turn massive data into usable information, analysis and visualization of large geospatial datasets remain difficult, as not all approaches scale appropriately (Anselin, 2012; Li et al., 2016). Visual analytics for spatial-temporal data is one area of research that aims to provide scalable methods for analyzing datasets that are too large to be contained within working memory (or random access memory, RAM). For example, Andrienko, Andrienko, Bak, Keim, and Wrobel (2013) outline a methodology for clustering of large movement datasets that begins with sub-setting the data and creating an iterative identification list of each event's neighbors that are "stored in the database, to be later retrieved on demand" (pg. 214). Similarly, Stange et al. (2011) employ various spatial and temporal filters and aggregations to prepare large movement datasets for clustering of trajectories and flows using data mining methods such as self-organized maps (SOM) and algorithms specific to mobile data. Additional solutions to the challenges posed by large datasets are being explored through the use of high performance computing environments for data wrangling and analysis (Leonard & Duffy, 2014; Li et al., 2016) and through geovisualization techniques integrated from visual analytics, or "geovisual analytics" (Anselin, 2012). These techniques include the use of multiple-linked views that allow users to work with multiple visualizations at once and human "vision-inspired" techniques such as foveation that aim to reduce information overload by varying detail depending on area of focus (Li et al., 2016, pg. 124). However, though advances in rendering have been made with emerging data formats (i.e. Vector Tiles, MBTiles, TileMapping Services), issues of optimizing geospatial data storage and querying remain. Tiles still require producer-side storage of raw data, and in general, spatial indexing techniques need to evolve for larger geospatial datasets, particularly in real-time applications (Li et al., 2016).

Even as computational techniques to extract and render information from data are improving, collaborative geospatial tools are limited by ongoing conceptual challenges to synthesizing information derived from large amounts of geospatial data. In particular, Miller and Goodchild (2015) point out that key issues resulting from the progression from a "data scarce to data-rich environment" are also longstanding challenges in geographic research: accuracy; uncertainty; representations of data and features; "populations (not samples), messy(not clean) data, and correlations (not causality)" (p. 450). While it is clear that these issues will continue to be ongoing challenges for both theory and technology, collaborative geospatial tools can serve as exploratory testing grounds of proposed solutions. For example, participatory data aggregators have already begun to integrate approaches to addressing issues of quality in VGI data such as biases in geographic coverage, user motivations, and knowledge levels (Quinn, 2015) through crowdsourced-based approaches (i.e. validation, repetition), social based approaches (i.e. trusted users), and geographic knowledge based approaches (i.e. spatial dependence and topological rules) (Goodchild & Li, 2012).

Developers of collaborative geospatial tools should also note ongoing concerns regarding the centralized production of technology and knowledge. It is clear that while collaborative geospatial tools are indeed becoming more interoperable and sophisticated, the development of

these tools require knowledge that is not equally shared, which serves as a barrier to including stakeholders in the tool development process. For example, Wright et al. (2009) explore how geospatial tools used in collaborative natural resource management projects can either reinforce the technical knowledge divide between scientists and the public or provide alternative ways for the citizens to engage in the storytelling process. In addition, through presenting a "hierarchy of hacking", Haklay (2013) identifies a key barrier to democratization within neogeography as the technical knowledge and skillsets needed for citizens to be empowered to create their own tools, in light of "the current corporatisation of the web" (p. 63). The author concludes that new geospatial tools have increased the access and use of geographic information only at the lower hacking levels; "the higher levels, where deep democratisation of technology is possible… require skills and aptitude that are in short supply and are usually beyond the reach of marginalised and excluded groups in society" (p. 67). Similarly concerned about corporate and top-down control of geospatial tool development, Miller and Goodchild (2015) argue: "We must be cognizant about where this research is occurring— in the open light of scholarly research where peer review and reproducibility is possible, or behind the closed doors of private-sector companies and government agencies, as proprietary products without peer review and without full reproducibility" (p. 460). Consistent with the concerns expressed in the literature, my analysis also indicates an overall high level of user knowledge needed to fully leverage the functionality offered by collaborative geospatial tools (average tool score = 2.55, along a gradient in which 1.0 indicates none needed and 3.0 indicates a high level needed). This score reflects the fact that many tools in primary cluster A of highly scalable and customizable tools and primary cluster B of participatory data aggregators provide both basic functionalities as well as capabilities for expansion of the tools by advanced users.

Looking into the future, continued development of collaborative geospatial tools requires a sustained focus on the eight dimensions of Open GIS proposed by Sui (2014): "Open Data, Open Software, Open Hardware, Open Standards, Open Research, Open Publication, Open Funding, and Open Education" (p. 4). In particular, Open Software and Open Standards have been critical for the previously highlighted strengths of collaborative geospatial tools: integration of open source technology and support of interoperability through tool integration and APIs. These aims are supported by ongoing evolution of Open Geospatial Consortium standards and other open data standards, combined with a renewed focus on standardized and queryable metadata (Sui, 2014). Similarly, Elwood et al. (2012) highlight that the required integration of data across differing formats and media can be a major challenge to data synthesis, which often "can only be achieved if systems are to a large degree interoperable" (p. 582). Steiniger and Hunter (2013) further argue for more open source APIs, as many popular "web-mapping tools work as black boxes and do not give users the freedom to study and modify them" (p. 145).

Finally, tools are but one component in the collaborative process, an iterative exercise in communication between people to "generate (ideas and options), negotiate, choose, and execute" solutions to community and global challenges (MacEachren & Brewer, 2004, p. 7). As such, the

process of stakeholders evaluating, implementing, and troubleshooting tools as a group may be more fundamental to the success of collaborative efforts than the functionality provided by the tools themselves. One likely reason is that while many environmental management and planning projects aspire to incorporate collaborative tasks (Cravens, 2016; Wright et al., 2009), tools are often chosen before project needs are understood, or are not evaluated until after projects are completed (Cravens, 2014). In addition, group discussion regarding the applicability and functionality tools can also serve a strong mechanism of stakeholder engagement, as the negotiation process can allow individuals to feel acknowledged and heard. While I have argued that tool functionality can be leveraged to provide technical support for collaborative tasks, future research can expand on this collaborative geospatial typology to focus on identifying which technical improvements are most critical for strengthening public engagement of non-scientists, particularly in the contexts of citizen science and collaborative environmental planning. It remains "a challenge for future research… how to combine computer technology with facilitation without stifling the creativity of participants" (Jankowski, 2009, p. 1971). In addition to focusing on expanding functionality, research can continue to explore additional ways that tools can empower stakeholders (i.e. further incorporation of theories of communication and decision-making, tool design, and user-computer interactions). As stakeholders become more involved in the applied process of technology design and creation, they can also highlight previously unrecognized barriers and impediments to collaboration (both social and technical) as well as help to redefine both conceptual frameworks and best practices for collaboration.

Conclusion

Spatial Data Science, which combines aspects of GI Science, Data Science, and CyberGIS, has emerged as an interdisciplinary field that supports collaborative geospatial research through an emphasis on leveraging cloud/web-based and open source geospatial tools that foster reproducible workflows and address long-standing technical barriers to collaboration. Here, I used a quantitative and repeatable approach to create an adaptable typology of collaborative geospatial tools based on their functionality for collaborative tasks. The resulting typological map reveals three key clusters composed of eight subclusters, across which divergence is driven by required infrastructure and user involvement. These clusters represent three primary types of collaborative geospatial tools: (1) highly scalable and customizable tools with heavier infrastructure needs, (2) participatory data aggregators and (3) content managers, the latter two with lighter infrastructure needs. As the process of collaboration is complex, one way (i.e. one cluster) is not better than another; these clusters represent discrete types of functionality that support communication and collaborative tasks for different needs and purposes. Overall, the development of a typology of collaborative geospatial tools can suggest key areas of future tool development and Spatial Data Science research, as well as help stakeholders evaluate tools by providing an understanding of the strengths of existing tools and highlighting areas of needed development. Thus, my example exploration of the emergent ecosystem of collaborative

geospatial tools is not only about tools per se; this work highlights the ongoing need to facilitate communication between scientists and stakeholders in order to support fruitful collaborations that address community and global challenges.

Chapter Three

Comparison of remotely-sensed vegetation disturbance products results in large differences in reported disturbance and representation of fire across California

Recent advances in high performance computing (HPC) have promoted the creation of standardized remotely-sensed products that map annual vegetation disturbance through two primary methods: (1) traditional approaches that integrate field data, public data on disturbance events, and vegetation indices derived from partially automated analyses, and (2) "big" data approaches based on automated HPC-based analyses. Given the recent proliferation of these annual products as well as their potential utility for understanding vegetation dynamics, it is important for product end-users (i.e. practitioners and researchers in domains other than remote sensing) to understand the differences in their representations of disturbance and the conditions under which they report it. I use fire in California as a case study to evaluate three widely used vegetation disturbance products created using different methods – LANDFIRE (representing the traditional approach), and Hansen Global Forest Change (GFC) and North America Forest Dynamics (NAFD), both created from automated approaches but with differing thresholds for reporting disturbance. Using Google's Earth Engine, I compared the reported amount of disturbance for 2001-2010 among these products and examined the products across differing environmental and burn conditions. I found that GFC reported the least amount of disturbed area in most years and across all habitat types, while LANDFIRE reported the highest amount across all years and habitat types. My comparison of environmental conditions (i.e. elevation, climate, habitat) did not reveal major differences in coverage by the products, but it did identify large differences in the coverage of burn conditions between GFC/NAFD (products created by automated methods) and LANDFIRE (the traditionally created product). Furthermore, I also identified the widest range in reported disturbance among the products (i.e. more uncertainty) in years with more fire incidence and in scrub/shrub habitat. Overall, rather than focusing on accuracy, my study can help end-users to evaluate these products based on the conditions under which they report disturbance, to understand these products as different representations of disturbance based on differing thresholds, and to identify drivers of uncertainty in reported disturbance.

Introduction

Recent advances in high performance computing (HPC; including distributed, parallel, clustered, and cloud-based methods) have provided new opportunities to analyze "big" remotely sensed data across broader spatial scales and finer temporal resolutions (Kalluri et al., 2015; Kang & Lee, 2016; Kumar et al., 2017; Lee, Gasster, Plaza, Chang, & Huang, 2011; Plaza & Chang, 2007). These HPC-based remote sensing analyses are increasingly being used to identify long-term vegetation changes using the Landsat Time Series (LTS) (Hermosilla et al., 2016; Soulard,

Albano, Villarreal, & Walker, 2016), with some efforts resulting in standardized maps of annual vegetation disturbance (i.e. annual changes in vegetation due to either natural and anthropogenic events) across the U.S. and globally (Goward et al., 2015; Hansen et al., 2013). These LTS-based products of annual vegetation disturbance have been primarily produced through two methods: (1) traditional approaches that integrate field data, data on disturbance events reported by public agencies, and remote sensing-derived vegetation indices from partially automated analyses, and (2) "big" data approaches based on algorithms and workflows that are automated with HPC to execute across a complete dataset (typically all LTS images collected for a given time period and spatial extent).

A key example of the traditional approach is Landscape Fire and Resource Management Planning Tools (LANDFIRE). Historically, LANDFIRE focused on providing spatially explicit data of canopy characteristics such as vegetation height and cover (Keane, Rollins, & Zhu, 2007; Reeves, Ryan, Rollins, & Thompson, 2009; Rollins, 2009; Ryan & Opperman, 2013). Through recent integration of vegetation indices calculated via HPC (such as differenced Normalized Burn Ratios (dNBR) to identify burned areas) and manual aggregation of data provided by public agencies on disturbance events (i.e. fire, harvest), LANDFIRE has released standardized products that report vegetation disturbance for each year between 1999 and 2014 across the entire U.S. In contrast to the traditional creation method of LANDFIRE, the big data approach relies on automated algorithms and workflows via HPC to identify annual changes in spectral signature (i.e. reflectance) at individual pixels across large multi-temporal stacks of images from LTS. Key examples are North American Forest Dynamics (NAFD) and Hansen Global Forest Change (GFC). Specifically, NAFD was produced using the Vegetation Change Tracker (VCT) (Huang et al., 2010) algorithm applied to the LTS within an HPC environment developed by NASA (NASA Earth Exchange, NEX) (Nemani, Votava, Michaelis, Melton, & Milesi, 2011). The resulting product is a collection of annual maps of vegetation disturbance across North America for 1984-2010 (Goward et al., 2015). In a similar vein, GFC was produced through collaboration between a group of academic researchers and Google that leveraged Google's cloud-based HPC infrastructure to produce an annual product of global forest change for 2000-2014, creating the first product to be mapped at the spatial and temporal resolution of LTS at a global extent (Hansen et al., 2013). Additional algorithms developed from the LTS, such as LandTrendr (Kennedy, Yang, & Cohen, 2010), Continuous Change Detection and Classification (CCDC) (Zhu and Woodcock 2014), and others (Cohen et al., 2017; Healey et al., 2017), may result in additional vegetation disturbance products in the future.

While there have been comparative evaluations of the algorithms used by remote sensing experts to map vegetation disturbance, such as VCT, CCDC, LandTrendr et al., (Cohen et al., 2017; Healey et al., 2017), there have been no systematic comparisons (at the time of this publication) of the vegetation disturbance products that are frequently being used by non-remote sensing experts (e.g. GFC, NAFD, LANDFIRE). Previously published papers evaluating vegetation disturbance products have focused on the accuracy or validation of an individual product

(Gudex-Cross, Pontius, & Adams, 2017; Hyde, Strand, Hudak, & Hamilton, 2015; K. Krasnow, Schoennagel, & Veblen, 2009; McKerrow, Dewitz, Long, Nelson, & Connot, 2016; Thomas et al., 2011; Tyukavina et al., 2015; Zhao et al., 2018; Zimmerman et al., 2013) or on integration of these products (or the algorithms used to create them) to improve the accuracy of disturbance identification (Healey et al., 2017, Schroeder et al., 2017, Soulard et al., 2017).

The lack of comparative evaluations is a notable omission in scientific literature, as these products (GFC, NAFD, and LANDFIRE) are widely used to study the impacts of vegetation change and disturbance on ecosystem processes such as carbon. For example, GFC has been used to examine the impacts of forest change on carbon dynamics both globally (Arneth et al., 2017; Tyukavina et al., 2015) and within the United States (U.S.) (Anderegg et al., 2016; Woodall et al., 2016). NAFD has also been frequently used to explore the impacts of forest disturbance on carbon dynamics within the U. (Dolan et al., 2017; Gu, Williams, Ghimire, Zhao, & Huang, 2016; Sleeter et al., 2018; Williams, Gu, MacLean, Masek, & Collatz, 2016). LANDFIRE has been applied more broadly across landscapes in the U.S. to explore impacts of past disturbance on hydrology (Boisramé, Thompson, Collins, & Stephens, 2017), subsequent fire(Parks, Miller, Nelson, & Holden, 2014) as well as carbon dynamics (Gonzalez, Battles, Collins, Robards, & Saah, 2015; Liu et al., 2011).

Given the recent proliferation of these annual products as well as their potential utility for non-remote sensing experts to explore the spatial-temporal impacts of vegetation disturbance on ecosystems, it is important to evaluate them to understand the differences in their representations of disturbance and the conditions under which they report it. To this end, this chapter targets end-users of remotely sensed vegetation disturbance products who are not remote sensing experts (i.e. practitioners and researchers of domains like ecology and conservation biology), but who rely on these products for their work and seek a greater understanding of how they identify and represent disturbance, beyond a discussion of accuracy. I provide a comparative evaluation of the three annual vegetation disturbance products derived from the LTS that have overlapping spatial and temporal extents at the time of this publication: Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE. I use fire in California across 2001-2010 as a case study to identify where and when these products identify disturbance, using two widely used reference datasets of fire across California: Monitoring Trends in Burn Severity (MTBS) and fire perimeters from the California Department of Forestry's Fire and Resource Assessment Program (FRAP) database. California is a fire-prone state, and large wildfires occur annually across the state, resulting in significant changes to forest, scrub/shrub and grass (Krasnow, Fry, & Stephens, 2017; Moritz & Stephens, 2008; Stephens, Martin, & Clinton, 2007), the three habitat types of focus in this study. I recognize that users are seeking these products to accurately characterize annual vegetation change and disturbance in their work, and this is the first comprehensive study to examine the key differences across these competing annual vegetation disturbance products. Rather than focusing on accuracy, this chapter compares the spatial and

temporal coverage of these products to identify the differences in the amounts and locations of reported disturbance.

Specifically, this chapter asked:

1. How comparable were the three vegetation disturbance products in their reported disturbance amounts across California, by year and by habitat type?
2. How different were the environmental conditions covered by the products (i.e. distribution of bioclimatic conditions and proportional areas for habitat types)?
3. How different were the burn conditions covered by the products (i.e. fire perimeter size and burn severity)?

To help end-users of these products better understand the differing methods used to create these products, I first review the key differences between the modern automated approaches of GFC and NAFD and the traditional creation approach of LANDFIRE and explain how their creation methods result in differing thresholds for reporting disturbance. For my comparative analysis of reported disturbance as well as environmental and burn conditions among the products, I employed Earth Engine (EE), a cloud-based, distributed HPC platform created by Google that provides a set of analytical functions for analyzing vector and raster-based geographic data via multiple cloud-based user interfaces (Gorelick et al., 2017). Even while limited to California, the disturbance products evaluated in this study are "big" data, as approximately 450 million pixels were analyzed for each year based on the LTS spatial resolution of 30 m. As such, I used the JavaScript API Code Editor to leverage the HPC capabilities of EE as well as the built-in functionality such as code-sharing and cloud data storage, which support reproducibility and collaboration (Palomino, Muellerklein, & Kelly, 2017). Overall, I believe my results can help researchers and practitioners to understand the impacts of disturbance identification methods (i.e. the traditional versus modern approaches) on the representation of disturbance, evaluate these products based on the conditions under which they report disturbance, and to choose the most appropriate data for their needs.

Study Data: Vegetation Disturbance Products

The three vegetation disturbance products included in this study are the only annual products that share an overlapping spatial and temporal extent at the time of this publication. A summary of the key differences in their creation methods are found in Table 3-1.

Table 3-1. Summary of Vegetation Disturbance Products

Disturbance Product	Time Period	Extent and Target Vegetation	Definition and Identification of Disturbance from LTS	Computing Environment
Hansen Global Forest Change (GFC)	2000-2014[1]	Global Forest	**Loss of cover (discrete):** "stand-replacement disturbance" leading to a non-forest state for the pixel (Hansen et al., 2013, supplemental material) **Identification method (automated):** supervised classification of forest loss; NDVI time series analysis to identify year of loss	Google Earth Engine (EE): cloud-based distributed computing platform (proprietary)
North American Forest Dynamics (NAFD)	1986-2010	North America Forest	**Disturbance of cover (continuous):** annual change in the integrated forest z-score (IFZ), an inverse measure of likelihood that a pixel is forested in a given year **Identification method (automated):** VCT algorithm applied to LTS, supplemented by dNBR analyses	NASA Earth Exchange (NEX): HPC cluster managed by NASA
LANDFIRE	1999-2014	United States All vegetation	**Loss and disturbance of cover (discrete and continuous):** depending on data integrated in that year **Identification method (manual integration):** year-by-year integration of disturbance events reported by public agencies and calculated indices from LTS including NDVI, dNBR, MTBS, VCT algorithm, and Multi-Index Integrated Change Algorithm (MIICA)	Custom multi-node HPC cluster managed by the USGS Earth Resources Observation and Science (EROS)

[1] first year of identifiable loss is 2001

Distributed HPC approach to identify discrete losses: Hansen Global Forest Change (GFC)

GFC (Hansen et al., 2013) is a key example of modern products that can be created at increasingly broader extents through the automation of standard remote sensing analyses on HPC. Specifically, GFC was produced as the first global-scale, annual forest loss product at spatial resolution of LTS (30 m) using Google's Earth Engine (EE), a cloud-based distributed computing platform that provides analytical capabilities of geospatial data (Gorelick et al., 2017). Within EE, a supervised classification process was conducted to identify locations of forest loss between 2000 and 2014, using training data of locations pre-labeled with known forest loss or no forest loss (i.e. discrete identification). As the baseline for forest, pixels containing tree cover with a height greater than five meters were identified as forested. In the training dataset, forest loss represented by pre-identified pixels that had experienced "stand-replacement disturbance" leading to a non-forest state for the pixel (Hansen et al., 2013). As such, forest degradation that did not result in a new cover type (i.e. any non-forest state) was not labeled as forest loss. The year of loss was identified through an analysis of a time series for Normalized Difference Vegetation Index (NDVI; an indicator of greenness calculated from the LTS bands for red and near-infrared); the year with the sharpest drop in NDVI was identified as the year of loss. The cause of loss, severity, and measurement of uncertainty were not provided.

Algorithmic approach to identify continuous disturbance with HPC: North American Forest Dynamics (NAFD)

NAFD (Goward et al., 2015) exemplifies a modern algorithmic approach to identifying continuous disturbance (i.e. reduction in vegetation cover) from the LTS, using the Vegetation Change Tracker (VCT) algorithm (Huang et al., 2010) on NASA Earth Exchange's computing facilities (Nemani et al., 2011). In contrast to the discrete disturbance (i.e. stand-clearing or replacement) reported by GFC, "the VCT approach captures most rapid stand-clearing events (including clearcut harvests and fire), as well as many non-stand-clearing events (partial harvest, thinning, storm damage, insect damage)" (Masek et al., 2013, pg. 1089). To create the NAFD data, the VCT algorithm was employed to identify annual forest cover and disturbance between 1986 and 2010 using a time series analysis of an integrated forest z-score (IFZ), an inverse measure of likelihood that a pixel is forested in a given year. The IFZ was informed by normalization indices that were calculated from a training dataset of known forest locations. A consistently low IFZ (close to zero) across the time period indicated relative stability in the forest cover, while a marked increase in IFZ indicated a disturbance in forest cover, ranging from partial to total stand disturbance (i.e. continuous identification of disturbance). To better incorporate disturbances specifically due to fire, NAFD also integrated differenced Normalized Burn Ratio (dNBR) analyses (i.e. the ratio of the difference between the near-infrared and short wave infrared bands of the LTS over the sum of these bands) that compared the burn indices between a pair of pre- and post-fire images. However, like GFC, NAFD also did not contain information on the cause, severity, or uncertainty of the disturbance.

Traditional year-by-year approach to data curation: LANDFIRE

While GFC and NAFD were produced from modern, automated analysis pipelines via HPC, the LANDFIRE disturbance data represents a more traditional, year-by-year approach: "developed through a multistep process employing a number of varied geospatial datasets to identify and label changes in vegetation cover" (LANDFIRE 2016). Specifically, a separate data layer for each year is independently created by combining all known data of reported disturbance in that given year: (1) point locations and perimeters of disturbance events provided by public agencies; (2) vegetation and burn indices calculated from remote sensing analyses of the LTS (e.g. NDVI, dNBR, Burned Area Reflectance Classification, Rapid Assessment of Vegetation Condition after Wildfire); and (3) other data integrated from MTBS, the VCT algorithm, and the Multi-Index Integrated Change Algorithm (MIICA) (Jin et al., 2013). A custom "multi-node cluster" managed by the U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) was used to process satellite imagery, calculate vegetation and burn indices, and integrate the data into one raster layer for each year (USGS 2016).

Methods

Comparison of reported disturbance across California for 2001-2010

For a standardized comparison of the reported disturbance, I created two sets of comparable raster images from the original GFC, NAFD, and LANDFIRE data using EE. The first set contained comparable annual rasters of reported disturbance for each year between 2001 and 2010, while the second set contained comparable aggregated-time rasters of reported disturbance across the study period of 2001-2010. Details on the standardization process are included in the supplemental material (Appendix 3-1). The spatial extent of reported disturbance by each product was mapped using the aggregated-time raster created for each product (Figure 3-1). Disturbed areas attributable to fire were identified by overlaying of the aggregated-time raster created for each disturbance product with a raster of fire occurrence derived from fire perimeters provided by CALFIRE Fire Resource and Assessment Program (FRAP) for the study period (more information on this derived raster of fire occurrence is included in Appendix 3-1).

I also used the derived fire occurrence raster in combination with the standardized annual rasters for each disturbance product to calculate the areas reported as disturbed (and attributable to fire) in each year (m^2/year) and across the study period using the EE function called ee.Image.pixelArea (Figure 3-2). This function provides the pixel areas of the two categories in binary images (e.g. where pixels both reported as disturbed and overlapping with the fire occurrence are labeled with a value of 1, and all others labeled value of 0). For GFC and NAFD, the sums of these annual values of reported disturbance (both attributable and non-attributable to fire) were equivalent the total area reported as disturbed by product across 2001-2010 (Figure 3-1). Due to the annual format of the original LANDFIRE data, pixels could be counted more than once in the sum across the time period (i.e. separate disturbances in different years); thus, two

sums are provided: the unique area reported as disturbed in the time period (Figure 3-1) as well as the duplicated total area reported as disturbed (Appendix 3-2). For the comparison across habitat type, I used the CALFIRE FVEG database to derive four major habitat categories across California: scrub/shrub, forest, grass, and other (e.g. desert, agriculture, wetlands, barren, urban). Definitions of the habitat types derived from FVEG are included in Appendix 3-1. To calculate the amounts of reported disturbance by habitat type, I applied ee.Image.pixelArea to binary images that combined the aggregated-time rasters for each disturbance product and rasters of the four habitat types derived from FVEG. These results provided the total pixel area identified as disturbed by each product across the study period for each habitat type (Figure 3-3; Appendix 3-2).

Environmental conditions

I compared the environmental conditions at pixels reported as disturbed by each of the three disturbance products, based on elevation from the National Elevation Dataset, climate water deficit (CWD) from the California Climate Commons, and mean temperature from the PRISM climate project (see Appendix 3-1 for more details on these environmental datasets). To calculate the distributions of these environmental conditions as covered by each disturbance product, I used the EE functions called ee.Reducer.percentile, ee.Reducer.mean and ee.Reducer.stdDev to produce multiple summary statistics for each disturbance product including minimum and maximum values, the 25th, 50th, and 75th percentiles, means, and standard deviations (Figure 3-4; Appendix 3-2). For a baseline reference comparison, the distributions of these environmental conditions across the total area of California were also calculated. I also compared the proportional areas of each habitat type within the disturbance products by dividing the calculated values for reported disturbance by habitat type by the overall area reported as disturbed by each product (Figure 3-5; Appendix 3-2). For another baseline reference comparison, the proportional areas of each habitat type across the total area of California were also calculated using ee.Image.pixelArea applied to the FVEG data.

Burn conditions

For the final portion of the analysis, I compared the products' spatial coverage across burn conditions as reported by the FRAP fire perimeters and the Monitoring Trends in Burn Severity (MTBS) data. For both datasets, I included the year 2000 to account for pixels that may have been reported as disturbed in the first year of the study (2001). For a baseline comparison of the reference data, I used ee.Image.pixelArea to calculate the reported amounts of fire disturbance across California and by habitat type using the fire occurrence raster derived from FRAP fire perimeters and the MTBS burn severity data (Table 3-2; Appendix 3-2). As the MTBS data were originally provided as annual rasters, I aggregated them to create a new single raster that contained the maximum burn severity at each pixel across the study period.

Since the creation method of LANDFIRE already incorporated versions of the FRAP and MTBS data (see section on Study Data), the primary intention of this analysis was to identify how comparable the burn conditions covered by GFC and NAFD were to those covered by LANDFIRE. To this end, the coverage of FRAP fire perimeters by size class and MTBS by burn severity level were calculated for each of the disturbance products (Figures 3-6 and 3-7). For the FRAP fire perimeters, I categorized the individual fire perimeters into six fire perimeter size classes based on acreage reported by FRAP (i.e. less than 100, 100-500, 500-1,000, 1,000-10,000, 10,000-90,000, greater than 90,000). Coverage of the FRAP fire perimeters was calculated using ee.Reducer.frequencyHistogram, which provided the pixel count for each disturbance product contained within each fire perimeter. These pixel counts were converted to percentages by dividing the number of pixels reported as disturbed by each disturbance product by the total number of pixels contained within the fire perimeter. For each size class of fire perimeters, a mean of these percentages was calculated to provide the average percentage of coverage in that size class for each disturbance product (Figure 3-6; Appendix 3-2). Last, I used ee.Image.pixelArea to calculate coverage by each of the disturbance products for each MTBS burn severity level (unburned to low, low, medium, high) by habitat type (Figure 3-7).

Results

LANDFIRE reported the highest amounts of disturbance across California for all years and habitat types

Across California between 2001 and 2010, GFC and NAFD reported similarly lower amounts of disturbance as compared to LANDFIRE (Figure 3-1). LANDFIRE reported the highest amount of disturbance at 8.41% of the total area of California (with 5.54% of the reported disturbance overlapping with a FRAP fire perimeter), while GFC reported the least amount of disturbance at 2.54% of the total area of California (with 1.71% of the reported disturbance overlapping with a FRAP fire perimeter). Overall, the products reported more similar amounts for disturbance that did not overlap with FRAP fire perimeters, ranging from 0.83% to 2.87% of the total area of California attributed to non-fire disturbance.

Comparing reported disturbance across years, LANDFIRE reported the highest amounts of disturbance in each year, while GFC and NAFD reported more similar amounts of disturbance, though NAFD generally reported slightly more disturbance in each year with the exception of 2006 (Figure 3-2). The greatest differences in reported disturbance between LANDFIRE and GFC/NAFD occurred in 2008, 2003, 2007, and 2006 (in descending order), all years for which FRAP reported the highest annual numbers of individual fire perimeters (between 309 and 425). In these years, the reported disturbances by LANDFIRE also demonstrated higher overlap with the FRAP fire perimeters, as compared to other years; similarly, the reported disturbances by GFC and NAFD demonstrated the highest overlaps with the FRAP fire perimeters in 2008, followed by 2007. The smallest difference between LANDFIRE and GFC/NAFD occurred in

2010 and 2001, years in which FRAP reported the lowest annual numbers of individual fire perimeters (204 and 200, respectively).

Across all habitat types, LANDFIRE also reported the highest amounts of disturbance as compared to NAFD, and even more so as compared to GFC (Figure 3-3). The products differed in their amounts of reported disturbance the most for scrub/shrub (ranging from approximately 6.5% to 20% of the total scrub/shrub area of California) and forest (ranging from approximately 4% to 14% of the total forest area of California). Overall, all of the products reported higher amounts of disturbances overlapping with FRAP fire perimeters in scrub/shrub (ranging from 5.7% to 17% of the total scrub/shrub area of California attributed to fire disturbance), as compared to forest (ranging from approximately 2.5 to 7% of the total forest area of California attributed to fire disturbance).

Disturbance 2001-2010	GFC*	NAFD*	LF*
Non-Fire	0.83%	1.64%	2.87%
Fire**	1.71%	2.13%	5.54%
Total	2.54%	3.77%	8.41%

* Percent area of California
** Based on overlap with CALFIRE Fire Resource Assessment Program (FRAP) Fire Perimeters

Figure 3-1. Total reported disturbance across California between 2001 and 2010. Disturbance attributed to fire is based on overlap with fire perimeters from CALFIRE Fire Resource and Assessment Program (FRAP).

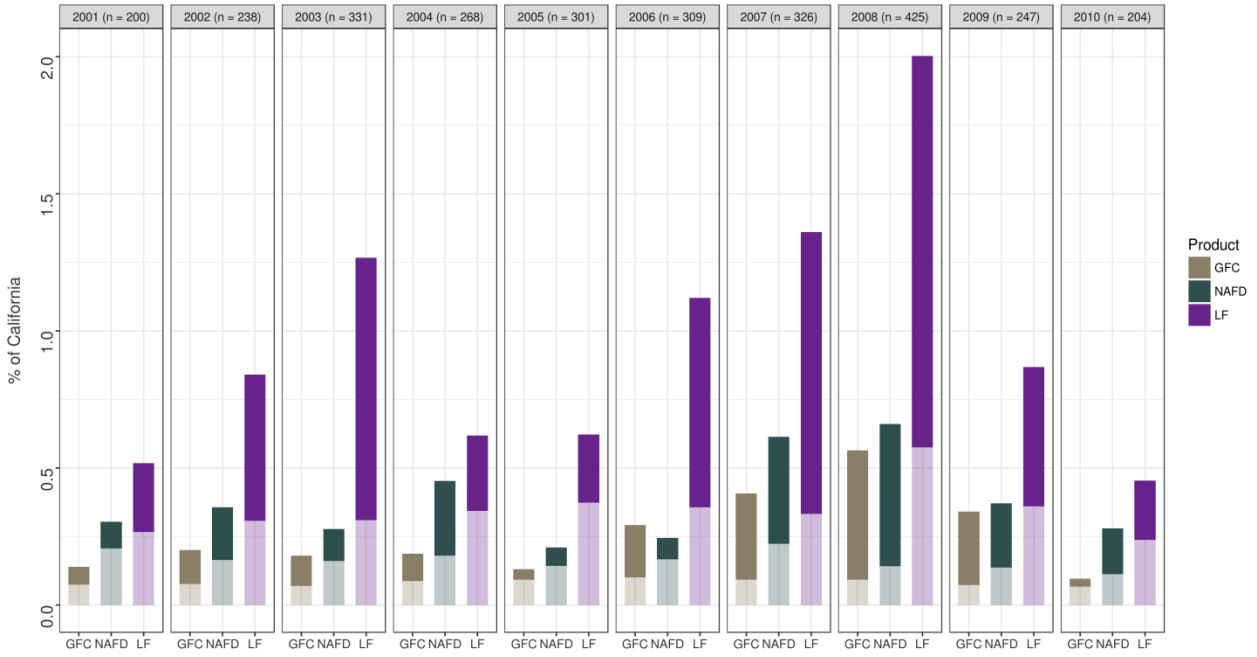


Figure 3-2. Annual reported disturbance across California for years 2001 to 2010. Darkest portion of each bar represents proportion of reported disturbed area attributed to fire, based on overlap with FRAP occurrence, for Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE (LF).

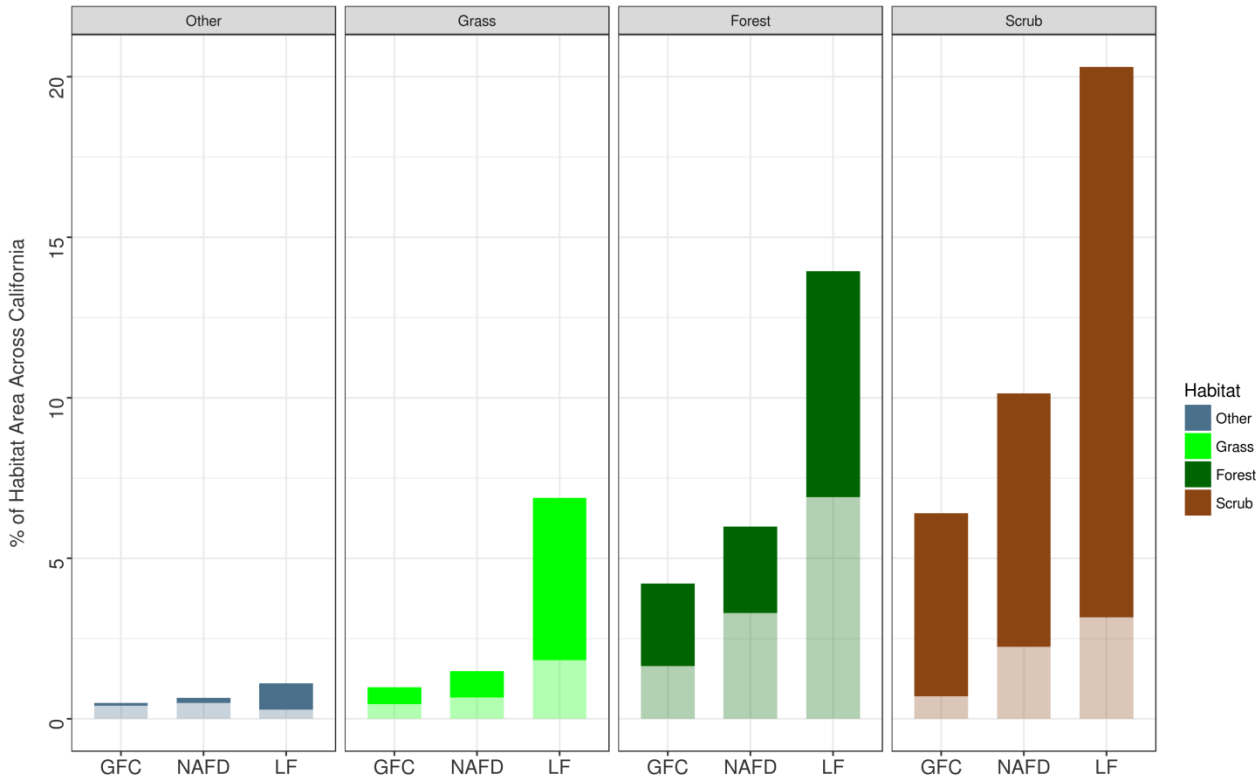


Figure 3-3. Reported disturbance by habitat type. Darkest portion of each bar represents proportion of reported disturbed area attributed to fire, based on overlap with FRAP occurrence, for Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE (LF).

Vegetation disturbance products covered similar environmental conditions

In my analysis of bioclimatic conditions covered by the disturbance products, I found that the products covered very similar distributions across elevation, climate water deficit (CWD), and mean temperature (Figure 3-4). Specifically, the products reported disturbance at higher elevations, lower CWD, and lower mean temperatures, as compared to the reference baselines for the total area of California. The products also covered similar proportions of habitat types that differed from the reference baselines for California (Figure 3-5). In particular, as proportions of their total areas, the products covered more forest (ranging from approximately 47.5% to 50% of their total areas) and notably more scrub/shrub (ranging from approximately 30 to 40% of their total areas), as compared to the proportions of those habitat types across California (approximately 30% for forest and 15% for scrub/shrub). The products demonstrated some variability in their coverage of grass and the other category, with GFC and NAFD reporting more similar values that were lower than that of LANDFIRE. Even with this variability, the products still reported less coverage in grass and notably less coverage in the other category than the reference baseline for California.

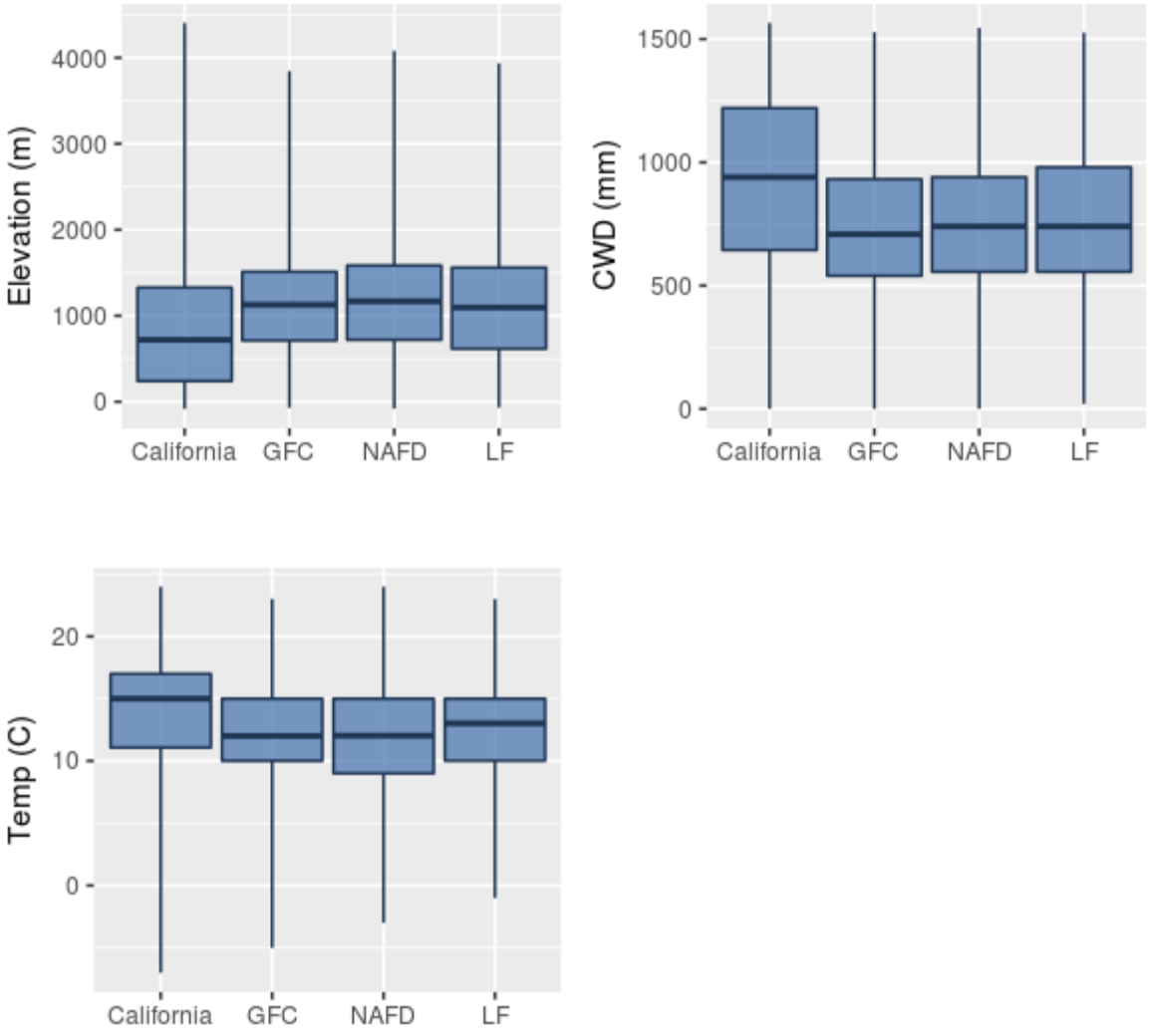


Figure 3-4. Distributions of bioclimatic conditions covered by the products. For reference comparison, the distributions of conditions across total area of California are also reported, alongside distributions for Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE (LF).

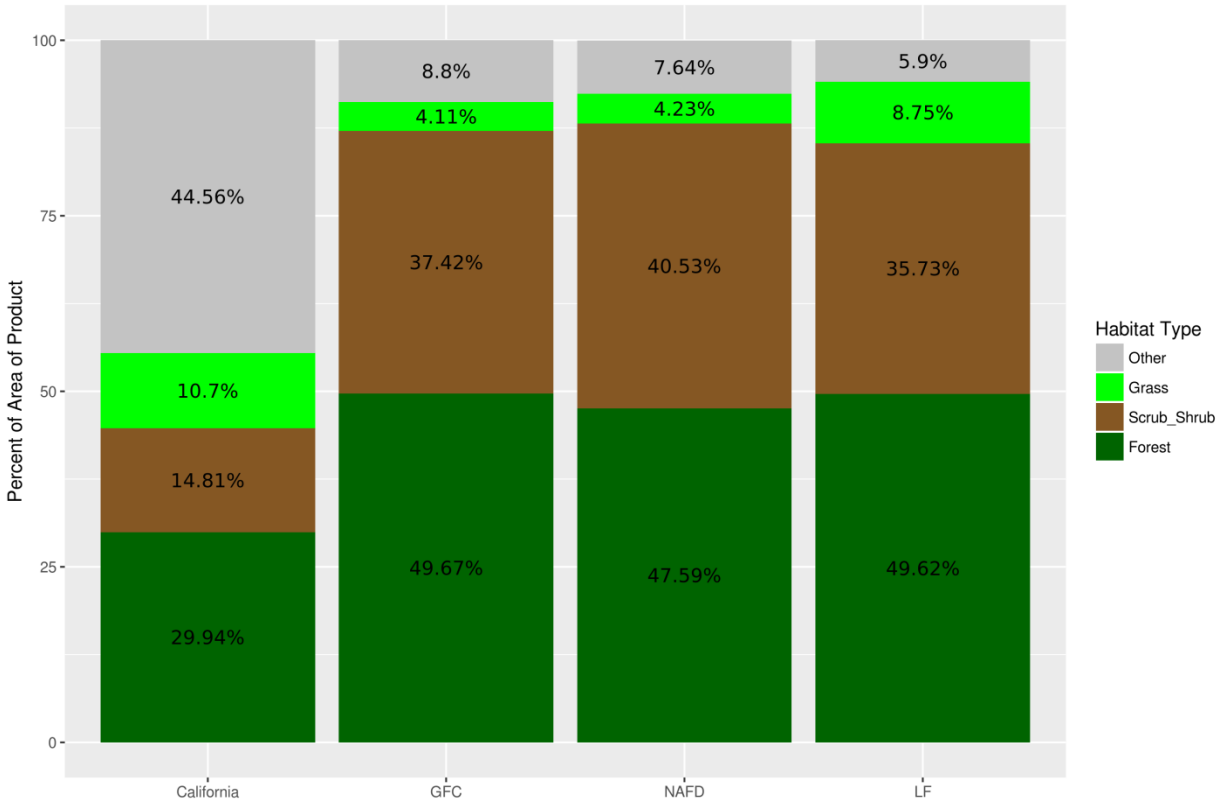


Figure 3-5. Habitat areas as proportions of total area of products. For reference comparison, habitat areas as proportions of total area of California are also reported, alongside proportions for Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE (LF).

Burn conditions covered by GFC/NAFD differed greatly from those of LANDFIRE

FRAP and MTBS reported similar amounts of fire disturbance in the study period, with both reporting approximately 5.8% of California as burned and reporting the most fire disturbance in scrub/shrub habitats (approximately 18% of all scrub/shrub habitat across California) (Table 3-2; Appendix 3-2). As LANDFIRE already incorporated versions of the FRAP fire perimeters and MTBS burn severity data, its coverage of the burn conditions reported by these reference datasets was close to 100% across all fire perimeter sizes and burn severity levels (Figures 3-6 and 3-7). In contrast, the automated disturbance products of GFC and NAFD demonstrated similar coverage of both the FRAP fire perimeters and MTBS burn severity data that were notably less than that of LANDFIRE. For fire perimeter sizes larger than 1,000 acres, GFC and NAFD reporting of disturbance within the fire perimeters increased with size, but reached a maximum coverage of approximately 40% at the largest fire perimeter size (Figure 3-6). As burn severity increased, both GFC and NAFD reported more disturbance overlapping with the MTBS

data across all habitat types, with the most disturbance reported for forest, which reached a maximum coverage of approximately 80% at the highest burn severity (Figure 3-7).

Table 3-2. FRAP and MTBS fire disturbance for 2000-2010

	All California	Scrub/Shrub	Forest	Grass	Other
FRAP Percent of Area Burned	5.83%	17.87%	7.31%	5.6%	0.89%
MTBS Percent of Area Burned	5.77%	18.10%	7.29%	4.99%	0.84%

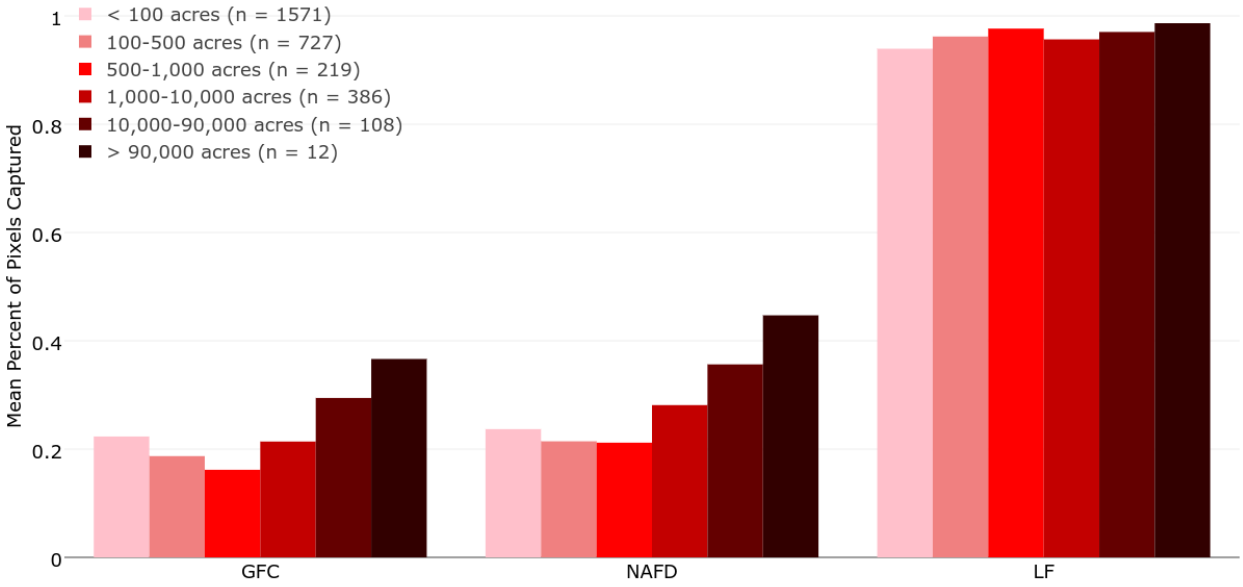


Figure 3-6. Comparison of covered burn conditions across fire perimeter size. Based on overlap with fire perimeters from CALFIRE Fire Resource and Assessment Program (FRAP) for Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE (LF).

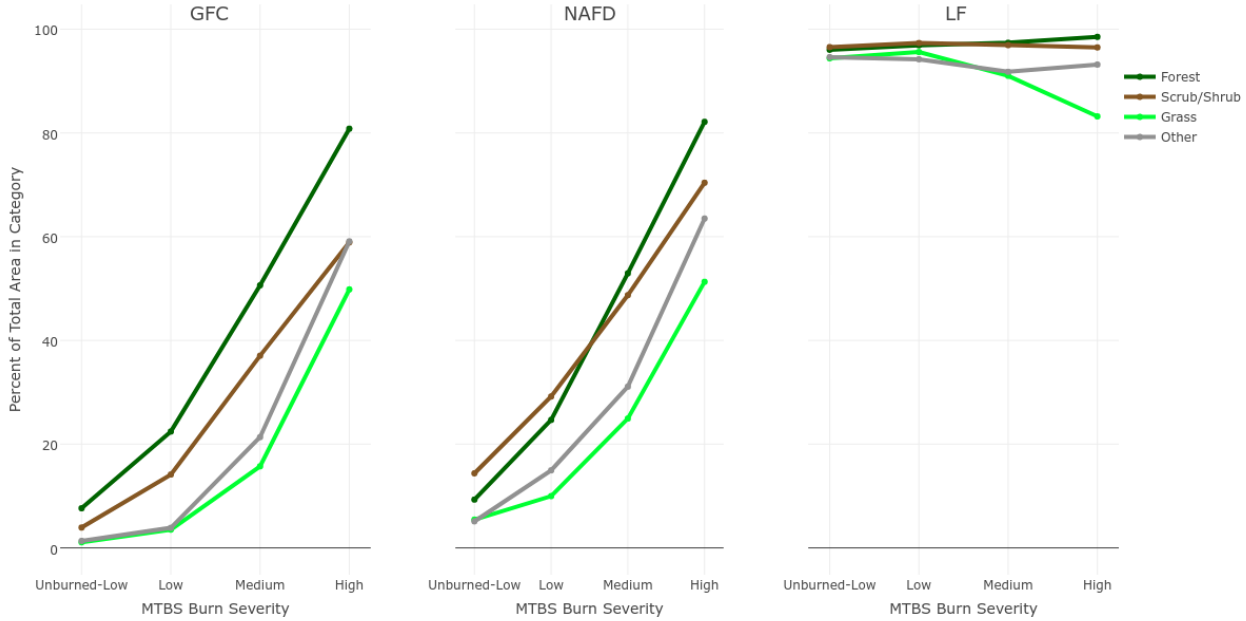


Figure 3-7. Comparison of covered burn conditions across burn severity and habitat type. Based on overlap with data from Monitoring Trends in Burn Severity (MTBS) for Hansen Global Forest Change (GFC), North American Forest Dynamics (NAFD), and LANDFIRE (LF).

Discussion

Differing methods of creation greatly impacted reported disturbance

Overall, the automated disturbance products of GFC and NAFD both reported significantly less disturbance across all years and habitat types than did the traditionally created product, LANDFIRE. While the comparison of environmental conditions (i.e. distributions of elevation and climate; proportional areas of habitat types) did not reveal major differences in spatial coverage by the products, my analysis of burn conditions identified notably less overlap with the reference data for GFC and NAFD, as compared to LANDFIRE. These results indicate that differences in reported disturbance by the products were not driven by differential coverage of bioclimatic or habitat conditions, but rather by the threshold for disturbance that is inherent to method of creation for each product. Specifically, as GFC and NAFD were both created from automated HPC-based analyses of only satellite imagery, rather than by a more manual data aggregation method like that of LANDFIRE, their thresholds for reporting disturbance are higher than that of LANDFIRE which incorporates disturbance events that have been reported and spatially delineated by other public agencies. The higher reporting of disturbance by NAFD as compared to GFC also reflects a key difference in their automated processes, namely that the threshold for discrete disturbance in GFC is higher than that of continuous disturbance in NAFD. Thereby, reductions in vegetation cover that do not result in a discrete change in vegetation cover would be captured by NAFD but not GFC.

These results clearly reflect the strengths and weaknesses of the different approaches to identifying disturbance and creating standardized annual products. Specifically, while the key strength of the distributed computing approach of GFC is that the workflow is both simplistic and replicatable (based on standard remote sensing techniques such as supervised classification and time series analysis of vegetation indices), it is clear that a major limitation of this approach is the identification of only discrete losses of vegetation. Reductions in vegetation cover are not identified until the reduction is significant enough to cause a change in cover type or notable drop in greenness, possibly limiting its applicability to certain habitat types. Furthermore, while the workflow for identifying loss is simple and approachable for non-remote sensing experts, it is not easily reproducible (i.e. based on a portable algorithm that can easily be copied to and executed on another platform). In contrast, the algorithmic approach of NAFD is based on an integrated-time comparison (through the VCT algorithm) that allows for the identification of partial disturbances and reductions in vegetation cover, and is completely reproducible and easily integrated with other analyses due to the portability of the VCT algorithm across platforms. However, a major limitation of this approach is that calculations of metrics like the IFZ score used by VCT require expert knowledge to interpret how the changes in values correspond to disturbance (i.e. identifying a partial disturbance from noise), making it less approachable for non-remote sensing experts to fully interpret or recreate products like NAFD. In general, as the approaches used by GFC and NAFD are automated remote sensing analyses focused on forest, there exists a notable potential for underreporting of disturbance, as smaller disturbances or disturbances in other vegetation types may be missed by these approaches that rely solely on spectral changes in satellite imagery.

In contrast to GFC and NAFD, the traditional approach of LANDFIRE does not target specific vegetation types and can easily incorporate both discrete and continuous disturbances because it is a manually curated product that aggregates multiple data types, including both HPC-based remote sensing analyses and publicly collected data on disturbance events. The variety of data used in this traditional curation approach also supports the labeling of a disturbance type, severity, and assigned uncertainty based on the data source. However, the major limitations of this approach are that it is neither easily replicatable nor reproducible (because it is not an automated process and aggregates data on a year-to-year basis) and there is a high potential for compounding data inaccuracies present in the datasets that are aggregated into LANDFIRE. Between years, this curation method can vary in data quality and accuracy, depending on the data that was received for that year by other public agencies (ranging from local to federal levels). For example, in the analysis of burn conditions, LANDFIRE did not demonstrate complete agreement with either the FRAP or MTBS, even though these reference data are stated to be integrated as part of the LANDFIRE creation process. Furthermore, there is a notable potential for overestimation of disturbance in products like LANDFIRE because disturbance event locations and perimeters reported by public agencies are not always ground-checked and can be hand-demarcated to include a larger area than the actual footprint of the disturbance. As both FRAP and MTBS follow a similar manual curation process as LANDFIRE through year-

by-year aggregation of data and multiple analyses combined into one product, it is not surprising that overlap with the reference data did not converge more strongly between LANDFIRE and GFC/NAFD, and higher overlap with the reference data should not be interpreted to mean that LANDFIRE is a better representation of burned areas than GFC and NAFD.

Implications for use of these products as representations of disturbance and fire

The differences in the creation methods of these disturbance products and the conditions under which they report disturbance have notable implications for their use as representations of disturbance in ecological studies, particularly for fire. For example, my results suggest that while a higher number of individual fires in a given year resulted in more area reported as disturbed by all products, it also resulted in a wider range in the total areas reported as disturbed (Figure 3-2). The largest difference among the products occurred in 2008, which is the year with the highest number of FRAP fire perimeters as well as the highest values for the areas reported as disturbed (and attributed to fire disturbance). The smallest difference among the products occurred in 2010, which corresponds with the second lowest number of fire perimeters as well as the lowest values for areas reported as disturbed for GFC and LANDFIRE. As smaller fires tend to occur more often than larger fires, it is likely that in the years with more recorded fire, smaller fires were driving the increased divergence in reported disturbance. As such, the difference between LANDFIRE and the other two products should be corroborated by higher spatial overlap of smaller fires by LANDFIRE than GFC and NAFD. Indeed, my results demonstrated much higher percent overlaps of the smallest fire size classes by LANDFIRE (Figure 3-7). These results highlight the impact of the differing thresholds for reporting disturbance, which will likely result in a wider gap (i.e. more uncertainty) among the products as the landscape experiences more disturbance or as reference data like FRAP and MTBS report more disturbance. Overall, higher coverage of smaller fires by LANDFIRE is primarily due to its inclusion of reference data on fire, while smaller fires likely do not impact the landscape significantly enough to be identified as disturbance by the spectral-based analyses behind the GFC and NAFD products.

Interestingly, the differential focus on forest by GFC and NAFD did not result in higher reported disturbance for forest as compared to scrub/shrub. All three disturbance products reported the most disturbed area in scrub/shrub (mostly attributed to fire), and the range of reported disturbance among the products was also widest for scrub/shrub (Figure 3-3). For forest, GFC and NAFD reported more similar totals for disturbed and burned areas that were lower than LANDFIRE, which reported more disturbance overall than GFC and NAFD but also reported less disturbed and burned area in forest than scrub/shrub. The greater difference in reported disturbance for scrub/shrub is somewhat surprising given that the products contained similar proportions of the habitat types and that scrub/shrub covers only 15% of the total area of California, as compared to approximately 30% for forest (Figure 3-5). My results show that scrub/shrub is most frequently reported as disturbed by both reference data and the disturbance products (Table 3-2; Figure 3-3), again suggesting that greater reporting of fire incidence by

reference data results in a wider range (i.e. more uncertainty) of reported disturbance among the disturbance products, in this case for scrub/shrub habitat.

Integral Role of Spatial Data Science and HPC

Previous to this study, comparative evaluations of disturbance products had likely been limited by two related factors: (1) the lack of overlap in products' coverage across space and time (i.e. limited and non-overlapping spatial and temporal extents); and (2) the inadequacy of traditional analytical tools to handle and analyze data at increasingly finer resolutions and broader extents. In other words, remotely-sensed products of vegetation disturbance with annual coverage at the spatial resolution of the LTS are recent developments that have been made possible by the creation of computational tools to create and compare them. In particular, the recent development of HPC tools specifically intended for geospatial analyses (such as EE) has helped to address computational challenges presented by large spatial-temporal datasets and has been encouraged by CyberGIS researchers interested in expanding geospatial tool interoperability and scalability for 'big' spatial data (Wang, 2016; Yang et al., 2011; Yang, Raskin, Goodchild, & Gahegan, 2010). Alongside these technological advances, integration of traditional geospatial methods and modern Data Science techniques (i.e. data mining/algorithms, machine learning) have arisen from the development of a Spatial Data Science to support the application of fundamental geospatial analyses to 'big' spatial-temporal data stacks such as the LTS (Palomino et al., 2017).

These developments in Spatial Data Science and HPC have resulted in both the creation of products at finer spatial and temporal resolutions and broader extents as well as an increased ability to compare and evaluate them. Even while limited to the California scale, the disturbance products evaluated in this study are "big" data, as approximately 450 million pixels were analyzed for each year to cover California at a 30 m spatial resolution. Just as the processing power and data handling capabilities of HPC were needed to create these disturbance products, a thorough interrogation and evaluation of these products also required the use of HPC to identify patterns across this complex multi-temporal data stack. The platform employed in this study, EE, is an exemplary tool emerging from these advances in HPC and Spatial Data Science, as it supports fundamental geospatial analyses such as raster stack calculations and zonal statistics on data that are not easily handled in traditional desktop tools.

While it is clear that the differences in the creation methods of these products have significant implications for their use in representing disturbance, the evolution from more curated (i.e. LANDFIRE) to more automated (i.e. GFC and NAFD) workflows is indicative of the overall trajectory of the Spatial Data Science field. As products of self-contained analysis pipelines or algorithms running autonomously on HPC, GFC and NAFD are key examples of data that will likely continue to be produced with high temporal frequency. In contrast to the lengthy protocols and manual data integration of a product like LANDFIRE, these modern Spatial Data Science products stream-line the identification of disturbance by focusing exclusively on changes in spectral characteristics using a data science approach that does not require a priori knowledge of

disturbance events. Despite lower overlap with the reference data as compared to LANDFIRE, GFC and NAFD both provide unique spatial coverage of vegetation disturbance that are not currently available in other products. At the time of this publication, GFC is the only dataset that maps annual vegetation change at a global extent with the spatial resolution of the LTS. While focusing on the more narrow extent of North America, NAFD uniquely provides identification of continuous disturbance (i.e. reduction of cover, not just discrete changes in vegetation) over a large continental-scale. Moving forward, these modern workflows and the products they create can also enable regional (and possibly global) analyses of fire return intervals and dynamics of burn intensity to identify generalizable trends (Stevens, Collins, Miller, North, & Stephens, 2017), beyond local analyses of individual fires(Collins et al., 2009; Collins, Kelly, van Wagtendonk, & Stephens, 2007).

Future Directions

This comparative study of these disturbance products provides an understanding of the implications of competing approaches to creating disturbance products and identifies how different approaches can possibly result in under- and over-estimation of reported disturbance. Expanding on this study to focus on spatially explicit agreement across the disturbance products, rather than on their differences, could provide the ability to more accurately identify burned areas within fire perimeters and other reference data as well as provide a spatially explicit measure of uncertainty for both fire and non-fire disturbances reported by the products. While there has been some research on integrating some of these products (Schroeder et al., 2017; Soulard et al., 2017) or the algorithms used to create them (Healey et al., 2017) to improve the accuracy of disturbance identification (i.e. potentially reducing uncertainty), the integrated products have not been used to quantify uncertainty across study areas. Through data integration based on spatial agreement, uncertainty could be identified in spatially explicit approach. For example, pixels that are reported as disturbed by all products (i.e. pixels of highest agreement) and are contained within a fire perimeter could be assigned a low uncertainty for a fire disturbance. Similarly, pixels that are reported as disturbed by all products (again, pixels of highest agreement) but are not contained within a fire perimeter could be assigned a low uncertainty for a non-fire disturbance. On the other hand, pixels that are not reported as disturbed in any product but are contained within a fire perimeter in reference data could be assigned a high uncertainty for fire disturbance (i.e. likely unburned area within the fire perimeter), helping to narrow down the true extent of fire events. These kind of spatial agreement metrics could provide a more automated and more objective identification of unburned areas (as compared to dNBR analyses which require defining thresholds for disturbance that vary by vegetation and ecosystem type) as well as highlight areas where individual disturbance products may be overzealous in reporting disturbance (i.e. errors of commission). These measures of uncertainty in disturbance could also help narrow the range of the amount of vegetation disturbance toward the “true” amount of disturbance in a given landscape.

Conclusion

I used Earth Engine to compare the reported amounts of disturbance for 2001-2010 among three widely used vegetation disturbance products and examined the products across differing environmental and burn conditions. Overall, GFC reported the least amount of disturbed area (2.54% of California) as well as least amount of disturbed area attributable to fire (1.71% of California) for 2001-2010, as compared to NAFD (3.77% and 2.13%) and LANDFIRE (8.41 and 5.55%). My analysis did not reveal major differences in the coverage of environmental conditions (i.e. elevation, climate, habitat) by the products, but I did identify notable differences in the coverage of burn conditions between GFC/NAFD (products created by automated methods) and LANDFIRE (the traditionally created product that manually incorporates some versions of the reference data on fire). These results indicate that differences in reported disturbance were driven by the differing methods of creation among the products, rather than by differential coverage of environmental conditions. Furthermore, I also identified notable differences in reported disturbance by year and by vegetation type. In particular, a higher incidence of fire as reported by the reference data in a given year resulted in more reported disturbance by all products, but also contributed to a wider range in the amount of reported disturbance (i.e. uncertainty). Similarly, both the reference data on fire and the disturbance products reported the most disturbance in scrub/shrub habitats over forest, grass, and other habitat types, again indicating that higher disturbance reported by reference data resulted in a greater difference in reported disturbance among the products (i.e. uncertainty).

Designed and executed within EE, my methodology provides a reproducible framework for comparative analyses of vegetation disturbance products to identify the conditions under which they report disturbance as well as drivers of uncertainty in reported disturbance. This chapter used fire in California as a case study to help end-users of these products understand how the identification methods of disturbance (i.e. the creation method of disturbance products) impact the amounts and locations of reported disturbance. Rather than focusing on accuracy, I interpret the differences in reported disturbance as the result of differing thresholds for reporting disturbance that are inherent to each product due its creation method. Thus, these disturbance products exist along a continuum of low to high thresholds for reporting disturbance (i.e. from LANDFIRE to GFC), with the "true" amount of disturbance existing somewhere on the continuum. Future work can explore spatial agreement among of these products (i.e. where they agree) as a data integration method for quantifying bounded estimates of vegetation disturbance, more accurate identification of burned areas and assigning spatially explicit metrics of uncertainty in disturbance across study areas.

Chapter Four

Low spatial agreement among remotely-sensed map products highlights high uncertainty of vegetation disturbance across California for 2001-2010

The Landsat Time Series (LTS) is increasingly being used to create standardized, remotely-sensed products that map annual occurrences of vegetation disturbance across the United States. These vegetation disturbance products are often used in research that aims to quantify the impact of disturbance (e.g. fire) on ecosystem processes, such as aboveground carbon dynamics. However, most carbon studies to date use only one product to represent disturbance in their study areas, without accounting for uncertainty in the amount and/or location of disturbance in that product. In this Chapter, I introduce a methodology for quantifying uncertainty in reported disturbance using an analysis of spatial agreement among competing vegetation disturbance products (i.e. data integration). My methodology uses basic raster calculations to identify which products overlap at each pixel, and then determines the amount of spatial agreement among products to identify uncertainty of disturbance (e.g. high uncertainty of disturbance due to only one product reporting disturbance versus low uncertainty where all products report disturbance). Using Google's Earth Engine, I present a case study of my methodology using the three vegetation disturbance products that have overlapping coverage across California between 2001 and 2010: Global Hansen Forest Change (GFC); North American Forest Dynamics (NAFD); and LANDFIRE. I examined the patterns of uncertainty of disturbance across habitat types, bioregions, and burn conditions as reported by reference data on fire. My results indicated low spatial agreement among the products, resulting in high uncertainty of disturbance across California. The low uncertainty of disturbance category covered only 15% of the total area in California that was reported as disturbed by at least one product. Most of the area with low uncertainty of disturbance was attributed to fire events, rather than non-fire events. Scrub/shrub was both most frequently reported as disturbed as well as identified to have the lowest uncertainty of disturbance, as compared to forest or grass. Across the California bioregions, uncertainty was universally high; on average, the areas with high uncertainty accounted for 64% or more of the total area reported as disturbed across all regions. Furthermore, my results highlighted large differences between the burned areas reported by the reference data and the areas with low uncertainty of disturbance derived from spatial agreement among the disturbance products. I also identified that LANDFIRE was solely responsible for approximately 50% of the total area reported as disturbed (the majority attributed to fire events), while GFC and NAFD individually accounted for less than 10% of the total area reported as disturbed (both attributing the majority to non-fire events). These results point to potential overestimation of disturbance by both the reference data on fire and LANDFIRE as well as potential underestimation by GFC and NAFD, particularly for fire events. Overall, my investigation of uncertainty of disturbance within California helps researchers and practitioners to identify the strengths and weaknesses of the individual and collective detection capabilities of these products and provides a clear case study

of a simple but powerful methodology for using a data integration approach to quantify uncertainty of disturbance across broad spatial and temporal extents.

Introduction

Remote sensing data have become critical components of research on interactions between vegetation and disturbance dynamics, such as fire, across the United States (U.S.) (Arroyo, Pascual, & Manzanera, 2008; Lentile et al., 2006). In particular, the Landsat Time Series (LTS), which provides satellite images at moderate spatial (i.e. 30 m) and temporal (i.e. approximately 2 weeks) resolutions, has been widely used by fire ecologists to identify site-specific connections between vegetation and fire severity through comparison of pre- and post-fire satellite images (Collins, Kelly, van Wagtendonk, & Stephens, 2007; Collins & Stephens, 2010; Miller, Safford, Crimmins, & Thode, 2009) and by remote sensing experts to develop LTS-specific algorithms aimed at identifying spatial-temporal dynamics of vegetation disturbance (i.e. quantifying changes in vegetation cover due to fire, drought, pestilence, harvesting, or land cover/use change) (Huang et al., 2010; Kennedy, Yang, & Cohen, 2010; Zhu, 2017; Zhu & Woodcock, 2014). Open access to the LTS coupled with its broad utility in mapping vegetation has led to the creation of multiple remotely-sensed products that map annual vegetation disturbance using different methods, including Hansen Global Forest Change (GFC) (Hansen et al., 2013), North American Forest Dynamics (NAFD) (Goward et al., 2015), and Landscape Fire and Resource Management Planning Tools (LANDFIRE) (Keane, Rollins, & Zhu, 2007; Reeves, Ryan, Rollins, & Thompson, 2009; Rollins, 2009; Ryan & Opperman, 2013). My previous work evaluating the differences among these products (Chapter Three) concluded that differing methods of creation (i.e. automated remote sensing analyses by GFC and NAFD versus a traditional approach by LANDFIRE that integrates field data and public data on disturbance events with remote sensing) resulted in different thresholds for reporting disturbance, from GFC on the low end of reported disturbance to LANDFIRE at the high end. Thus, rather than relying on one product to accurately represent disturbance, researchers and practitioners should view these products as differing representations of disturbance along a continuum of low to high thresholds for reporting disturbance, and instead, seek ways to represent uncertainty of reported disturbance in their work.

While these vegetation disturbance products are increasingly being used by researchers (including many non-remote sensing experts) to estimate changes in vegetation biomass and quantify impacts on regional-scale aboveground carbon dynamics across the U.S., most studies to date use only one product to represent the amount and location of disturbance in their study areas. For example, GFC has been used to identify changes in forest cover and carbon stocks within Forest Inventory Analysis (FIA) plots in the Eastern U.S. (Woodall et al., 2016) and to explore the impact of disturbance on tree mortality and related changes in water and carbon fluxes in the Western U.S. (Anderegg et al., 2016). NAFD has been used to identify the impact of time since disturbance on biomass and carbon in in the Pacific Northwest (Gu, Williams, Ghimire, Zhao, & Huang, 2016), and more broadly, to explore the potential impact of

disturbance history on U.S. forests' ability to remain a net carbon sink (Williams, Gu, MacLean, Masek, & Collatz, 2016). In California, LANDFIRE has been used to project future wildfire emissions under different climate, population and development scenarios (Hurteau, Westerling, Wiedinmyer, & Bryant, 2014), and in combination with other data such as FIA or Monitoring Trends in Burn Severity (MTBS) to identify the influence of disturbance on carbon dynamics across short (2001-2010) (Gonzalez, Battles, Collins, Robards, & Saah, 2015) and longer time periods (1951-2000) (Liu et al., 2011). While some of these studies accounted for uncertainty of certain variables, such as habitat type or canopy characteristics, none accounted for uncertainty in the amount and/or location of disturbance in the product that they used to represent disturbance.

The lack of accounting for uncertainty in the amount and location of disturbance presents challenges for applying the results of carbon studies based on only one disturbance product. For example, using LANDFIRE data, a previous study found that the majority (i.e. two-thirds) of the aboveground carbon loss across California between 2001 and 2010 was specifically due to wildfires in the predominantly forested Sierra Nevada and Klamath-Siskiyou mountains, while the remaining one-third was identified to have occurred in shrub/shrub habitats, primarily attributed to fire events in Central and Southern California chaparral (Gonzalez et al., 2015). However, my previous work (Chapter Three) identified notable differences in the amounts of reported disturbance among various products (i.e. GFC, NAFD, LANDFIRE) in both scrub/shrub and forest habitat areas and in years with high fire incidence, highlighting the uncertainty introduced by the choice in disturbance product used to represent disturbance in a given study.

Recognizing the role of uncertainty in mapping and quantifying disturbance, other researchers have started to explore methodologies for addressing uncertainty in the amount and location of disturbance through integrating the algorithms used to identify disturbance (i.e. ensemble modeling) (Healey et al., 2017), the development of multi-step statistical models that incorporate many of these products as predictors of disturbance (Schroeder et al., 2017), and map integration approaches (Soulard et al., 2017) that improve the accuracy in the identification of disturbance. While these emerging approaches have contributed to a greater ability for remote sensing experts to identify and quantify disturbance (and will possibly led to more robust disturbance products in the future), they present new limitations for accounting for uncertainty: (1) not all disturbance products are the results of reproducible algorithms (e.g. the widely used LANDFIRE; see Chapter Three, for a detailed description); (2) the complexity of many of these approaches make them unrealistic for non-remote sensing experts, who need a simple way to incorporate measures of uncertainty of disturbance into their work; or (3) the approaches improve accuracy but do not provide a measure of uncertainty of reported disturbance that can be used for further research on the implications of disturbance on ecosystems.

In this Chapter, I present a simplified methodology specifically for quantifying uncertainty of disturbance based on spatial agreement among competing vegetation disturbance products (i.e. a

data integration approach). As quantifying uncertainty is directly related to how it is defined, this Chapter is guided by a definition that highlights the contribution of competing representations to uncertainty: "The degree to which the measured value of some quantity is estimated to vary from the true value. Uncertainty can arise from a variety of sources, including limitations on the precision or accuracy of a measuring instrument or system; measurement error; the integration of data that uses different scales or that describe phenomena differently; *conflicting representations of the same phenomena*; the variable, unquantifiable, or indefinite nature of the phenomena being measured; or the limits of human knowledge" (ESRI 2018). My methodology begins with basic raster calculations to quantify spatial agreement at each pixel by overlaying the disturbance products and labeling each pixel based on the products that report disturbance at that pixel (e.g. only LANDFIRE reported disturbance at the given pixel). These spatial agreement metrics are then converted to measures of uncertainty based on the number of products that report disturbance (e.g. high uncertainty of disturbance due to only one product reporting disturbance at a given pixel versus low uncertainty at a given pixel where all products report disturbance), which can also be overlaid with reference data on disturbance events to label disturbance type. I use a case study of spatial agreement among the three vegetation disturbance products that have overlapping coverage across California (i.e. GFC, NAFD, and LANDFIRE) to quantify uncertainty of disturbance between 2001 and 2010. Motivated by previously discussed research highlighting the strong influence of habitat, regional area, and fire on carbon dynamics across California, this study further examines uncertainty of disturbance across habitat types (including scrub/shrub and forest), bioregions (including Central and Southern California and the Sierra Nevada and Klamath-Siskiyou mountains), and burn conditions as reported by state and federal reference data on fire.

Specifically, I address the following questions:

1. Where and how much did the three vegetation disturbance products agree across California between 2001 and 2010?
2. Using spatial agreement among the products as measures of uncertainty of disturbance at a given location, how did uncertainty vary by habitat type and bioregion across California?
3. How did uncertainty of disturbance vary across burn conditions reported by reference data on fire (i.e. fire perimeter size, burn severity)?

While I use fire in California as a case study to examine the implications of uncertainty in reported disturbance, my methodology of mapping and quantifying uncertainty through data integration (i.e. spatial agreement among disturbance products) can easily be expanded to new study sites with other data sources and be reproduced for new disturbance products derived from the LTS and other satellite sensors as they become available. I believe that my simple methodology can encourage and empower non-remote sensing experts to easily account for

uncertainty in the amount and location of disturbance in analyses that aim to understand the potential impacts of disturbances on ecosystems.

Methods

Spatial agreement approach to identifying measures of uncertainty of disturbance

The methodology presented in this Chapter for quantifying uncertainty of disturbance is based on a series of basic raster calculations that can be completed in any Geographic Information System (GIS) or programming language (e.g. R, Python) that supports operations on georeferenced image files such as pre-classified raster layers (Figure 4-1). As all disturbance products have a different labeling system for disturbance, any analysis should begin with a standardization of the raster layers to define which values constitute disturbance. This standardization can easily be accomplished by relabeling all pixels that are reported as disturbed within each raster layer, using a new value for each raster (such as 2, 3, 4 if using three raster products), and labeling all non-disturbed pixels in all raster layers with a different value (such as 0). Following Figure 4-1, the standardized raster layers can then be overlaid to (a) create a raster stack from which (b) overlaps are identified through a simple addition or multiplication across the raster images (e.g. a multiplied value of 6 would indicate that the first two layers labeled 2 and 3 reported disturbance at that pixel). If desired, the resulting (c) spatial agreement map can be overlaid with a raster layer of (d) reference data on disturbance (e.g. fire occurrence) and used to assign a disturbance type. The final spatial agreement map (with or without disturbance type) are translated to measures of uncertainty of disturbance based on the number of products that reported disturbance at that pixel. For example, if using three disturbance products, only one product reporting disturbance at a given pixel indicates a high uncertainty of disturbance, while two products or all products reporting disturbance at a given pixel indicates a medium or low uncertainty of disturbance, respectively.

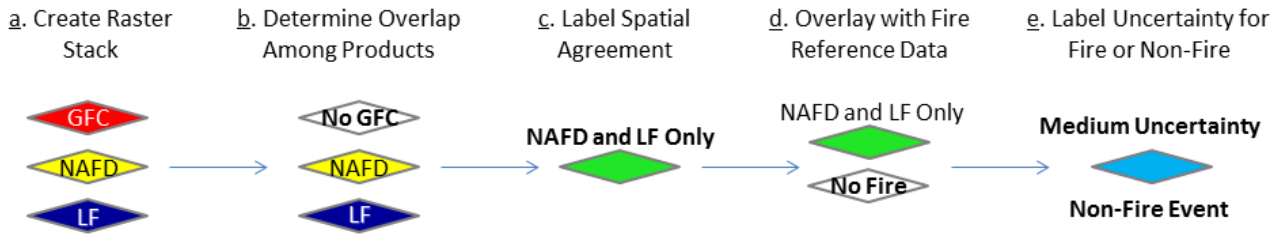


Figure 4-1. Methodology exemplified with products used in the case study. In this applied example of the methodology, spatial agreement at the given location exists only between NAFD and LANDFIRE (LF) and there is no spatial overlap with the fire reference data. As such, uncertainty is identified as a medium level of uncertainty for a non-fire disturbance.

Application of methodology to case study of California fire

For this Chapter, I conducted my analysis using Google Earth Engine (EE), a cloud-based distributed computing platform that provides geospatial functionality, such as pixel-based calculations based on overlaying multiple raster layers and calculating pixels areas contained within pre-defined zones of raster layers or features within vector layers (Gorelick et al., 2017). I chose to use GEE due to the number of pixels that would be analyzed for the study area of California: approximately 450 million at a 30 m pixel resolution for each raster layer included in the raster calculation. The initial map of spatial agreement was created by overlaying standardized versions of the vegetation disturbance products (from Chapter Three), which labeled pixels based on reported disturbance at each pixel anytime between 2001 and 2010. The resulting map contains eight categories of spatial agreement (Figure 4-2): (1) no reported disturbance; (2-4) disturbance reported by only one product (GFC, NAFD, or LANDFIRE, exclusively); (5) by GFC and NAFD only; (6) GFC and LANDFIRE only; (7) NAFD and LANDFIRE only; and (8) by all three products. The spatial agreement categories were then aggregated to label uncertainty of disturbance based on the number of products that reported disturbance at a given pixel (Figure 4-3): (1) no reported disturbance; (2) high uncertainty of disturbance (only one product reported disturbance); (3) medium uncertainty of disturbance (two products reported disturbance); and (4) low uncertainty of disturbance (all products reported disturbance).

The areas covered by each agreement and uncertainty level were calculated using the EE function called ee.Image.pixelArea, which calculates areas using binary images in which pixels with the chosen attribute are labeled with a 1 (e.g. the chosen spatial agreement category), while all other pixels are given a value of 0. The results are provided as percentages of the total area of California, based on the contiguous area of California (i.e. approximately 408,642 square kilometers, excluding islands), which was derived from a publicly available vector dataset of state boundaries provided by ESRI. Additionally, the results are provided as percentages of the total area that was reported as disturbed by at least one product (i.e. all levels with the exception of "no reported disturbance", approximately 40,021 square kilometers). Detailed results are provided in Appendix 4-1.

Next, I overlaid the calculated spatial agreement map with a raster dataset of fire occurrence, derived from the vector-based FRAP fire perimeter dataset for 2000-2010, to quantify uncertainty of disturbance across fire and non-fire events. Additional information on the FRAP fire perimeter data is provided in supplemental material for Chapter Three (Appendix 3-1). Fire occurrence in 2000 was included to account for pixels in the vegetation disturbance products that likely would have been reported as disturbed in the first year of analysis (2001). Pixels that overlapped with the fire occurrence raster were labeled with a fire disturbance event, while pixels that did not overlap with the fire occurrence raster but were reported as disturbed by at least one product were attributed to non-fire disturbance events. As reliable and complete reference data on non-fire disturbance (e.g. harvesting or tree mortality due to drought or

pestilence) are not readily available with full coverage across the state, non-fire disturbance was treated as an aggregated category of disturbance.

To examine biogeographical variations in uncertainty of disturbance across California, the results were summarized by habitat type derived from the CALFIRE FVEG raster dataset (Table 4-1) and by California bioregions as defined by a vector dataset from the Jepson Herbarium Project (Table 4-2). Details regarding the bioregions used in this study can be found in the supplemental material (Appendix 4-1); definitions of the habitat types (scrub/shrub; forest; grass; and other) can be found in the supplemental material for Chapter Three (Appendix 3-1). For habitat type, the area calculations were completed using the EE function called ee.Image.pixelArea to calculate the area of binary images for each combination of uncertainty, disturbance type, and habitat type (e.g. total area of pixels labeled with high uncertainty attributed to a fire event in scrub/shrub). For the bioregions, I used the EE function, ee.Reducer.frequencyHistogram (intended for vector datasets), to complete a zonal calculation that provided a count of the pixels of each combination of uncertainty and disturbance type by bioregion and region (the aggregated unit of multiple bioregions) (e.g. number of pixels labeled with high uncertainty attributed to a fire event in the bioregions of the Sierra Nevada region). The results are provided as percentages of the total area or number of pixels in the respective habitat type or bioregion (e.g. the proportion of all scrub/shrub area across California) as well as percentages of the total reported disturbed area in that habitat type or bioregion (e.g. the proportion of scrub/shrub area across California reported as disturbed by at least one product) (Appendix 4-1).

The final portion of this analysis examined uncertainty of disturbance across burn conditions by comparing the areas with low and high uncertainty of disturbance across California to the areas reported as burned by reference data on fire in California, organized by FRAP fire perimeter size and MTBS maximum burn severity for 2000-2010 (as described in supplemental material for Chapter Three, Appendix 3-1). The areas with low, medium and high uncertainty of disturbance were calculated across four maximum burn severity levels (unburned to low; low; medium; and high) using ee.Image.pixelArea, and across six fire perimeter size classes based on acreage (less than 100; 100-500; 500-1,000; 1,000-10,000; 10,000-90,000; and greater than 90,000) using ee.Reducer.frequencyHistogram. The results of uncertainty of disturbance by burn severity are provided as percentages of the total area in the burn severity class, while uncertainty of disturbance within the fire perimeters are provided as the average proportions within the size class (Figure 4-2).

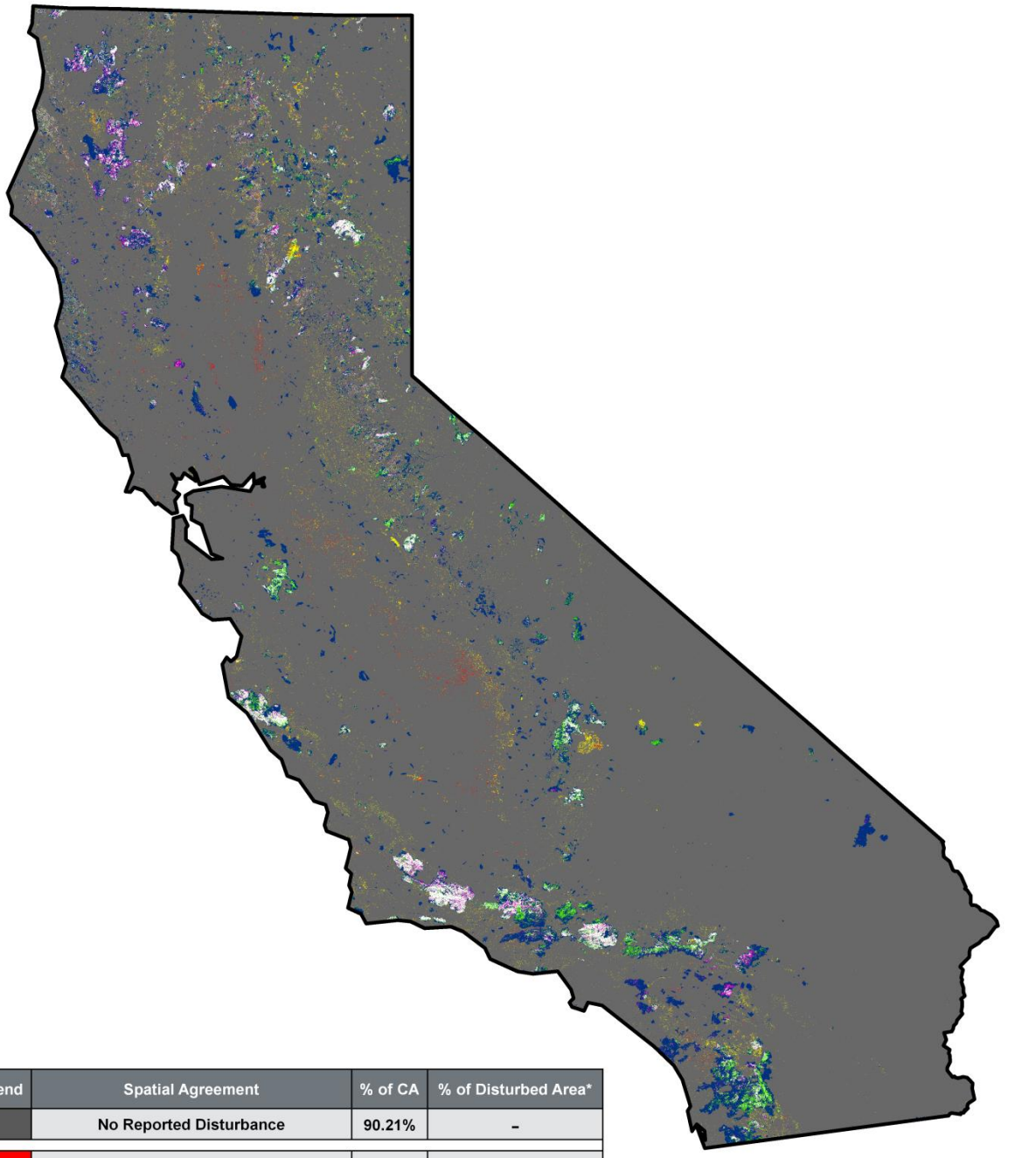
Results

Low spatial agreement highlights high uncertainty of disturbance across California

Spatial agreement among the disturbance products indicated that 9.79% of the total contiguous area of California was reported as disturbed by at least one product between 2001 and 2010, covering approximately 40,021 square kilometers (i.e. total area reported as disturbed) (Figure 4-

2; Appendix 4-1). The highest agreement category (i.e. all products reported disturbance) accounted for only 1.47% of California (approximately 15% of the total area reported as disturbed), indicating overall low spatial agreement among the products. LANDFIRE and NAFD more frequently agreed with each other (12.39% of the total area reported as disturbed) than did either with GFC; agreement between GFC and NAFD was especially low (only 1.15% of the total area reported as disturbed). Furthermore, the disturbance products significantly varied in the amount of area for which they were the sole product to report disturbance. By itself, LANDFIRE accounted for 51.79% of the total area reported as disturbed (30.34% specifically attributed to fire), while GFC and NAFD individually accounted for much lower percentages at 3.02% and 9.92% of the total area reported as disturbed (only 0.07% and 0.58% attributed to fire), respectively (Appendix 4-1).

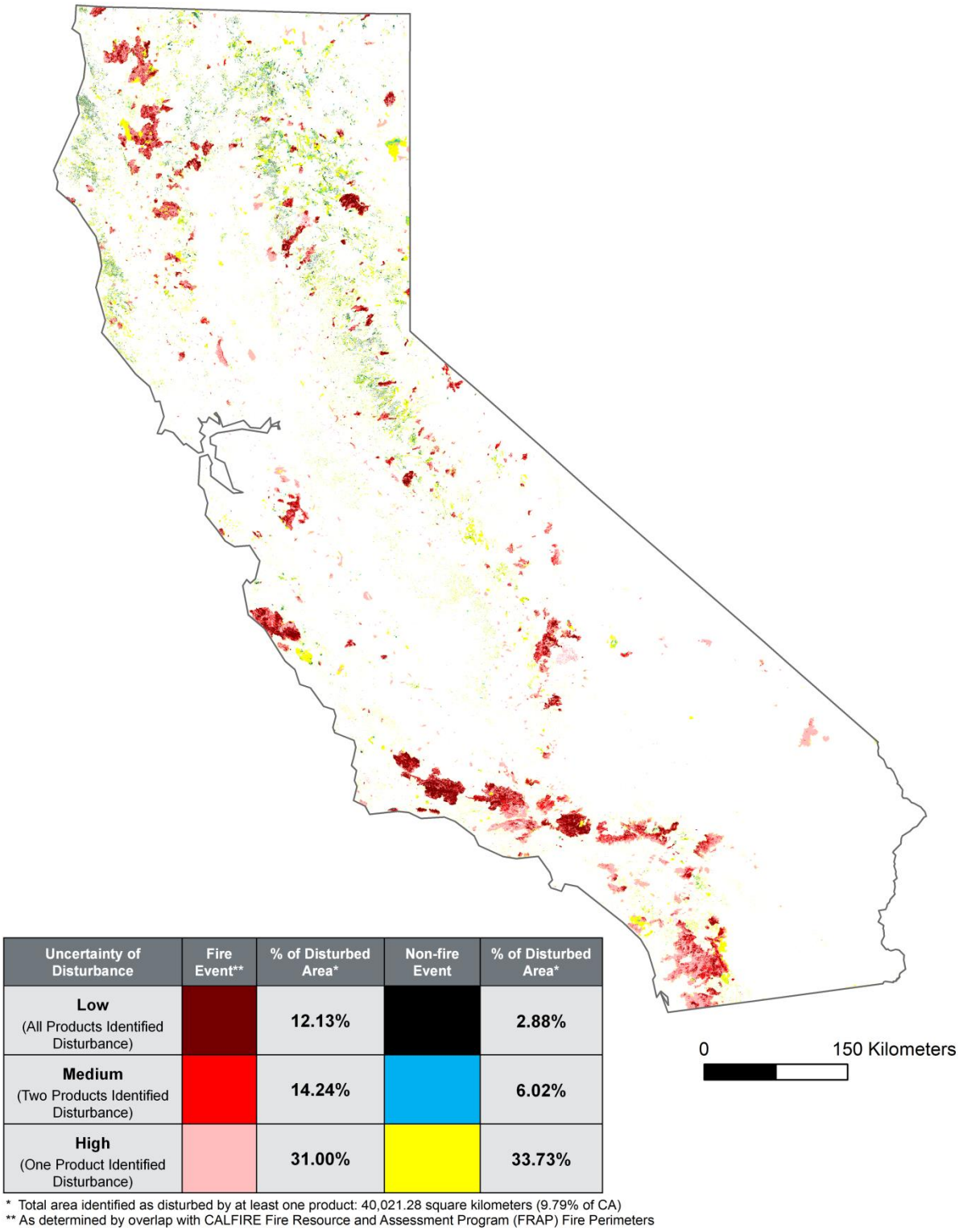
The uncertainty of disturbance map highlighted high uncertainty of reported disturbance for California between 2001 and 2010, particularly for non-fire events (Figure 4-3). The highest uncertainty level (i.e. only one product reported disturbance) covered 64.73% of the total area reported as disturbed across California, relatively split between fire and non-fire events (31% and 33.73% of the total area reported as disturbed, respectively). However, at the medium uncertainty level (i.e. two products reported disturbance), the area attributed to fire (14.24%) was more than double of the area attributed to non-fire events (6.02%), resulting in medium uncertainty of disturbance covering 20.26% of the total area reported as disturbed across California. Agreement between NAFD and LANDFIRE was the greatest contributor to the medium uncertainty level (12.39% of the total area reported as disturbed), particularly regarding fire events (8.99%) (Appendix A). Furthermore, low uncertainty of disturbance (only 15% of the total area reported as disturbed across California) was primarily attributed to fire events (12.13%), rather than non-fire events. While the areas with low and medium uncertainty were higher for fire events than non-fire events, the overall low area totals in the low and medium uncertainty categories across California indicate high uncertainty regardless of disturbance type.



Legend	Spatial Agreement	% of CA	% of Disturbed Area*
	No Reported Disturbance	90.21%	–
	GFC Only	0.30%	3.02%
	NAFD Only	0.97%	9.92%
	LANDFIRE Only	5.07%	51.79%
	GFC and NAFD	0.11%	1.15%
	GFC and LANDFIRE	0.66%	6.72%
	NAFD and LANDFIRE	1.21%	12.39%
	All Products Reported Disturbance	1.47%	15.01%

* Total area identified as disturbed by at least one product: 40,021.28 square kilometers (9.79% of CA)

Figure 4-2. Map of spatial agreement among the vegetation disturbance products. Agreement between GFC and NAFD (orange) is negligible at displayed map scale.



Uncertainty of Disturbance	Fire Event**	% of Disturbed Area*	Non-fire Event	% of Disturbed Area*
Low (All Products Identified Disturbance)		**12.13%**		**2.88%**
Medium (Two Products Identified Disturbance)		**14.24%**		**6.02%**
High (One Product Identified Disturbance)		**31.00%**		**33.73%**

* Total area identified as disturbed by at least one product: 40,021.28 square kilometers (9.79% of CA)
** As determined by overlap with CALFIRE Fire Resource and Assessment Program (FRAP) Fire Perimeters

Figure 4-3. Map of uncertainty of disturbance derived from spatial agreement. Low uncertainty of non-fire events (black) is highly scattered, primarily throughout Northern California and is not easily visible at displayed map scale.

The examination of uncertainty of disturbance by habitat type across California indicated that scrub/shrub was most frequently reported as disturbed as well as identified to have the highest amount of area in the lowest uncertainty of disturbance category (Table 4-1). Overall, the proportion of scrub/shrub reported as disturbed by at least one product (22.51%) was the highest of all habitat types, and of all scrub/shrub across California, 4.49% was identified to have a low uncertainty of disturbance (almost entirely attributed to fire, 4.21%), more than double the amount across all forest (2.35%, with 1.67% attributed to fire). Scrub/shrub and forest shared more similar values for low uncertainty of disturbance when limiting their areas to only habitat reported as disturbed by at least one product, with 19.96% of disturbed scrub/shrub identified to have a low uncertainty of disturbance (almost entirely attributed to fire events at 18.71%) and 14.60% of disturbed forest area identified to have low uncertainty of disturbance (with a smaller proportion attributed to fire events, 10.36%). Overall, 44.51% of the scrub/shrub reported as disturbed had a low or medium uncertainty, as compared to only 35.32% of forest.

Table 4-1. Uncertainty of disturbance by habitat type

Habitat Type	Total % Habitat Area Reported as Disturbed	% Low Uncertainty	% Medium Uncertainty	% High Uncertainty
Percentage of all habitat type area across California				
Scrub/Shrub	22.51%	4.49% (4.21%)[1]	5.52%	12.5%
Forest	16.10%	2.35% (1.67%)[1]	3.33%	10.41%
Grass	7.45%	0.59% (0.36%)[1]	0.72%	6.15%
Other	1.89%	0.09% (0.06%)[1]	0.2%	1.6%
Percentage of habitat type area reported as disturbed only				
Scrub/Shrub	-	19.96% (18.71%)[1]	24.55%	55.49%
Forest	-	14.6% (10.36%)[1]	20.72%	64.68%
Grass	-	7.91% (4.86%)[1]	9.63%	82.47%
Other	-	4.5% (3.1%)[1]	10.67%	84.82%

[1] percentage attributed to fire events based on overlap with fire occurrence derived from FRAP fire perimeters

Across the California bioregions, uncertainty was universally high, as the average areas with high uncertainty accounted for 64% or more of the total area reported as disturbed across all regions (the aggregated unit of multiple bioregions) (Table 4-2). Only two bioregions (San Gabriel Mountains in the Southwestern California region and Outer South Coast Range in the Central Western California region) of all the bioregions had a low uncertainty covering more than 30% of the total area reported as disturbed (Appendix 4-1). The scrub-dominant Southwestern California region had the highest average for reported disturbance across its bioregions (33.52% of the bioregion areas reported as disturbed by at least one product) and was identified to have the second highest average area with low uncertainty of disturbance across its bioregions (14.04% of the total area reported as disturbed had low uncertainty of disturbance, mostly attributed to fire, 13.69%). In particular, uncertainty of disturbance was lowest in the bioregions of San Gabriel Mountains, Western Transverse Ranges, and San Bernardino Mountains (with low uncertainty covering 31.72%, 20.14%, and 16.52% of their total areas reported as disturbed, respectively) (Appendix 4-1).

In the middle range of reported disturbance, the regions of forest-dominant Cascade Ranges, Northwestern California (which includes the North Coast and Klamath Ranges), and Sierra Nevada had similar values for the average area reported as disturbed, ranging from 10.55% to 15.34% of the bioregion area totals. Looking only at the total area reported as disturbed, the average proportions in the low uncertainty category were also similar, ranging from 8.34% to 11.64% across the bioregions (2.25% to 9.43% attributed to fire). Within these regions in the middle-range of reported disturbance, uncertainty of disturbance was lowest in the bioregions of the Central Sierra Nevada Foothills District, Northern High Sierra Nevada District, and Klamath Ranges (with low uncertainty covering 20.03%, 19.42%, 18.82% of their total areas reported as disturbed, respectively) (Appendix 4-1).

At the lowest end of reported disturbance, the regions of Central Western California (which includes the Central and South Coast ranges as well as the San Francisco Bay Area) and of Great Central Valley, as well as the Great Basin (which includes the White, Inyo, and Warner Mountains) and Desert Provinces had similar low averages for the area reported as disturbed by at least one product, ranging from 2.24% to 8.89% of the bioregion area totals. However, looking only at the total area reported as disturbed, the average proportion with low uncertainty ranged widely, from 0.38% (the lowest average across all regions) to 16.64% (the highest average across all regions), with 0.26% to 15.37% attributed to fire. Within these regions in the low-range of reported disturbance, outlier bioregions with low uncertainty of disturbance included the Outer South Coast Ranges District in the Central Western California Region and Warner Mountains Bioregion of the Great Basin Province (low uncertainty covering 34.61% and 25.25% of their total area reported as disturbed, respectively) (Appendix 4-1).

Overall, the scrub-dominant areas of Central Western and Southwestern California had lower uncertainty as compared to the forest-dominant Sierra Nevada and Klamath mountains, and all of

these areas attributed the majority of the area with low uncertainty to fire disturbance. However, even with outlier bioregions, uncertainty was high across these and all other California bioregions (i.e. 64% or more of the total area reported as disturbed across all regions; Table 4-2).

Table 4-2. Uncertainty of disturbance by California region

California Region[1]	% Region Area Reported as Disturbed	% Low uncertainty[2]	% Medium uncertainty[2]	% High uncertainty[2]
Region with Highest Reported Disturbance				
Southwestern California	33.52%	14.04% (13.69%)[3]	21.60%	64.36%
Regions in Middle-range of Reported Disturbance				
Cascade Ranges	15.34%	8.34% (2.25%)[3]	15.83%	75.83%
Northwestern California	14.36%	11.64% (6.95%)[3]	18.41%	69.95%
Sierra Nevada	10.55%	11.28% (9.43%)[3]	16.73%	71.99%
Regions with Lowest Reported Disturbance				
Central Western California	8.89%	16.64% (15.37%)[3]	18.26%	65.09%
Great Basin Province	6.83%	10.7% (8.83%)[3]	15.85%	73.45%
Great Central Valley	2.78%	0.38% (0.26%)[3]	6.20%	93.42%
Desert Province	2.24%	0.65% (0.62%)[3]	10.83%	88.53%

[1] average provided across all bioregions in the region; see Appendix A for detailed results by bioregion
[2] percentage calculated from area reported as disturbed only (values for low, medium and high sum to 100%)
[3] percentage attributed to fire events based on overlap with fire occurrence derived from FRAP fire perimeters

High uncertainty of disturbance across all fire perimeter sizes and most burn severity classes

My results indicated that uncertainty of disturbance decreased more with burn severity than with fire perimeter size, with the proportions of low uncertainty of disturbance more than doubling with each increase in burn severity (Figure 4-4). At the highest burn severity, 55% of the area had low uncertainty of disturbance, though 16.47% of the area still had a high uncertainty of disturbance. Across the fire perimeter size classes, high uncertainty of disturbance comprised a large proportion across all fire perimeter size classes (Figure 4-4). While there was an observable decrease in the area with high uncertainty for fire perimeters larger than 1,000 acres, the highest average value for low uncertainty was only 27.59% of the total perimeter area, occurring in largest fire perimeter size class (greater than 90,000 acres). In general, while low uncertainty of disturbance incrementally increased with both fire perimeter size and burn severity (Figure 4-4), my results also highlighted large differences between the burned areas reported by the reference data and the areas with low uncertainty of disturbance derived from spatial agreement among the three disturbance products (e.g. Figure 4-5, MTBS burn severity overlaid by low uncertainty of disturbance map).

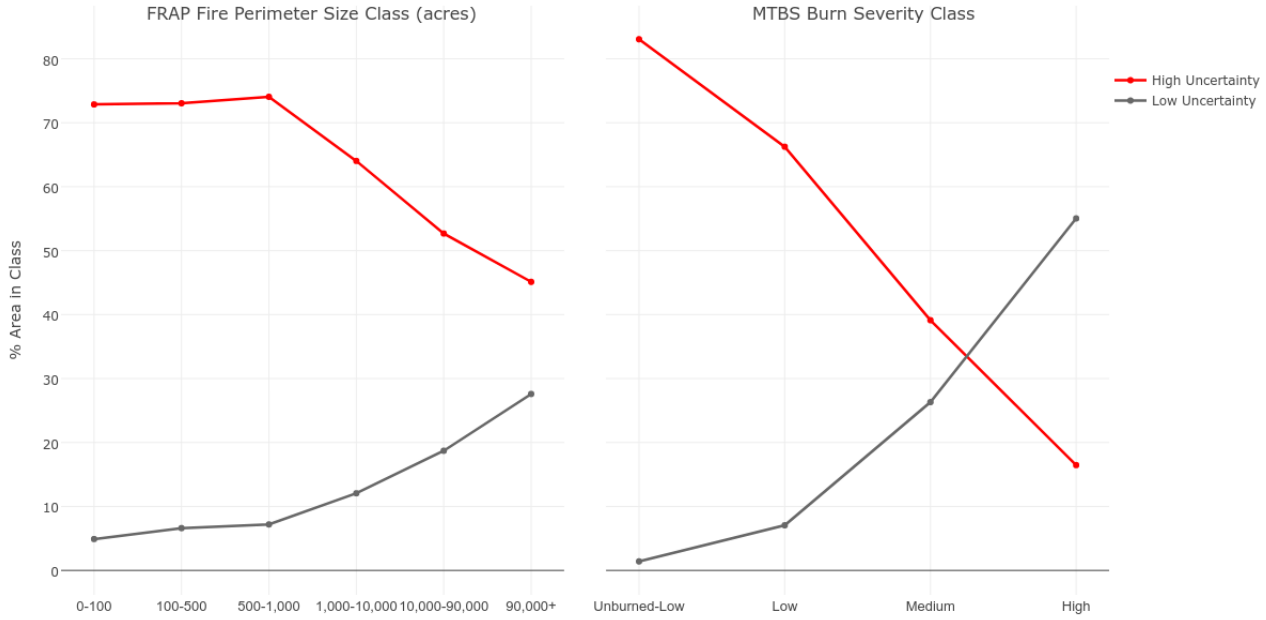


Figure 4-4. Uncertainty of disturbance by burn condition. High uncertainty of disturbance is identified where only one product reported disturbance, while low uncertainty is identified where all products report disturbance. Percent of areas in each fire perimeter size class are calculated based on overlap with fire perimeters from CALFIRE Fire Resource and Assessment Program (FRAP), while percent of areas in each burn severity class are based on overlap with data from Monitoring Trends in Burn Severity (MTBS).

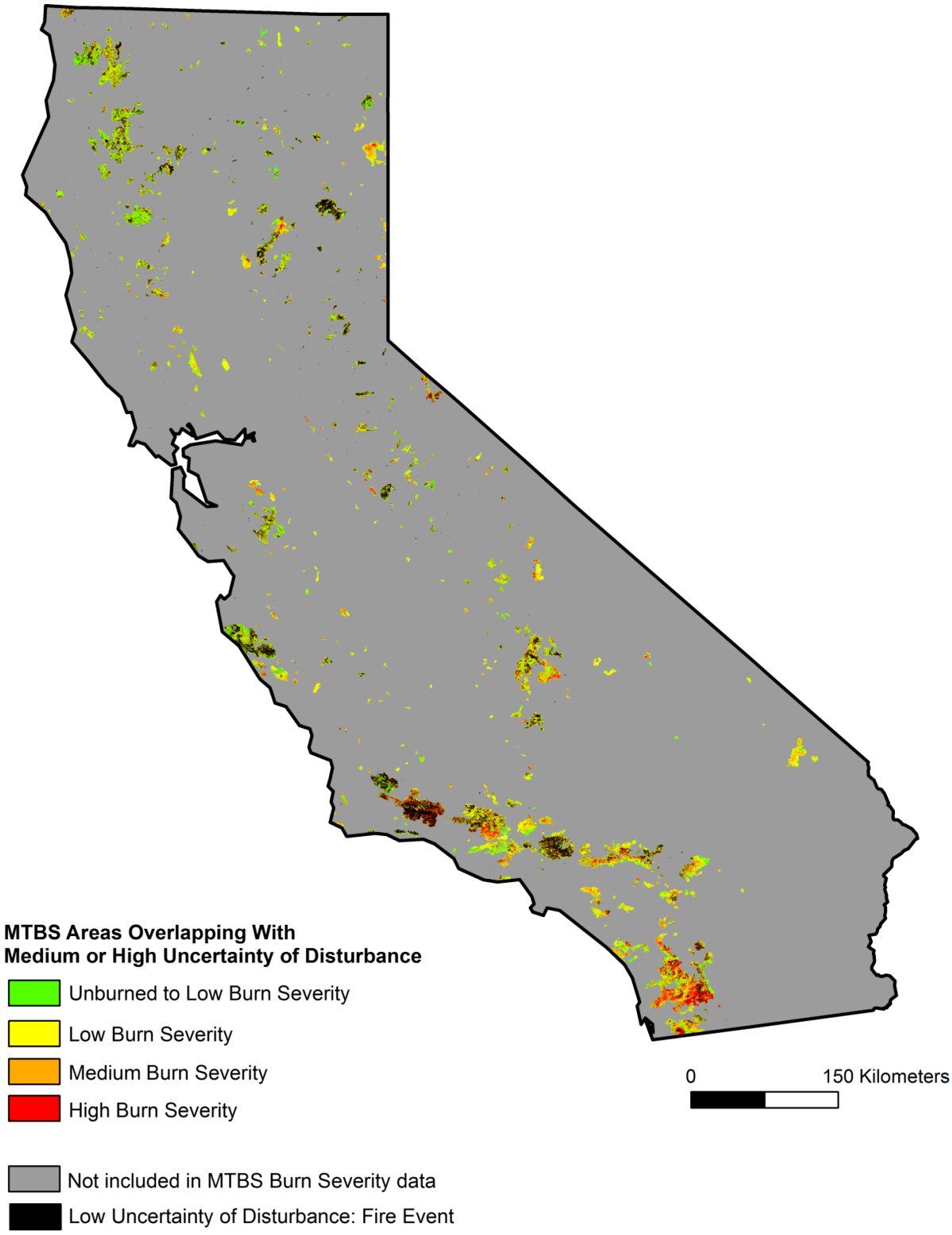


Figure 4-5. Map of burn severity overlaid by low uncertainty of disturbance areas attributed to fire events (black). Based on overlap with data from Monitoring Trends in Burn Severity (MTBS).

Discussion

Implications of high uncertainty of disturbance in products and reference data

While the recent proliferation of these disturbance products has supported research examining relationships between vegetation disturbance and aboveground carbon dynamics, uncertainty in the location and amounts of disturbance have strong implications for determining the carbon trajectory of a given area (i.e. toward more loss or more storage). While my study indicated high uncertainty of disturbance across all bioregions (Table 4-2; Appendix 4-1), including the forest-dominant Sierra Nevada and Klamath mountains, several recent studies of carbon dynamics in California have concluded that these same areas have been and will continue to primary contributors to carbon loss due to disturbances such as fire and harvesting, followed by fire in the scrub-dominant areas of Central and Southern California (Gonzalez et al., 2015; Hurteau et al., 2014; Liu et al., 2011). All of these studies used LANDFIRE as the primary dataset, with some including reference data from MTBS and FIA. Furthermore, my results also indicated the uncertainty was generally high across all habitat types (Table 4-1), though scrub/shrub was identified to have lower uncertainty of disturbance than forest. While one study did acknowledge notable uncertainty related to the low accuracy of the LANDFIRE vegetation classification as well as the ordinal attribution of new characteristics to pixels as they are disturbed (i.e. changes to vegetation height or composition only occur due to a discrete change in vegetation type), uncertainty in the location or amount of disturbance is not explicitly considered (Gonzalez et al., 2015). Rather, the study definitively concluded that "wildfires on 6% of the state analysis area produced two-thirds of the live carbon stock loss" (pg. 68). As such, it is clear that uncertainty in the location and amount of disturbance needs to be more thoroughly accounted for in studies that estimate aboveground carbon dynamics across habitat types and bioregions.

In addition to the uncertainty of disturbance products like LANDFIRE, it must be noted that the FRAP and MTBS reference data are also representations of fire with their own biases and uncertainties that introduce uncertainty. For example, because manually created fire perimeters (such as those provided by FRAP) enforce the assumption that all pixels within the demarcated perimeter burned in that fire event, researchers have recognized that fire perimeters often include large amounts of unburned areas, as they are often delineated at larger extents than the area likely to have actually burned (Kolden & Weisberg, 2007; Kolden, Lutz, Key, Kane, & van Wagtendonk, 2012). Similarly, researchers have questioned the ability of Difference Normalized Burn Ratio (dNBR) analysis (used by MTBS) to accurately classify burn severity across all ecosystem types or to even be consistently comparable across a single landscape (French et al., 2008; Kolden, Smith, & Abatzoglou, 2015; Parker, Lewis, & Srivastava, 2015; Roy, Boschetti, & Trigg, 2006)

In comparing the results of this Chapter to my past work (Chapter Three), I also identify that both FRAP and MTBS reported higher amounts of burned areas than the total area that was found to have low uncertainty of disturbance. The total burned area in FRAP and MTBS between

2000 and 2010 represented approximately 6% of the area of California (Chapter Three), while the total area with the low uncertainty of disturbance between 2001 and 2010 identified in this analysis was less than 1.5% of California (Figure 4-1). Furthermore, FRAP and MTBS each reported similar proportions of burned area within each habitat type (Chapter Three) that were noticeably larger than the amount of that habitat area with low uncertainty of disturbance. In particular, both reference data indicated that approximately 18% of all scrub/shrub across California experienced a fire event between 2000 and 2010, while less than 5% of all scrub/shrub was identified in this study to have a low uncertainty of disturbance between 2001 and 2010. Similarly, the reference data indicated that approximately 7% of forest and 5% of grass had experienced a fire event, while less than 2.5% of forest and 1% of grass were identified to have a low uncertainty of disturbance (Table 4-1).

High uncertainty of disturbance across all fire perimeter sizes and all but the highest burn severity (Figure 4-4) further indicate that the reference data may actually be overestimating the true amount of burned area in California. On average, high uncertainty of disturbance covered over 45% of the areas of the individual fire perimeters across all fire perimeter size classes. These results are consistent with other research that have identified overestimation (by an average of 18%, up to 37%) by manually mapped fire perimeters (such as those from FRAP), as compared to remotely-sensed fire perimeters (Kolden & Weisberg, 2007; Kolden et al., 2012). The discrepancy between burned area reported by MTBS and areas with low uncertainty of disturbance (Figure 4-5) is also supported by research concluding that independent classification of burned areas resulted in higher accuracy than use of MTBS data to identify burned areas (Meddens, Kolden, & Lutz, 2016); these researchers reported that the average proportion of unburned areas within remotely-sensed fire perimeters was 20%, with typically higher proportions of unburned area within non-forest.

The uncertainty of disturbance highlighted by this study also points to the interesting dilemma introduced by a "data-rich" world, in which there are many competing datasets that represent the landscape based on the assumptions and processes by which they are created. For example, in contrast to research in the U.S. that points to overestimation of burned areas in state and federal reference data, international researchers argue that global remote sensing-based products (for which reference data are limited) greatly underestimate burned areas due to the exclusion of small fires (Nogueira, Ruffault, Chuvieco, & Mouillot, 2016). Furthermore, my previous work (Chapter Three) demonstrated that LANDFIRE reported much higher disturbance than GFC and Hansen primarily due to the inclusion of small FRAP fire perimeters that may not be picked up by the automated analyses of the latter two. However, fire perimeters smaller than 1,000 acres were shown in this study to have high uncertainty of disturbance. In addition, as LANDFIRE was solely responsible for approximately 50% of the total area reported as disturbed (primarily attributed to fire events), it is likely that the greater inclusion of smaller fires actually contributed to over-reporting of disturbance by LANDFIRE. In contrast, GFC and NAFD were individually responsible for only 3% and 10% of the total area reported as disturbed, respectively, primarily

attributed to non-fire events. Additional research investigating the omission of burned areas by GFC and NAFD could identify the extent to which these products may be underreporting disturbance by fire events at local and regional scales.

While it is possible that an individual product (e.g. LANDFIRE) could be the sole reporter of disturbance at a particular location because it accurately identified disturbance at that location, the goal of this Chapter was to quantify and examine uncertainty of disturbance, rather than accuracy. For example, for LANDFIRE, being the sole reporter at a given location could also be due to the product having a lower threshold for disturbance (as it integrates data reported by other agencies without additional review). This is supported by that fact that it was the sole reporter of 50% of the total area reported as disturbed, while NAFD and GFC were only sole reporters for 3% and 10%, respectively (Figure 4-1). For NAFD, sole reporting could be attributed to higher sensitivity to continuous disturbance as compared to GFC, and to continuous disturbance not being readily identified and reported by agencies contributing data to LANDFIRE. For GFC, sole reporting could be due to a high "salt and pepper" effect that often results from pixel-based supervised classification methods (Liu & Xia, 2010; Yu et al., 2006). As such, the creation method of each product (see detailed descriptions of products in Chapter Three) contributes to uncertainty of the reporting of disturbance at a given location and is an important measurement that is separate from an accuracy assessment of whether disturbance truly occurred at that location.

Data integration as a methodology to account for uncertainty of disturbance

This study presented a simple data integration approach (based on spatial agreement) for dealing with the data deluge in remotely-sensed products of vegetation change and disturbance. In contrast to other approaches such as algorithm and model ensembling, I believe that my methodology is not only easier to implement but is also more flexible and able to accommodate a variety of data, regardless of how they are created. Specifically, while remote sensing experts are using algorithm ensemble modeling to account for uncertainty of disturbance (Healey et al., 2017), not all disturbance products are the results of reproducible algorithms that can be ensembled. For example, LANDFIRE, a widely used product, is created by combining manually gathered data on disturbance events reported by federal, state, and local agencies (often vector-based data), with remotely-sensed vegetation and burn indices calculated from the LTS through semi-automated analyses (see Chapter Three for more details). Furthermore, the complexity of algorithm ensembling and other methods such as multi-step predictive modeling (Schroeder et al., 2017) can make these frameworks unapproachable for users of these disturbance products who may have limited programming or remote sensing expertise. To this end, a data integration approach through basic raster calculations across chosen products can help address the “black-box” of ensembling approaches, particularly some machine learning methods that do not provide information on which features included in the models are most important (and to what extent) for accurate identification of disturbance. Overall, my data integration approach puts the power of uncertainty analysis back into the hands of users of these disturbance products, particularly non-

remote sensing experts who do not have the resources to conduct a more complex analyses and/or a subsequent accuracy assessment.

This data integration approach through spatial agreement can also provide insight into the contributions of individual products to overall uncertainty of disturbance, similar to how some ensemble models report the contribution of each algorithm to overall accuracy. This information can help users to further understand whether and how these products can be used in complementary ways. For example, my case study concluded that NAFD and LANDFIRE more frequently agreed with each other than did either with GFC, and that the area covered by NAFD-LANDFIRE agreement was almost as much as the area covered by the complete agreement category (i.e. low uncertainty of disturbance) (Figure 4-1). This is perhaps not surprising given that the algorithm behind NAFD is focused on identifying continuous disturbance (i.e. any change or reduction in green vegetation), and LANDFIRE also attempts to incorporate continuous disturbance, while GFC only identifies discrete disturbances (i.e. stand-replacement change) (see Chapter Three for more details). However, it is surprising that agreement between NAFD and GFC was very low, given that both NAFD and GFC have a focus on forested areas, while LANDFIRE does not focus on any specific habitat type. As GFC was individually responsible for the least amount of the total area reported as disturbed, any spatial agreement combination with GFC would also result in low area totals. However, the conservative reporting of disturbance by GFC can be considered to have narrowed uncertainty even more, given that some of the continuous disturbance identified by NAFD and LANDFIRE may have minimal ecological impact (i.e. slight reductions of greenness due to natural ecological fluctuations). Furthermore, very little of the area with low uncertainty of disturbance was attributed to non-fire events, indicating that these products may not collectively provide useful data about the location and amount of non-fire disturbance such as harvesting or mortality from drought and pestilence. Given the dynamics of multi-disturbance regimes as well as potential overestimation of fire by the reference data, it is likely that incorporating reference data on non-fire events would actually result in large overlaps with the reference data on fire events, thereby introducing new uncertainties that need to be examined.

While I acknowledge that this Chapter is not the first study to explore data integration of remotely sensed products, other studies have focused more on improving accuracy in identification of disturbance, rather than on quantifying or providing measures of uncertainty. For example, Schroeder et al. (2017) presented a different methodology for data integration based on a multi-step predictive model that integrates LANDFIRE and MTBS with other data (such as imagery from Google Earth and the National Agriculture Imagery Program, NAIP) as predictors to more accurately map five different types of forest disturbance. While this is an interesting and robust example of data integration, the complexity of this methodology likely resulted in the limited implementation to only ten LTS scenes across the U.S. Similarly, Soulard et al. (2017) also provided a useful data-driven methodology to improve accuracy in identification of forest disturbance and quantify magnitude of disturbance through “harmonized

maps" (pg. 169) that integrate multiple products (including GFC, NAFD, and LANDFIRE) and reduce commission errors. However, as pixels where only one product reported disturbance were not included in their final product, it is possible that high omission errors in their final product were primarily due to not including these pixels that my study labeled to have high uncertainty. Other researchers such as Kolden et al. (2015) have explored the integration of LANDFIRE and MTBS (both of which may be overestimating disturbance according to my results) to more accurately and consistently identify thresholds of burn severity across different habitat types and bioregions, though they concluded that more integration of data collected in the field, rather than remotely-sensed data, was needed for higher accuracy. In future work, my methodology and results could be combined with these other data integration approaches to address both accuracy and uncertainty of a given study area or time period.

Moving forward, there are likely to be even more standardized disturbance products that can be incorporated into uncertainty analyses: from potential products developed from the LandTrendr algorithm (Kennedy et al., 2010) and the Landsat Burned Area Essential Climate Variable (BAECV) (Vanderhoof, Fairaux, Beal, & Hawbaker, 2017) to the upcoming Land Change Monitoring, Assessment and Projection (LCMAP), which is based on Continuous Change Detection and Classification (CCDC) algorithm (Zhu & Woodcock, 2014) and currently in validation phase before publication by USGS. As new vegetation disturbance metrics and products continued to be developed from LTS and possibly from other satellite sensors such as Sentinel 2 or WorldView 4 in the future, my methodology for data integration based on spatial agreement can be expanded and adapted to include new data options as they become available. Future work using my data integration approach could also examine patterns of uncertainty (e.g. spatial clustering; characteristics of individual fires with low or high uncertainty) as well as adapt my approach to integrate data across different sensor platforms and measurement scales such as non-standardized data from UAVs and micro-satellites. My flexible methodology allows for the combination of spatial agreement into uncertainty categories in different ways than were presented in this study (particularly if using a much higher number of products or if some products are identified to be complementary), and exploration of spatial clustering within and across uncertainty categories to explore potential issues introduced by raster stacking (particularly if trying to integrate across different satellite sensors and platforms for which pixel alignment may not match completely).

Conclusion

Low spatial agreement among the vegetation disturbance products resulted in less than 1.5% of California having a low uncertainty of disturbance between 2001 and 2010, despite the fact that almost 10% of California was reported as disturbed by at least one product in the time period. Most of the reported area with low uncertainty of disturbance was attributed to fire events, rather than non-fire events, indicating that spatial agreement among these products may not be appropriate for investigating non-fire disturbance (such as harvesting or mortality due to drought and pestilence). While I found that LANDFIRE and NAFD more frequently agreed with each

other than did either product with GFC (agreement between GFC and NAFD being especially low), I also identified that LANDFIRE was solely responsible for approximately 50% of the total area reported as disturbed by at least one product, attributing the majority of the disturbed area to fire events. On the other hand, GFC and NAFD both individually accounted for less than 10% of the total area reported as disturbed, and both attributed the majority of the disturbed area to non-fire events. These results point to both potential overestimation in disturbance by LANDFIRE as well as underestimation by GFC and NAFD, particularly for fire events. In my examination of uncertainty across biogeographical divisions, I found that scrub/shrub had a lower uncertainty of disturbance than forest, while uncertainty was universally high across all bioregions. As such, my results highlighted important implications of uncertainty of disturbance for studies of aboveground carbon dynamics in California, which have previously concluded that the forest-dominant Sierra Nevada and Klamath-Siskiyou regions have been and will continue to be the primary contributors of aboveground carbon loss, followed by scrub-dominant Central and Southern California. Furthermore, I identified large differences between the burned areas reported by the reference data and the areas with low uncertainty of disturbance, as derived from spatial agreement among the disturbance products. As such, my results indicated potential over-estimation of disturbance by the FRAP and MTBS reference data themselves. Overall, my investigation of uncertainty of disturbance for California identified the strengths and weaknesses of the individual and collective detection capabilities of these products and provided a clear case study of a simple but powerful methodology for using a data integration approach based on spatial agreement to quantify uncertainty across broad spatial and temporal extents.

Chapter Five

Conclusions and Future Directions for Spatial Data Science

Spatial Data Science presents both opportunities and challenges for geospatial researchers and practitioners. There are more options for geospatial data and analytical tools than ever before (particularly for earth and environmental monitoring and assessment), but these new tools have also helped to create new challenges surrounding how to evaluate and choose between data, software and tool options, how to integrate them into unified solutions, and how to evaluate their quality and uncertainty. Furthermore, as research teams investigate environmental challenges such as climate change, wildfire management, and the loss of biodiversity and natural areas that require increasingly more interdisciplinary collaboration as well as more complex geospatial analyses at broader extents and finer spatial resolutions, they also need to be able to evaluate analytical tools specifically on their functionality that supports collaborative completion of geospatial tasks. This dissertation addressed in part these challenges by successfully applying Spatial Data Science concepts and techniques to develop new frameworks for evaluating geospatial tools based on collaborative potential and for evaluating and integrating competing remotely-sensed data products on vegetation change and disturbance.

Motivated by the lack of quantitative evaluation of geospatial tools based on their collaborative potential, Chapter Two developed a reproducible framework for evaluating multi-user geospatial tools based on their support for collaborative tasks and provided the first published typology of collaborative geospatial tools. While there had been efforts in the past to qualitatively evaluate and cluster geospatial software packages by various capacities, there had not been an attempt to evaluate the growing landscape of tools in a quantitative way. In Chapter Two, I outlined a collaborative Spatial Data Science workflow as the backbone of my evaluation framework to score and cluster multi-user geospatial tools (particularly open source and web/cloud-based options) based on their technical functionality that supports collaborative completion of geospatial tasks (i.e. setting up a working environment, data wrangling, analysis, and visualization/publication). I presented my results as a map of the emergent ecosystem of collaborative geospatial tools with three primary niches of tools: (1) participatory data aggregators; (2) content managers; and (3) highly scalable and customizable tools. While I discussed the advantages and disadvantages of each niche (based on user involvement and needed infrastructure), I also concluded that no single tool can meet all project needs, and that my framework can help researchers and practitioners to evaluate how collaborative geospatial tools meet certain needs and to explore ways that multiple tools can be integrated within the overall Spatial Data Science workflow.

One of the key challenges facing researchers and analysts in the spatial data science discipline is meaningful integration of multiple data streams. There exist increasing amounts and types of spatially referenced products derived from remote sensing imagery, and often it is difficult to

understand their relative quality, utility, and provenance. Motivated by the absence of comparative evaluations of remotely-sensed disturbance products of vegetation change and disturbance, in Chapter Three, I compared the three published products with overlapping geographic and temporal extents for California (at the time of this dissertation) to identify drivers of uncertainty in vegetation disturbance (particularly for fire) and help end-users of these products understand how they report disturbance and the conditions under which they identify it. I found large differences in reported disturbance by the products that resulted from their differing methods of creation (i.e. explicit inclusion of reference data by LANDFIRE due to its manual creation process), rather than differences in their coverage of environmental conditions. Overall, LANDFIRE (a product which uses a more traditional workflow in its production through combining vector-based of disturbance events with remote sensing analyses) reported more disturbance across California for all years and all habitat types than did the two products that used automated, big data approaches to identifying disturbance in satellite imagery (i.e. Hansen Global Forest Change, GFC, and North American Forest Dynamics, NAFD). When I compared GFC and NAFD, I found the former reported the least disturbance across all years and habitat types, reflecting the major difference in these automated approaches: GFC provides discrete identification of disturbance, while NAFD focuses on continuous disturbance. Specifically, while the algorithm used to create NAFD (Vegetation Change Tracker, VCT) can identify reductions in vegetation and green cover (i.e. continuous disturbance), the workflow used to create GFC (through a combination of standard remote sensing techniques such as supervised classification and time series analysis of vegetation indices) can only identify changes in vegetation that result in a complete change to a new cover type (i.e. discrete disturbance). I also concluded that the differences between the products were greatest in scrub/shrub areas (which were more frequently reported as disturbed by reference data) and in years with more fire incidence (as reported by reference data), indicating that more disturbance reported by reference data actually resulted in more differences among the products due to their differing thresholds for identifying disturbance (from GFC with the highest threshold based on discrete disturbance only to LANDFIRE with the lowest threshold as it incorporates reference data without limitations).

These vegetation disturbance products are used widely in ecology and earth system science as the sole representation of disturbance in carbon studies, and often without regard to uncertainty in these datasets. This sort of product acceptance illustrates another key challenge in spatial data science: understanding and communicating uncertainty in spatial data and model products. In Chapter Four, I provided a simple but powerful methodology for accounting for uncertainty in disturbance based on quantifying spatial agreement among disturbance products. My framework used basic raster calculations (i.e. spatial overlays that can be calculated in any platform that can support satellite images) to identify where the products agreed in reporting disturbance, and then converted spatial agreement to measures of uncertainty of disturbance (e.g. high uncertainty where only one product reported disturbance, but low uncertainty where all products reported disturbance). I found that despite 10% of California being reported as disturbed by at least one product between 2001 and 2010, only 15% of that area had a low uncertainty of disturbance.

Across biogeographical divisions, I generally found that uncertainty was high across all bioregions and habitat types, though I did find that scrub/shrub had lower uncertainty than forest, particularly for fire events. My results also highlighted potential over-estimation of disturbance by both LANDFIRE and the reference data on fire, as LANDFIRE was found to be solely responsible for approximately 50% of the total area reported as disturbed in the study period, and large differences existed between burned areas reported by reference data and the areas with low uncertainty derived from the spatial agreement analysis. Based on my results indicating high uncertainty across California, I concluded that accounting for uncertainty in disturbance is particularly important for studies focused on the ecological implications of disturbance, such as impacts on aboveground biomass and carbon dynamics; rather than choosing only one product to represent disturbance, users of these products (in particular, non-remote sensing experts) can use simple methodologies for data integration (such as the presented framework) to address uncertainty in the location and amount of disturbance.

Directions for future research

Just as each chapter of this dissertation leveraged Spatial Data Science concepts and techniques to develop these frameworks for evaluating and integrating competing options from the plethora of available geospatial tools and data, each chapter also highlighted technical and conceptual challenges that remain for Spatial Data Science. Identifying additional challenges for the future of collaborative geospatial work, Chapter Two provided a list of technical developments needed for stronger functionality within collaborative geospatial tools (such as more integration of the cloud and high performance computing, HPC, for big data handling, stronger spatial-temporal integration, and controlled versioning of individual data features) and concluded with more general discussion of conceptual challenges for Spatial Data Science, including scaling of methods, data synthesis, how to develop tools to best serve users' needs, and centralized production of technology and knowledge. Chapter Three highlighted the critical need to view and use competing remotely-sensed vegetation disturbance products as different representations based on differing thresholds for reporting disturbance and called for new analytical frameworks that can help researchers and practitioners to quantify and account for uncertainty in disturbance products. While Chapter Four addressed this call by providing a new user-friendly framework (particularly for non-remote sensing experts) to account for uncertainty based on data integration, it concluded with a call for more analyses of patterns of uncertainty (e.g. spatial clustering; characteristics of individual fires with low or high uncertainty) and for frameworks to integrate data across different satellite sensors and measurement scales, particularly non-standardized data from UAVs, micro-satellites, and newer satellites that are used for global environmental research such as Sentinel 2 and Worldview 4. While each dissertation chapter addressed different challenges in Spatial Data Science and identified research priorities in specific arenas, I conclude this dissertation by highlighting three avenues of future research that can provide key contributions toward data synthesis needed to address environmental challenges

in the 21st century: scaling of geospatial methods; spatial-temporal integration; and academic curriculum focused explicitly on skills needed by Spatial Data Scientists.

Scaling of methods

In addition to the technical challenges posed by big data handling for scalability of geospatial tools, Spatial Data Science is also challenging the methodologies and techniques traditionally used to identify spatial patterns and processes. As geospatial datasets become larger (through coverage of broader extents, finer resolutions, or more dimensions), researchers are questioning whether the same geospatial methods that were previously used at smaller scales (such as spatial statistics and exploratory visualization techniques) are still useful at a global scale (Anselin 2015). Furthermore, as the complexity of many environmental challenges require the analyses of increasingly multi-dimensional and multi-scale data with larger volumes, these analyses introduce new concerns regarding statistical significance of results from big data. Specifically, researchers such as Gandomi and Hader (2015) and Miller and Goodchild (2015) have identified concerns regarding noise in large data, the applicability of statistical methods when dealing with whole populations rather than samples, and other issues resulting from messy and disparate data. In particular, Miller and Goodchild also highlight the importance of remembering that even in big data, correlations are not causation, while Gandomi and Hader (2015) highlight work from others (Fan and Lv, 2008) to point out that stronger but spurious statistical correlations can result from randomness in big data simply as a result of the size of the dataset. As it is clear that these issues regarding scale, extent, and size are critical areas for Spatial Data Science, future research needs to focus on identifying new principles and methods for working with large datasets, such as the exploration of nested and stacked analyses to create an integrated view of a phenomena over many scales, and alternatives to long-held scientific tenets, such as questioning the relevance of significance values (i.e. *p* values), for large and complex data.

Spatial-temporal integration

Recognizing the complexity and feedback between ecosystem processes across both space *and time*, researchers have been calling for stronger spatial-temporal integration within environmental and ecological analyses, noting particular weakness in the temporal components of most research (Watson et al., 2013; Wolkovich et al., 2014). Within Spatial Data Science applications for identifying and quantifying landscape change, the gains that have been made in the spatial domain of landscape change have not necessarily been matched in the temporal domain. In particular, time series analyses of landscape change have been limited to either year-to-year comparisons or identifications of the year of maximum or last loss, likely due to the lack of computational support for complex spatial-temporal analyses within existing geospatial tools. Future research can continue to leverage HPC to address this disparity through new frameworks that represent vegetation disturbance as multi-year processes, rather than individual, unrelated events. This reframing to an integrated spatial-temporal representation of landscape change provides two critical capabilities: (1) mapping of repeated disturbance as ecological regimes with

successional stages; and (2) differentiation of the impacts of disturbance regimes that appear to be similar but differ due to the historical context of a location. In addition, the rate of disturbance across years (i.e. a comparable metric of acceleration towards increasing or decreasing disturbance) can be interpreted as the temporal stability of vegetation in a given location. As an important application of this framework, identifying these differing trajectories of landscape change can address current limitations of existing land carbon balance studies, resulting from inadequate accounting of vegetation disturbance. For example, total loss has typically been calculated as the sum of point-in-time losses (i.e. aggregation of yearly losses), based on pixels of the same vegetation being assigned equal values of aboveground carbon loss. With spatially-temporally integrated frameworks of vegetation disturbance, total carbon loss can be differentially estimated within vegetation types, based on whether a location is experiencing stability or deceleration (i.e. pixels unlikely to increase in carbon loss or remain as carbon sinks), or acceleration (i.e. pixels that are likely to remain carbon sources).

New curriculum for Spatial Data Science

In addition to collaboration, researchers have also noted the general need for more communication between scientists who model ecological and environmental processes and those who collect data and conduct experiments, citing differences in jargon and technical knowledge as well as issues in data quality, standardization and sharing (Heuschele et al., 2017). Furthermore, other researchers have specifically noted the lack of statistical training of doctoral students in ecological and environmental sciences, in sharp contrast to the advanced techniques that are being used in published ecological work (Touchon and McCoy, 2016). For Spatial Data Science, overcoming similar challenges is critical for achieving the data synthesis needed to tackle environmental challenges in the 21st century (and beyond) and should be a priority in the curriculum that is used to prepare future Spatial Data Scientists. Spatial Data Scientists, like their ecological counterparts, need comparable technical knowledge and language to collaborate effectively on big or complex geospatial data projects. One of the main bifurcations in background knowledge for Spatial Data Science is whether a students' knowledge base originates within the computer science or statistics side (resulting in more technical skills but no domain knowledge) or within geography or other fields such as environmental science (resulting in more understanding of how space and time influence landscapes but weaker technical skills).

To address this current limitation in academic curriculum, three key areas need to integrated within a new cross-disciplinary curriculum for Spatial Data Science: (1) computational training (i.e. scripting, cloud-based and HPC techniques, data management tools for storage, access, and versioning); (2) statistical training for both big data (i.e. machine learning, data mining) and spatial domains (i.e. remote sensing techniques and spatial statistics such as spatial autocorrelation and geographically weighted regression); and (3) visualization and publication tools that support collaboration (i.e. interactive web mapping using web frameworks that support both geospatial and non-geospatial data; development of application programming interfaces).

While there are university and online courses that individually cover these skills, there is not yet a unified curriculum or program that builds on these concepts (particularly as they apply to spatial data) to advance students' knowledge *from* introductory *to* advanced levels needed to become Spatial Data Scientists. Rather, both students and professionals often have to acquire these skills on their own without formal guidance that can help them to appropriately target their learning. As our societal and environmental challenges grow more pressing and complex, I need new curriculum to prepare both students and professionals as Spatial Data Science leaders who can work collaboratively and successfully leverage data and tools to develop the needed policies, research, tools, and/or applications to find solutions at local to global scales.

In conclusion, as Spatial Data Science continues to evolve through the development of new methods and tools for harnessing the ever-increasing amounts of data being collected through new instrumentation and volunteered by citizens, it is clear that interdisciplinarity will continue to be at the core of this emerging discipline. This interdisciplinarity specifically promotes the translation of methods and tools from related disciplines (such as Data and Computer Science) to geospatial questions as well as the applications of these methods and tools across multiple spatial and temporal scales and boundaries (both administrative and ecological). These avenues of interdisciplinarity also support increased scientific and policy collaborations through the development of collaborative tools for geospatial work and guidance on evaluating and integrating competing data and tools. In this light, the grand challenges of Spatial Data Science (i.e. choosing from the plethora of data and tool options, scaling of methods, spatial-temporal integration, achieving data synthesis) become grand opportunities to engage in data discovery, interdisciplinary explorations of spatial-temporal questions, and the harnessing of cutting-edge computational advances, thereby stimulating the breakthrough ideas and solutions needed to address our societal and environmental challenges in the 21st century and beyond.

References

Akamani, K., Holzmueller, E. J., & Groninger, J. W. (2016). Managing Wicked Environmental Problems as Complex Social-Ecological Systems: The Promise of Adaptive Governance. In *Landscape Dynamics, Soils and Hydrological Processes in Varied Climates* (pp. 741-762). Springer International Publishing.

Allen, G. M., & Gould Jr, E. M. (1986). ComRlexi, Wickedness.

Anderegg, W. R. L., Martinez-Vilalta, J., Cailleret, M., Camarero, J. J., Ewers, B. E., Galbraith, D., … Trotsiuk, V. (2016). When a Tree Dies in the Forest: Scaling Climate-Driven Tree Mortality to Ecosystem Water and Carbon Fluxes. *Ecosystems* , *19*(6), 1133–1147.

Andrienko, G., Andrienko, N., Bak, P., Keim, D., & Wrobel, S. (2013). Visual Analytics Focusing on Spatial Events. In *Visual Analytics of Movement* (pp. 209-251). Springer Berlin Heidelberg.

Andrienko, N., & Andrienko, G. (2013). Visual analytics of movement: An overview of methods, tools and procedures. *Information Visualization*, *12*(1), 3-24.

Anselin, L. (2012). From SpaceStat to CyberGIS twenty years of spatial data analysis software. *International Regional Science Review*, *35*(2), 131-157.

Anselin, L. (2015). Spatial Data Science for an Enhanced Understanding of Urban Dynamics. http://citiespapers.ssrc.org/spatial-data-science-for-an-enhanced-understanding-of-urban-dynamics/. Accessed on October 3. 2016.

Arneth, A., Sitch, S., Pongratz, J., Stocker, B. D., Ciais, P., Poulter, B., … Zaehle, S. (2017). Historical carbon dioxide emissions caused by land-use changes are possibly larger than assumed. *Nature Geoscience*, *10*, 79.

Arroyo, L. A., Pascual, C., & Manzanera, J. A. (2008). Fire models and methods to map fuel types: The role of remote sensing. *Forest Ecology and Management*, *256*(6), 1239–1252.

Balint, P. J., Stewart, R. E., & Desai, A. (2011). *Wicked environmental problems: managing uncertainty and conflict*. Island Press.

BBC News. (2013). Forest change mapped by Google Earth. *http://www.bbc.com/news/science-environment-24934790*. Accessed on October 26, 2016.

Beecham, R., Wood, J., & Bowerman, A. (2014). Studying commuting behaviours using collaborative visual analytics. *Computers, Environment and Urban Systems*, *47*, 5-15.

Boisramé, G., Thompson, S., Collins, B., & Stephens, S. (2017). Managed Wildfire Effects on Forest Resilience and Water in the Sierra Nevada. *Ecosystems* , *20*(4), 717–732.

Carroll, M. S., Blatner, K. A., Cohn, P. J., & Morgan, T. (2007). Managing fire danger in the forests of the US Inland Northwest: A classic "wicked problem" in public land policy. *Journal of Forestry*, *105*(5), 239-244.

Ciss, S. (2015). Random Uniform Forests for Classification, Regression and Unsupervised Learning.

Cohen, W. B., Healey, S. P., Yang, Z., Stehman, S. V., Brewer, C. K., Brooks, E. B., … Zhu, Z. (2017). How Similar Are Forest Disturbance Maps Derived from Different Landsat Time Series Algorithms? *Forests, Trees and Livelihoods*, *8*(4), 98.

Collins, B. M., Kelly, M., van Wagtendonk, J. W., & Stephens, S. L. (2007). Spatial patterns of large natural fires in Sierra Nevada wilderness areas. *Landscape Ecology*, *22*(4), 545–557.

Collins, B. M., Miller, J. D., Thode, A. E., Kelly, M., van Wagtendonk, J. W., & Stephens, S. L. (2009). Interactions Among Wildland Fires in a Long-Established Sierra Nevada Natural Fire Area. *Ecosystems* , *12*(1), 114–128.

Collins, B. M., & Stephens, S. L., & 2010. (2010). Stand-replacing patches within a "mixed severity"fire regime: quantitative characterization using recent fires in a long-established natural fire area. *Springer*. Retrieved from http://link.springer.com/article/10.1007/s10980-010-9470-5.

Connors, J. P., Lei, S., & Kelly, M. (2012). Citizen science in the age of neogeography: Utilizing volunteered geographic information for environmental monitoring. *Annals of the Association of American Geographers*, *102*(6), 1267-1289.

Cravens, A. E. (2014). Needs before Tools: Using Technology in Environmental Conflict Resolution. *Conflict Resolution Quarterly*, *32*(1), 3-32.

Cravens, A. E. (2016). Negotiation and Decision Making with Collaborative Software: How MarineMap 'Changed the Game' in California's Marine Life Protected Act Initiative. *Environmental management*, *57*(2), 474-497.

Cruz, D., Wieland, T., & Ziegler, A. (2006). Evaluation criteria for free/open source software products based on project analysis. *Software Process: Improvement and Practice*, *11*(2), 107-122.

Daniels, S. E., & Walker, G. B. (2001). Working through environmental conflict: The collaborative learning approach.

Dickinson, J. L., Zuckerberg, B., & Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annual review of ecology, evolution and systematics*, *41*, 149-72.

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T., & Purcell, K. (2012). The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, *10*(6), 291-297.

Dolan, K. A., Hurtt, G. C., Flanagan, S. A., Fisk, J. P., Sahajpal, R., Huang, C., … Masek, J. G. (2017). Disturbance Distance: quantifying forests' vulnerability to disturbance under current and future conditions. *Environmental Research Letters: ERL [Web Site]*, *12*(11), 114015.

Eftelioglu, E., Ali, R. Y., Tang, X., Xie, Y., Li, Y., & Shekhar, S. (2017). Geospatial Data Science: A Transdisciplinary Approach. In *Geospatial Data Science Techniques and Applications* (pp. 17-56). CRC Press.

Elwood, S. (2006). Beyond Cooptation or Resistance: Urban Spatial Politics, Community Organizations, and GIS-Based Spatial Narratives. *Annals of the association of American geographers*, *96*(2), 323-341.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the association of American geographers*, *102*(3), 571-590.

ESRI. https://support.esri.com/en/other-resources/gis-dictionary/term/uncertainty. Last accessed 04/20/2018.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293-314.

Frame, T. M., Gunton, T., & Day, J. C. (2004). The role of collaboration in environmental management: an evaluation of land and resource planning in British Columbia. *Journal of environmental planning and management*, *47*(1), 59-82.

Francis, J., Easterday, K., Scheckel, K., & Beissinger, S. R. (2017). The World Is a Park: Using Citizen Science to Engage People in Parks and Build the Next Century of Global Stewards. In *Science, Conservation, and National Parks* (pp. 275- 293). The University of Chicago Press.

French, N. H. F., Kasischke, E. S., Hall, R. J., Murphy, K. A., Verbyla, D. L., Hoy, E. E., & Allen, J. L. (2008). Using Landsat data to assess fire and burn severity in the North American boreal forest region: an overview and summary of results. *International Journal of Wildland Fire*, *17*(4), 443–462.

Gahegan, M., Luo, J., Weaver, S. D., Pike, W., & Banchuen, T. (2009). Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. *Computers & Geosciences*, *35*(4), 836-854.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Gonzalez, P., Battles, J. J., Collins, B. M., Robards, T., & Saah, D. S. (2015). Aboveground live carbon stock changes of California wildland ecosystems, 2001–2010. *Forest Ecology and Management*, *348*, 68–77.

Goodchild, M. F. (1992). Geographical information science. *International journal of geographical information systems*, *6*(1), 31-45.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110-120.

Goodman, K. J., Parker, S. M., Edmonds, J. W., & Zeglin, L. H. (2014). Expanding the scale of aquatic sciences: the role of the National Ecological Observatory Network (NEON). *Freshwater Science*, *34*(1), 377-385.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.06.031

Goward, S.N., C. Huang, F. Zhao, K. Schleeweis, K. Rishmawi, M. Lindsey, J.L. Dungan, & A. Michaelis. (2015). NACP NAFD Project: Forest Disturbance History from Landsat, 1986-2010. *ORNL DAAC, Oak Ridge, Tennessee, USA.* https://doi.org/10.3334/ORNLDAAC/1290

Gudex-Cross, D., Pontius, J., & Adams, A. (2017). Enhanced forest cover mapping using spectral unmixing and object-based classification of multi-temporal Landsat imagery. *Remote Sensing of Environment*, *196*, 193–204.

Gu, H., Williams, C. A., Ghimire, B., Zhao, F., & Huang, C. (2016). High-resolution mapping of time since disturbance and forest carbon flux from remote sensing and inventory data to assess harvest, fire, and beetle disturbance legacies in the Pacific Northwest. *Biogeosciences; Katlenburg-Lindau*, *13*(22), 6321–6337.

Haklay, M., Singleton, A., & Parker, C. (2008). Web mapping 2.0: The neogeography of the GeoWeb. *Geography Compass*, *2*(6), 2011-2039.

Haklay, M. (2013). Neogeography and the delusion of democratisation. *Environment and Planning A*, *45*(1), 55-69.

Hall, A. C. (2014). GI science, not GIScience. *Journal of Spatial Information Science*, *2014*(9), 129-131.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R. & Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160), 850-853.

Head, B.W. & Alford, J. (2015). Wicked Problems Implications for Public Policy and Management. *Administration & Society*, *47*(6), 711-739.

Healey, S. P., Cohen, W. B., Yang, Z., Kenneth Brewer, C., Brooks, E. B., Gorelick, N., … Zhu,

Z. (2017). Mapping forest change using stacked generalization: An ensemble approach. *Remote Sensing of Environment*. https://doi.org/10.1016/j.rse.2017.09.029

Hermosilla, T., Wulder, M. A., White, J. C., Coops, N. C., Hobart, G. W., & Campbell, L. B. (2016). Mass data processing of time series Landsat imagery: pixels to data products for forest monitoring. *International Journal of Digital Earth*, *9*(11), 1035–1054.

Heuschele, J., Ekvall, M. T., Mariani, P., & Lindemann, C. (2017). On the missing link in ecology: improving communication between modellers and experimentalists. *Oikos*, *126*(8), 1071-1077.

Huang, C., Goward, S. N., Masek, J. G., Thomas, N., Zhu, Z., & Vogelmann, J. E. (2010). An automated approach for reconstructing recent forest disturbance history using dense Landsat time series stacks. *Remote Sensing of Environment*, *114*(1), 183–198.

Hurteau, M. D., Westerling, A. L., Wiedinmyer, C., & Bryant, B. P. (2014). Projected effects of climate and development on California wildfire emissions through 2100. *Environmental Science & Technology*, *48*(4), 2298–2304.

Hyde, J., Strand, E. K., Hudak, A. T., & Hamilton, D. (2015). A CASE STUDY COMPARISON OF LANDFIRE FUEL LOADING AND EMISSIONS GENERATION ON A MIXED CONIFER FOREST IN NORTHERN IDAHO, USA. *Fire Ecology*, *11*(3). Retrieved from https://www.fs.fed.us/rm/pubs_journals/2015/rmrs_2015_hyde_j001.pdf

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485-488.

Jankowski, P. (2009). Towards participatory geographic information systems for community-based environmental decision making. *Journal of environmental management*, *90*(6), 1966-1971.

Jepson Flora Project (eds.) 2017. Jepson eFlora. **http://ucjeps.berkeley.edu/ef**. Last accessed October 23, 2017.

Jiang, B. & Thill, J.C. (2015). Volunteered Geographic Information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems*, *53*, 1-3.

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., & Xian, G. (2013). A comprehensive change detection method for updating the National Land Cover Database to circa 2011. *Remote Sensing of Environment*, *132*(Supplement C), 159–175.

Kalluri, S., Gundy, J., Haman, B., Paullin, A., Van Rompay, P., Vititoe, D., & Weiner, A. (2015). A High Performance Remote Sensing Product Generation System Based on a Service Oriented Architecture for the Next Generation of Geostationary Operational Environmental Satellites. *Remote Sensing*, *7*(8), 10385–10399.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., ... & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, *10*(4), 271-288.

Kang, S., & Lee, K. (2016). Auto-Scaling of Geo-Based Image Processing in an OpenStack Cloud Computing Environment. *Remote Sensing*, *8*(8), 662.

Keane, R. E., Rollins, M., & Zhu, Z.-L. (2007). Using simulated historical time series to prioritize fuel treatments on landscapes across the United States: The LANDFIRE prototype project. *Ecological Modelling*, *204*(3), 485–502.

Kelly, M., Ferranto, S., Lei, S., Ueda, K. I., & Huntsinger, L. (2012). Expanding the table: the web as a tool for participatory adaptive management in California forests. *Journal of environmental management*, *109*, 1-11.

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr—Temporal segmentation algorithms. *Elsevier Oceanography Series*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0034425710002245

Kolden, C. A., Lutz, J. A., Key, C. H., Kane, J. T., & van Wagtendonk, J. W. (2012). Mapped versus actual burned area within wildfire perimeters: Characterizing the unburned. *Forest Ecology and Management*, *286*(Supplement C), 38–47.

Kolden, C. A., Smith, A. M. S., & Abatzoglou, J. T. (2015). Limitations and utilisation of Monitoring Trends in Burn Severity products for assessing wildfire severity in the USA. *International Journal of Wildland Fire*, *24*(7), 1023–1028.

Kolden, C. A., & Weisberg, P. J. (2007). Assessing accuracy of manually-mapped wildfire perimeters in topographically dissected areas. *Fire Ecology*, *3*. Retrieved from http://www.fireecology.net/Journal/pdf/Volume03/Issue01/022.pdf

Krasnow, K. D., Fry, D. L., & Stephens, S. L. (2017). Spatial, temporal and latitudinal components of historical fire regimes in mixed conifer forests, California. *Journal of Biogeography*, *44*(6), 1239–1253.

Krasnow, K., Schoennagel, T., & Veblen, T. T. (2009). Forest fuel mapping and evaluation of LANDFIRE fuel maps in Boulder County, Colorado, USA. *Forest Ecology and Management*, *257*(7), 1603–1612.

Kumar, U., Ganguly, S., Nemani, R. R., Raja, K. S., Milesi, C., Sinha, R., … Gayaka, S. (2017). Exploring Subpixel Learning Algorithms for Estimating Global Land Cover Fractions from Satellite Data Using High Performance Computing. *Remote Sensing*, *9*(11), 1105.

LANDFIRE Disturbance Metadata. https://landfire.cr.usgs.gov/distmeta/servlet/gov.usgs.edc.MetaBuilder?TYPE=HTML&DATASET=FA9. Last accessed 01/06/2017.

Lee, C. A., Gasster, S. D., Plaza, A., Chang, C. I., & Huang, B. (2011). Recent Developments in High Performance Computing for Remote Sensing: A Review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *4*(3), 508–527.

Lentile, L. B., Holden, Z. A., Smith, A. M. S., Falkowski, M. J., Hudak, A. T., Morgan, P., … Benson, N. C. (2006). Remote sensing techniques to assess active fire characteristics and post-fire effects. *International Journal of Wildland Fire*, *15*(3), 319.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ... & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119-133.

Liu, D., & Xia, F. (2010). Assessing object-based classification: advantages and limitations. *Remote Sensing Letters* , *1*(4), 187–194.

Liu, J., Vogelmann, J. E., Zhu, Z., Key, C. H., Sleeter, B. M., Price, D. T., … Jiang, H. (2011). Estimating California ecosystem carbon change using process model and land cover disturbance data: 1951–2000. *Ecological Modelling*, *222*(14), 2333–2341.

Leonard, L., & Duffy, C. J. (2014). Automating data-model workflows at a level 12 HUC scale: Watershed modeling in a distributed computing environment. *Environmental Modelling & Software*, *61*, 174-190.

MacEachren, A. M., & Brewer, I. (2004). Developing a conceptual framework for visually-enabled geocollaboration. *International Journal of Geographical Information Science*, *18*(1), 1-34.

MacEachren, A. M., Pike, W., Yu, C., Brewer, I., Gahegan, M., Weaver, S. D., & Yarnal, B. (2006). Building a geocollaboratory: supporting Human–Environment Regional Observatory (HERO) collaborative science activities. *Computers, Environment and Urban Systems*, *30*(2), 201-225.

Malczewski, J. (2006). GIS-based multicriteria decision analysis: a survey of the literature. *International Journal of Geographical Information Science*, *20*(7), 703-726.

Masek, J. G., Goward, S. N., Kennedy, R. E., Cohen, W. B., Moisen, G. G., Schleeweis, K., & Huang, C. (2013). United States Forest Disturbance Trends Observed Using Landsat Time Series. *Ecosystems* , *16*(6), 1087–1104.

McKerrow, A., Dewitz, J., Long, D. G., Nelson, K., & Connot, J. A. (2016). A comparison of NLCD 2011 and LANDFIRE EVT 2010: Regional and national summaries. Retrieved from https://pubs.er.usgs.gov/publication/70177839

Meddens, A. J. H., Kolden, C. A., & Lutz, J. A. (2016). Detecting unburned areas within wildfire perimeters using Landsat and ancillary data across the northwestern United States. *Remote Sensing of Environment*, *186*, 275–285.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449-461.

Miller, J. D., Safford, H. D., Crimmins, M., & Thode, A. E. (2009). Quantitative Evidence for Increasing Forest Fire Severity in the Sierra Nevada and Southern Cascade Mountains, California and Nevada, USA. *Ecosystems* , *12*(1), 16–32.

Morin, A., Urban, J., Adams, P. D., Foster, I., Sali, A., Baker, D., & Sliz, P. (2012). Shining light into black boxes. *Science*, *336*(6078), 159-160.
Moritz, M. A., & Stephens, S. L. (2008). Fire and sustainability: considerations for California's altered future climate. *Climatic Change*. Retrieved from http://www.springerlink.com/index/5411702235mx5432.pdf

Nagy, D., Yassin, A. M., & Bhattacherjee, A. (2010). Organizational adoption of open source software: barriers and remedies. *Communications of the ACM*, *53*(3), 148-151.

Nemani, R., Votava, P., Michaelis, A., Melton, F., & Milesi, C. (2011). Collaborative Supercomputing for Global Change Science. *Eos, Transactions American Geophysical Union*, *92*(13), 109–110.

Neset, T. S., Opach, T., Lion, P., Lilja, A., & Johansson, J. (2016). Map-Based Web Tools Supporting Climate Change Adaptation. *The Professional Geographer*, *68*(1), 103-114.

National Geographic. (2016). The Great Nature Project. http://nationalgeographic.org/projects/great-nature-project/. Accessed on October 26, 2016.

Nogueira, J. M. P., Ruffault, J., Chuvieco, E., & Mouillot, F. (2016). Can We Go Beyond Burned Area in the Assessment of Global Remote Sensing Products with Fire Patch Metrics? *Remote Sensing*, *9*(1), 7.

Palomino, J., Muellerklein, O. C., & Kelly, M. (2017). A review of the emergent ecosystem of collaborative geospatial tools for addressing environmental challenges. *Computers, Environment and Urban Systems*, *65*(Supplement C), 79–92.

Parker, B. M., Lewis, T., & Srivastava, S. K. (2015). Estimation and evaluation of multi-decadal fire severity patterns using Landsat sensors. *Remote Sensing of Environment*, *170*(Supplement C), 340–349.

Parks, S. A., Miller, C., Nelson, C. R., & Holden, Z. A. (2014). Previous Fires Moderate Burn Severity of Subsequent Wildland Fires in Two Large Western US Wilderness Areas. *Ecosystems* , *17*(1), 29–42.

Pedersen, B., Kearns, F., & Kelly, M. (2007). Methods for facilitating web-based participatory research informatics. *Ecological Informatics*, *2*(1), 33-42.

Plaza, A. J., & Chang, C.I. (2007). *High Performance Computing in Remote Sensing*. Chapman & Hall/CRC.

Quinn, S. (2015). Using small cities to understand the crowd behind OpenStreetMap. *GeoJournal*, 1-19.

Reeves, M. C., Ryan, K. C., Rollins, M. G., & Thompson, T. G. (2009). Spatial fuel data products of the LANDFIRE Project. *International Journal of Wildland Fire*, *18*(3), 250–267.

Rinner, C. (2003). Web-based spatial decision support: status and research directions. *Journal of Geographic Information and Decision Analysis*, *7*(1), 14-31.
Roberts, J. J., Best, B. D., Dunn, D. C., Treml, E. A., & Halpin, P. N. (2010). Marine Geospatial Ecology Tools: An integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modelling & Software*, *25*(10), 1197-1207.

Rocchini, D., & Neteler, M. (2012). Let the four freedoms paradigm apply to ecology. *Trends in ecology & evolution*, *27*(6), 310-311.

Rollins, M. G. (2009). LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment. *International Journal of Wildland Fire*, *18*(3), 235–249.

Roy, D. P., Boschetti, L., & Trigg, S. N. (2006). Remote sensing of fire severity: assessing the performance of the normalized burn ratio. *IEEE Geoscience and Remote Sensing Letters*, *3*(1), 112–116.

Ryan, K. C., & Opperman, T. S. (2013). LANDFIRE – A national vegetation/fuels data base for use in fuels treatment, restoration, and suppression planning. *Forest Ecology and Management*, *294*(Supplement C), 208–216.

Saveri, A., Rheingold, H., & Vian, K. (2005). Technologies of cooperation. *Institute for the Future. Retrieved on July*, *9*, 2005.

Schroeder, T. A., Schleeweis, K. G., Moisen, G. G., Toney, C., Cohen, W. B., Freeman, E. A., … Huang, C. (2017). Testing a Landsat-based approach for mapping disturbance causality in U.S. forests. *Remote Sensing of Environment*, *195*, 230–243.

Schutt, R. & O'Neil, C. (2014). Doing Data Science. Sebastopol, O'Reilly.

Seasketch. (2016). http://www.seasketch.org. Accessed on October 26, 2016.

Selin, S., & Chevez, D. (1995). Developing a collaborative model for environmental planning and management. *Environmental management*, *19*(2), 189-195.

Sieber, R. E. (2000).Conforming (to) the opposition: the social construction of geographical information systems in social movements. *International Journal of Geographical Information Science*, 14(8), 775-793.

Sieber, R. (2006). Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers*, *96*(3), 491-507.

Sleeter, B. M., Liu, J., Daniel, C., Rayfield, B., Sherba, J., Hawbaker, T. J., … Loveland, T. R. (2018). Effects of contemporary land-use and land-cover change on the carbon balance of terrestrial ecosystems in the United States. *Environmental Research Letters: ERL [Web Site]*, *13*(4), 045006.

Stallman, R. (1985). The GNU manifesto.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., & Andrienko, N. (2011, November). Analytical workflow of monitoring human mobility in big event settings using bluetooth. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on indoor spatial awareness* (pp. 51-58). ACM.

Steiniger, S., & Hay, G. J. (2009). Free and open source geographic information tools for landscape ecology. *Ecological Informatics*, *4*(4), 183-195.

Steiniger, S., & Weibel, R. (2009). GIS software–a description in 1000 words. *Encyclopeadia of Geography*, 1-2.

Steiniger, S., & Hunter, A. J. (2013). The 2012 free and open source GIS software map–A guide to facilitate research, development, and adoption. *Computers, Environment and Urban Systems*, *39*, 136-150.

Soulard, C. E., Acevedo, W., Cohen, W. B., Yang, Z., Stehman, S. V., & Taylor, J. L. (2017). Harmonization of forest disturbance datasets of the conterminous USA from 1986 to 2011. *Environmental Monitoring and Assessment*, *189*(4), 170.

Soulard, C. E., Albano, C. M., Villarreal, M. L., & Walker, J. J. (2016). Continuous 1985–2012 Landsat Monitoring to Assess Fire Effects on Meadows in Yosemite National Park, California. *Remote Sensing*, *8*(5), 371.

Stephens, S. L., Martin, R. E., & Clinton, N. E. (2007). Prehistoric fire area and emissions from California's forests, woodlands, shrublands, and grasslands. *Forest Ecology and Management*, *251*(3), 205–216.

Stevens, J. T., Collins, B. M., Miller, J. D., North, M. P., & Stephens, S. L. (2017). Changing spatial patterns of stand-replacing fire in California conifer forests. *Forest Ecology and Management*, *406*(Supplement C), 28–36.

Sui, D. (2014). Opportunities and impediments for open GIS. *Transactions in GIS*, *18*(1), 1-24.

Temby, O., Sandall, J., Cooksey, R., & Hickey, G. M. (2016). How do civil servants view the importance of collaboration and scientific knowledge for climate change adaptation?. *Australasian Journal of Environmental Management*, *23*(1), 5-20.

Thomas, N. E., Huang, C., Goward, S. N., Powell, S., Rishmawi, K., Schleeweis, K., & Hinds, A. (2011). Validation of North American Forest Disturbance dynamics derived from Landsat time series stacks. *Remote Sensing of Environment*, *115*(1), 19–32.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423.

Touchon, J. C., & McCoy, M. W. (2016). The mismatch between current statistical practice and doctoral training in ecology. *Ecosphere*, *7*(8).

Tyukavina, A., Baccini, A., Hansen, M. C., Potapov, P. V., Stehman, S. V., Houghton, R. A., … Goetz, S. J. (2015). Aboveground carbon loss in natural and managed tropical forests from 2000 to 2012. *Environmental Research Letters: ERL [Web Site]*, *10*(7), 074002.

USGS Earth Resources Observation and Science (EROS) User Services. Email correspondence on October 13, 2016.

Vanderhoof, M. K., Fairaux, N., Beal, Y.-J. G., & Hawbaker, T. J. (2017). Validation of the USGS Landsat Burned Area Essential Climate Variable (BAECV) across the conterminous United States. *Remote Sensing of Environment*, *198*, 393–406.

Voss, A., Denisovich, I., Gatalsky, P., Gavouchidis, K., Klotz, A., Roeder, S., & Voss, H. (2004). Evolution of a participatory GIS. *Computers, Environment and Urban Systems*, *28*(6), 635-651.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., & Nyerges, T. L. (2013). CyberGIS software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science*, *27*(11), 2122-2145.

Wang, S. (2016). CyberGIS and Spatial Data Science. *GeoJournal*. Online issue. doi:10.1007/s10708-016-9740-0

Watson, S. J., Luck, G. W., Spooner, P. G., & Watson, D. M. (2014). Land-use change: incorporating the frequency, sequence, time span, and magnitude of changes into ecological research. *Frontiers in Ecology and the Environment*, *12*(4), 241-249.

White, R. L., Sutton, A. E., Salguero-Gomez, R., Bray, T. C., Campbell, H., Cieraad, E., Geekiyanage, N., Gherardi, L., Hughes, A.C., Jørgensen, P.S. & Poisot, T. (2015). The next generation of action ecology: novel approaches towards global ecological research. *Ecosphere*, *6*(8), 1-16.

Williams, C. A., Gu, H., MacLean, R., Masek, J. G., & Collatz, G. J. (2016). Disturbance and the carbon balance of US forests: A quantitative review of impacts from harvests, fires, insects, and droughts. *Global and Planetary Change*, *143*, 66–80.

Wolkovich, E. M., Cook, B. I., McLauchlan, K. K., & Davies, T. J. (2014). Temporal ecology in the Anthropocene. *Ecology letters*, *17*(11), 1365-1379.

Woodall, C. W., Walters, B. F., Russell, M. B., Coulston, J. W., Domke, G. M., D'Amato, A.

W., & Sowers, P. A. (2016). A Tale of Two Forest Carbon Assessments in the Eastern United States: Forest Use Versus Cover as a Metric of Change. *Ecosystems , 19*(8), 1401–1417.

Wright, D. J., Duncan, S. L., & Lach, D. (2009). Social power and GIS technology: a review and assessment of approaches for natural resource management. *Annals of the Association of American Geographers, 99*(2), 254-272.

Wright, D. J., & Wang, S. (2011). The emergence of spatial cyberinfrastructure. *Proceedings of the National Academy of Sciences, 108*(14), 5488-5491.

Wulf, W.A. (1993). The collaboratory opportunity. *Science*, 261, 854–855.

Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., … Fay, D. (2011). Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? *International Journal of Digital Earth, 4*(4), 305–329.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: past, present and future. *Computers, Environment and Urban Systems, 34*(4), 264-277.

Yu, Q., Gong, P., Clinton, N., Biging, G., Kelly, M., & Schirokauer, D. (2006). Object-based Detailed Vegetation Classification with Airborne High Spatial Resolution Remote Sensing Imagery. *Photogrammetric Engineering & Remote Sensing*, *72*(7), 799–811.

Yuan, M. (2016). 30 years of IJGIS: the changing landscape of geographical information science and the road ahead. *International Journal of Geographical Information Science*, 1-10.

Zhao, F., Huang, C., Goward, S. N., Schleeweis, K., Rishmawi, K., Lindsey, M. A., … Michaelis, A. (2018). Development of Landsat-based annual US forest disturbance history maps (1986–2010) in support of the North American Carbon Program (NACP). *Remote Sensing of Environment, 209*, 312–326.

Zhao, H., Salloum, S., Cai, Y., & Huang, J. Z. (2015). Ensemble subspace clustering of text data using two-level features. *International Journal of Machine Learning and Cybernetics*, 1-16.

Zhu, Z. (2017). Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing: Official Publication of the International Society for Photogrammetry and Remote Sensing , 130*, 370–384.

Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment, 144*(Supplement C), 152–171.

Zimmerman, P. L., Housman, I. W., Perry, C. H., Chastain, R. A., Webb, J. B., & Finco, M. V. (2013). An accuracy assessment of forest disturbance mapping in the western Great Lakes. *Remote Sensing of Environment, 128*, 176–185.

Appendices

Appendix 2-1. Scoring Rubric

Feature Name (metric)	Group (Subset)	Description of Group (Subset)	Score of 1	Score of 2	Score of 3
mobile support	1	Setting Up the Working Environment	none - no mobile functionality (i.e. no application or cannot display UI on phone web browser)	some - no specific mobile functionality or application provided but could be developed by user or used on phone web browser	full - mobile application (i.e. Google Play or App store) or SDK available
reproducibility of working environment	1	Setting Up the Working Environment	none - platform does not reproduce working environment	some - platform supports reproducibility of working environment via shared code or custom application installs but not required; or user interface is not the only primary working environment	full - platform automatically reproduces working environment (i.e. working environment the same regardless of how users access user interface); or user interface is the primary working environment
scalability	1	Setting Up the Working Environment	none - platform is limited by tool provider (i.e. functionality or data limits, number of users, applications/operating systems supported)	some - platform is somewhat limited by tool provider but not completely (i.e. data limits, number of users, applications/operating systems supported); or there are not many limitations but the platform is exclusively hosted on cloud by provider	full - platform not limited by tool provider (i.e. data limits, number of users, applications supported) or user can deploy their own cloud services (even if for a cost)

shared working environment	1	Setting Up the Working Environment	none - platform is either local install only or does not have this capability	some - platform can provide a shared working environment if desired but it is not necessary for full tool functionality	full - platform requires users to use the same working environment in order to work on the same data jointly (i.e. the same UI within a same session) or there is only one dataset available for all users to work on (e.g. OpenStreetMap)
web-based	1	Setting Up the Working Environment	none - platform relies on either local (Desktop or Mobile) or server based install	some - platform has web-based or cloud-based component (i.e. there are both desktop/mobile and web-based/cloud-based options available or needed)	full - platform is completely web-based or cloud-based (i.e. no installs needed or available)
"live" data dissemination and database dependence	2	Data Wrangling	none - "live" data not supported	some - "live" data supported but not required	full - "live" data is required for full tool functionality
raster - create	2	Data Wrangling	none - user cannot create raster data	some - user can create traditional raster formats (i.e. GeoTIFF)	full - user can create non-traditional and open data raster formats and services (i.e. NetCDF, HDF5, Tile Mapping Services)
raster - edit	2	Data Wrangling	none - user cannot edit raster data	some - user can edit traditional raster formats (i.e. GeoTIFF)	full - user can edit non-traditional and open data raster formats (i.e. NetCDF, HDF5, Tile Mapping Services)

raster - upload	2	Data Wrangling	none - user cannot upload raster data	some - user can upload traditional raster formats (i.e. GeoTIFF)	full - user can upload or consume non-traditional and open data raster formats (i.e. NetCDF, HDF5, Tile Mapping Services)
vector - create	2	Data Wrangling	none - user cannot create vector data	some - user can create traditional vector formats (i.e. SHP, KML)	full - user can create non-traditional and open data vector formats and services (i.e. GeoJSON, WMS/WFS, Vector Tiles, MBTiles)
vector - edit	2	Data Wrangling	none - user cannot edit vector data	some - user can edit traditional vector formats (i.e. SHP, KML)	full - user can edit non-traditional and open data vector formats and services (i.e. GeoJSON, WMS/WFS, Vector Tiles, MBTiles)
vector - upload	2	Data Wrangling	none - user cannot upload vector data (i.e. formats beyond CSV)	some - user can upload traditional vector formats (i.e. SHP, KML)	full - user can upload or consume non-traditional and open data vector formats and services (i.e. GeoJSON, WMS/WFS, Vector Tiles, MBTiles)
spatial analysis	3	Data Analysis	none - platform does not focus on data analysis	some - platform provides some basic tools for analyzing data (i.e. geoprocessing, spatial filtering queries, creating new visualizations like heat maps)	full - platform provides advanced tools for analyzing data (i.e. statistical functions)

integration of time	3	Data Analysis	none - time is not an integral component of the platform	some - platform provides some support for time data (i.e. collecting data over time, querying for or visualizing data over time)	full - platform supports analysis over time (i.e. calculating change over time, statistical functions)
scripting capabilities	3	Data Analysis	none - platform does not provide options for scripting (i.e. not possible to integrate Python, R, C, JavaScript code)	some - platform provides some functionality for integrating scripting for some spatial analysis (i.e. Python or R for geoprocessing); or API available for creating new analysis applications for analyzing data	full - platform provide options for advanced scripting such as existing or custom statistical functions (i.e. Python, R, C, JavaScript)
integration of high performance computing	3	Data Analysis	none - platform is not able to or does not support distributed/parallel computing (i.e. user cannot or would not benefit much from implementing this functionality)	some - platform could support distributed/parallel computing but does not currently provide a mechanism/tool to the user for this functionality	full - platform supports or incorporates distributed/parallel computing
reproducibility of workflow (analysis)	3	Data Analysis	none - platform does not provide tools for reproducing workflow; or platform is not focused on analyzing data	some - platform supports reproducibility of workflow via shared code but not required	full - platform automatically reproduces workflow (i.e. workflow is a saved product of any session)
custom visualization options	4	Data Visualization and Publication	none - user can only create standardized maps	some - user can modify existing code for map or results (i.e. CSS, HTML)	full - user can incorporate advanced visualization tools (i.e. D3, Leaflet, other JavaScript libraries) or create custom styles
data publication options	4	Data Visualization and Publication	none - user can only download maps or calculated results	some - spatial data are downloadable	full - API available to access data

raster - download	4	Data Visualization and Publication	none - user cannot download raster data	some - user can download traditional raster formats (i.e., GeoTIFF); for mobile, ability to use "big data" formats and services offline	full - user can download non-traditional and open data raster formats and services (i.e. NetCDF, HDF5, Tile Mapping Services) to be consumed in other applications
vector - download	4	Data Visualization and Publication	none - user cannot download vector data	some - user can download traditional vector formats (i.e., SHP, KML); for mobile, ability to use "big data" formats and services offline	full - user can download non-traditional and open data vector formats and services (i.e. GeoJSON, WMS/WFS, Vector Tiles, MBTiles) to be consumed in other applications
free of cost	5	FOSS4G and Accessibility	none - payment required (i.e. free is limited to trial basis only)	some - free to a certain level (i.e. data or functionality limits) or free but restricted to a limited set of users for full functionality (i.e. tool is open but API is restricted)	full - completely cost-free for full functionality and open to anyone
interoperability	5	FOSS4G and Accessibility	none - platform does not take inputs from or provide outputs to other platforms (i.e. only imports/exports to SHP or KML)	some - platform can import from or provide output to a limited set of platforms (at least one) (i.e. ArcGIS online, database management systems, WMS/WFS)	full - platform provides tools open to consuming/being consumed by a wide variety of platforms via API, SDK, or SOAP web services; or connects to at least 3 tools/kernels (i.e. Jupyter kernels for R, Python, MatLab)

open source integration (back-end)	5	FOSS4G and Accessibility	none - proprietary platform	some - platform relies on both proprietary and open source components	full - platform is composed entirely of open source components
privacy and access control	5	FOSS4G and Accessibility	none - all data, analysis and projects are public by default	some - platform allows for some restriction to data, analysis workflow, and projects but does not include a user/content management system; or privacy can be provided for a cost; or it is private by default because it is a local install or data is used locally	full - platform includes a user/content management system that allows full control over access to data, analysis workflow, and projects (even if paid)
tools for asynchronous tasks	6	User Involvement	none - platform does not provide tools to support asynchronous tasks (i.e. sharing code, versioning, off-line capabilities, sharing of status updates)	some - platform provides some tools (at least 1) to support asynchronous tasks (i.e. sharing code, versioning, off-line capabilities, sharing of status updates)	full - platform provides many tools to support asynchronous tasks (i.e. sharing code, versioning, off-line capabilities, sharing of status updates)
open source modifications by users	6	User Involvement	none - user cannot add or modify existing open source functionality (i.e. platform is locked down)	some - user can add open source functionality (i.e. add D3 code, use of API) but not modify existing open source functionality (i.e. modify existing functions)	full - released under an open source license (e.g. GNU GPL); or user can add to and modify existing open source functionality (full source code available)
user involvement in technology building	6	User Involvement	none - users are not involved in creating or expanding the platform	some - users can expand or build off of the platform (i.e. SDK provided, ability to build custom forms, leverage WFS, script new tools/processes)	full - users are active participants in creating and expanding the platform (i.e. collaborative coding or customization)

user knowledge needed for modification of technology	6	User Involvement	none - users do not need much technical knowledge to expand the platform (i.e. no programming) but there are limited or no modifications possible	some - users need some technical knowledge to expand the platform (i.e. programming knowledge is useful), allowing for more modifications, such as form building	full - users need standard technical knowledge to expand the platform (i.e. programming knowledge is required), allowing for advanced modification options such as integrating custom functions

Appendix 2-2. Feature Scores by Tool

Label	Name	mobile support	reproducibility of working environment	scalability	shared working environment	web-based
T30	AmigoCloud	3	3	2	2	2
T28	ArcGIS Online	2	3	2	2	3
T31	ArcGIS Open Data	2	2	2	2	3
T1	CARTO	3	2	3	2	2
T23	Data Basin	2	2	2	2	3
T20	eBird	2	3	2	2	3
T24	ESRI Collector for ArcGIS	3	2	2	2	2
T16	Field Papers	2	3	2	3	3
T12	FME Suite	2	3	3	2	2
T21	GeoLocate	2	2	3	2	2
T25	Geopaparazzi	3	2	1	2	1
T9	Global Forest Watch	2	2	3	2	3
T13	Google Earth Engine	2	3	2	2	3
T22	HOLOS	2	2	3	2	3
T17	iNaturalist	3	3	3	2	2

T4	Jupyter Hub	2	3	3	2	2
T26	Locus Map	3	2	1	1	1
T14	Madrona	2	3	3	2	2
T15	MapBox Studio	3	3	3	2	2
T2	MapGuide	2	2	3	2	2
T5	NASA NEX sandbox	1	2	2	2	2
T10	NextGIS	3	2	2	2	2
T18	OpenDataKit/ GeoODK	3	2	1	2	2
T19	OpenStreetMap	2	3	2	3	3
T27	Orux Maps	3	1	1	1	1
T6	OS Geo Live	1	3	3	2	2
T11	QGIS Cloud	2	2	3	2	2
T7	ROpenSci	2	2	3	2	2
T8	Rshiny	2	2	3	2	2
T29	Seasketch	2	3	2	2	3
T3	XchangeCore	2	2	3	2	2
	Average Tool Score	2.258	2.387	2.355	2	2.226

Label	**Name**	**"live" data dissemination and database dependence**	**raster - create**	**raster - edit**	**raster - upload**	**vector - create**
T30	AmigoCloud	3	1	1	2	3
T28	ArcGIS Online	3	1	1	3	2
T31	ArcGIS Open Data	3	1	1	3	1
T1	CARTO	3	1	1	1	2
T23	Data Basin	2	1	1	3	2

T20	eBird	3	1	1	1	1
T24	ESRI Collector for ArcGIS	2	1	1	3	2
T16	Field Papers	2	1	1	1	2
T12	FME Suite	2	3	3	3	3
T21	GeoLocate	2	1	1	1	3
T25	Geopaparazzi	2	1	1	3	2
T9	Global Forest Watch	2	2	1	3	3
T13	Google Earth Engine	2	2	1	2	3
T22	HOLOS	3	1	1	1	1
T17	iNaturalist	3	1	1	1	3
T4	Jupyter Hub	2	3	3	3	3
T26	Locus Map	2	1	1	3	2
T14	Madrona	3	1	1	2	3
T15	MapBox Studio	3	1	1	3	3
T2	MapGuide	3	1	1	3	2
T5	NASA NEX sandbox	2	3	3	3	3
T10	NextGIS	2	2	2	3	3
T18	OpenDataKit/ GeoODK	2	1	1	1	3
T19	OpenStreetMap	3	1	1	1	2
T27	Orux Maps	2	1	1	3	2
T6	OS Geo Live	2	3	3	3	3
T11	QGIS Cloud	2	2	2	3	3
T7	ROpenSci	2	3	2	2	3
T8	Rshiny	2	3	2	3	3
T29	Seasketch	3	1	1	3	2

T3	XchangeCore	2	1	1	2	2
	Average Tool Score	2.387	1.516	1.387	2.323	2.419

Label	Name	vector - edit	vector - upload	spatial analysis	integration of time	scripting capabilities
T30	AmigoCloud	2	3	2	2	2
T28	ArcGIS Online	2	3	2	1	2
T31	ArcGIS Open Data	1	3	2	2	2
T1	CARTO	2	2	2	2	2
T23	Data Basin	2	3	2	2	1
T20	eBird	1	1	2	2	1
T24	ESRI Collector for ArcGIS	2	3	1	2	1
T16	Field Papers	2	1	1	2	1
T12	FME Suite	3	3	2	1	2
T21	GeoLocate	1	1	2	1	1
T25	Geopaparazzi	2	3	1	2	1
T9	Global Forest Watch	2	3	3	3	3
T13	Google Earth Engine	1	2	3	3	3
T22	HOLOS	1	1	2	2	2
T17	iNaturalist	1	1	1	2	1
T4	Jupyter Hub	3	3	3	3	3
T26	Locus Map	2	3	2	2	1
T14	Madrona	3	3	3	2	3
T15	MapBox Studio	3	3	3	2	3
T2	MapGuide	2	3	2	2	2
T5	NASA NEX sandbox	2	3	3	3	3

T10	NextGIS	3	3	2	2	3
T18	OpenDataKit/ GeoODK	3	3	1	2	2
T19	OpenStreetMap	2	1	1	2	1
T27	Orux Maps	2	3	1	2	1
T6	OS Geo Live	3	3	3	3	3
T11	QGIS Cloud	3	3	3	2	2
T7	ROpenSci	3	3	3	3	3
T8	Rshiny	3	3	3	3	3
T29	Seasketch	2	3	3	3	3
T3	XchangeCore	2	3	2	2	2
	Average Tool Score	2.129	2.549	2.129	2.161	2.032

Label	**Name**	**integration of HPC**	**reproducibility of workflow (analysis)**	**custom visualization options**	**data publication options**	**raster – download**
T30	AmigoCloud	1	2	2	3	1
T28	ArcGIS Online	1	2	2	2	2
T31	ArcGIS Open Data	1	1	2	3	1
T1	CARTO	3	2	3	3	1
T23	Data Basin	1	2	1	2	2
T20	eBird	1	1	1	1	1
T24	ESRI Collector for ArcGIS	1	2	1	1	2
T16	Field Papers	1	2	1	2	1
T12	FME Suite	3	3	1	3	3
T21	GeoLocate	1	2	1	3	1
T25	Geopaparazzi	1	1	1	2	1

T9	Global Forest Watch	3	3	3	3	2
T13	Google Earth Engine	3	3	3	3	2
T22	HOLOS	2	2	3	3	1
T17	iNaturalist	1	1	1	3	1
T4	Jupyter Hub	2	3	3	2	3
T26	Locus Map	1	1	1	3	1
T14	Madrona	2	2	3	3	1
T15	MapBox Studio	2	2	3	3	3
T2	MapGuide	1	2	3	3	2
T5	NASA NEX sandbox	3	2	3	2	3
T10	NextGIS	2	1	2	3	2
T18	OpenDataKit/ GeoODK	1	2	1	3	1
T19	OpenStreet Map	2	2	1	3	1
T27	Orux Maps	1	1	1	2	1
T6	OS Geo Live	2	2	3	2	3
T11	QGIS Cloud	2	2	1	2	2
T7	ROpenSci	2	2	3	3	2
T8	Rshiny	2	2	3	2	2
T29	Seasketch	2	2	1	2	1
T3	XchangeCore	2	2	1	2	2
	Average Tool Score	1.7097	1.903	1.903	2.484	1.677

Label	Name	vector - download	free	interoperability	open source integration	privacy and access control
T30	AmigoCloud	3	1	3	2	3
T28	ArcGIS Online	3	2	2	2	3

T31	ArcGIS Open Data	3	2	3	2	3
T1	CARTO	3	2	3	3	3
T23	Data Basin	2	2	2	2	3
T20	eBird	3	2	3	2	1
T24	ESRI Collector for ArcGIS	2	1	2	2	3
T16	Field Papers	3	3	3	3	2
T12	FME Suite	3	1	3	2	3
T21	GeoLocate	3	3	3	2	3
T25	Geopaparazzi	2	3	2	3	2
T9	Global Forest Watch	3	3	3	3	2
T13	Google Earth Engine	3	2	3	3	3
T22	HOLOS	2	3	3	3	1
T17	iNaturalist	3	3	3	3	1
T4	Jupyter Hub	3	3	3	3	3
T26	Locus Map	3	2	3	2	2
T14	Madrona	3	3	3	3	3
T15	MapBox Studio	3	2	3	3	2
T2	MapGuide	3	3	2	3	3
T5	NASA NEX sandbox	3	2	2	2	2
T10	NextGIS	3	2	3	3	3
T18	OpenDataKit/ GeoODK	3	3	3	3	2
T19	OpenStreet Map	3	3	3	3	1
T27	Orux Maps	2	3	2	2	2
T6	OS Geo Live	3	3	3	3	2
T11	QGIS Cloud	3	2	2	3	2

T7	ROpenSci	3	3	3	3	1
T8	Rshiny	3	2	2	3	2
T29	Seasketch	2	1	2	2	3
T3	XchangeCore	3	2	3	3	3
	Average Tool Score	2.806	2.323	2.677	2.613	2.323

Label	**Name**	**tools for asynchronous tasks**	**open source modification by users**	**user involvement in technology building**	**user knowledge needed for modification of technology**
T30	AmigoCloud	2	2	2	3
T28	ArcGIS Online	2	1	2	2
T31	ArcGIS Open Data	2	2	2	2
T1	CARTO	2	3	2	3
T23	Data Basin	3	1	1	1
T20	eBird	1	1	1	1
T24	ESRI Collector for ArcGIS	3	1	1	1
T16	Field Papers	3	3	2	3
T12	FME Suite	3	2	2	3
T21	GeoLocate	2	2	2	2
T25	Geopaparazzi	2	2	2	2
T9	Global Forest Watch	2	3	2	3
T13	Google Earth Engine	3	2	3	3
T22	HOLOS	2	2	2	3
T17	iNaturalist	2	3	2	3

T4	Jupyter Hub	2	3	2	3
T26	Locus Map	2	2	2	3
T14	Madrona	3	3	3	3
T15	MapBox Studio	2	2	2	3
T2	MapGuide	2	3	2	3
T5	NASA NEX sandbox	2	2	3	3
T10	NextGIS	2	3	2	3
T18	OpenDataKit/ GeoODK	2	3	3	3
T19	OpenStreet Map	3	2	3	3
T27	Orux Maps	2	1	1	1
T6	OS Geo Live	2	3	3	3
T11	QGIS Cloud	2	2	2	2
T7	ROpenSci	2	3	3	3
T8	Rshiny	2	3	2	3
T29	Seasketch	3	2	3	2
T3	XchangeCore	3	3	2	3
	Average Tool Score	2.258	2.258	2.129	2.548

Appendix 3-1. Supplemental Material to Chapter 3

All datasets used in this study (including the disturbance products, reference data on fire, and environmental conditions) were used or calculated at a 30 m spatial resolution (resampled, if needed) and were clipped to the boundary of California using a vector file publicly provided by ESRI. All pre-processing was completed in Google Earth Engine (EE), unless otherwise noted. Once uploaded to EE, all data were projected and stored as WGS84 Web Mercator (EPSG 3857).

Standardization of Vegetation Disturbance Products

The standardization process was adjusted for the temporal structure of the individual disturbance products. Specifically, LANDFIRE was only provided as individual annual rasters, while GFC was only provided as a single raster with distinct values for each year. As NAFD was available

in both temporal formats, I decided to use the single raster NAFD provided for last year of disturbance to easily query the data by year. To create the annual rasters for GFC and NAFD, I queried the original rasters by year (2001 to 2010) and created new annual rasters, in which all pixels reported as disturbed in a given year were labeled with a value of 1. Undisturbed pixels in that year were labeled with a value of 0. To create comparable aggregated-time rasters of reported disturbance for GFC and NAFD, I relabeled the original rasters, so that all pixels reported as disturbed any time between 2001 and 2010 were given a value of 1, while undisturbed pixels or pixels disturbed outside of time frame (i.e. before 2001 or after 2010) were labeled with a value of 0. Similarly, to create comparable annual and aggregated-time rasters for LANDFIRE, I relabeled the original annual rasters such that all pixels reported as disturbed in a given year were given a value of 1 (all others given a value of 0), and then aggregated the annual rasters to create an aggregated-time raster in which all pixels reported as disturbed anytime between 2001 and 2010 were labeled with a value of 1 (all others given a value of 0).

For GFC, the integrated-time raster of forest loss was available in Google Earth Engine as part of the raster image stack named Hansen Global Forest Change v1.2 (2000-2014). A layer called Lossyear (band 3) provided a value at each pixel representing the year in which loss was identified in the LTS, beginning with a value of 1 for 2001 to a value of 14 for 2014. A value of 0 meant that no loss was identified between 2001 and 2014. I limited Lossyear to values greater than or equal 1 (representing 2001) and for values less than or equal to 10 (representing 2010). The individual pixel values were then queried to create rasters by year (i.e. the raster for year 2001 was created by querying for pixels with a value 1, and so on). In my aggregated-time raster, pixels with values between 1 and 14 were set to a value of 1 to indicate identified disturbance in my study period, while all other pixels were given a value of 0.

For NAFD, the aggregated-time raster of forest disturbance was available on the NAFD-NEX webpage (https://daac.ornl.gov/NACP/guides/NAFD-NEX_Forest_Disturbance.html) as VCT_30m_last.tif from version R1 of the Forest Disturbance History dataset. It provided a value at each pixel representing the last year of forest disturbance, calculated as the time passed since 1970 (i.e. value = last year of disturbance - 1970). A value of 4 meant that no disturbance was identified in the time period. I limited VCT_30m_last.tif to values greater than or equal 31 (representing 2001) and for values less than or equal to 40 (representing 2010). Mirroring my process for GFC, the individual pixel values were also queried to create rasters by year (i.e. the raster for year 2001 was created by querying for pixels with a value 31, and so on). In my aggregated-time raster, pixels with the values between 31 and 40 were then set to a value of 1 to indicate identified disturbance in my study period, while all other pixels were given a value of 0.

For LANDFIRE, the yearly rasters of disturbance were available on the LANDFIRE website (https://www.landfire.gov/disturbance_2.php) for 2001 to 2010 (version 1.4.0.). Each raster was named US_DISTYEAR.tif, where YEAR is the year of disturbance. Each raster labeled pixels with values for various disturbance types ranging from 11 to 1133, with the exception of value

equal to 15 which represented increased greenness after a previous disturbance. A value of 0 meant that no disturbance was identified in that year. While each yearly raster provided values at each pixel labeling the disturbance type and a measure of uncertainty, my processing of the LANDFIRE data treated all disturbance types equally, as GFC and NAFD did not identify disturbance type or uncertainty. As such, following my process for GFC and NAFD, I queried the pixels disturbed in each yearly raster (i.e. values between 11 and 1133 with the exception of 15) and relabeled them with a value of 1 to indicate identified disturbance in that year, while all other pixels were given a value of 0. These yearly rasters were then aggregated in a raster calculation via the sum function to create the aggregated-time raster for LANDFIRE. All pixels identified as disturbed in any of the yearly rasters from 2000-2010 were given a value of 1, while all other pixels were given a value of 0.

Use of FRAP Fire Perimeters for Fire Occurrence and Coverage Analysis by Size Class

For a reference dataset on fire perimeters, this study used the vector database provided by the CALFIRE Fire and Resource Assessment Program (FRAP), version 15.1. My study included fire perimeters recorded for 2000 to 2010, which allowed for inclusion of areas that were already burned in year 1 of my study. The FRAP data are compiled in collaboration with the Bureau of Land Management, National Park Service, and U.S. Forest Service (USFS) and represent "the most complete digital record of fire perimeters in California" for my study period (http://frap.fire.ca.gov/projects/fire_data/fire_perimeters_index). Each fire perimeter had associated attributes including name, year of occurrence, and size in acres (a standard unit of measurement within the field of fire management). I queried the original FRAP database for fire perimeters between 2000 and 2010 (n = 3,131). As the name attribute in the FRAP dataset did not contain unique values and sometimes contained null values, a unique identifier for each fire perimeter was created by combining the name and the object identifier, or OBJECTID, assigned in the table (ex: Barrel_360). Fires with extents occurring exclusively outside of the California boundary from ESRI (i.e. fires on islands, in other states, or in Mexico) were excluded from the analysis. I also excluded fires with extents smaller than 1 acre (approximately 4,046 square meters); for fires with partial extents in California, the size class was determined by the original acreage, though only the area contained within California was analyzed. The analyzed fire perimeters (n = 3023) were converted from vector data in the ESRI File Geodatabase format to GeoTIFF rasters using ESRI ArcGIS. For each year, I created a new raster of the fire perimeters occurring in that year, using a 30 m pixel resolution to match the disturbance products and the MTBS data. In these yearly rasters, pixels occurring within a fire perimeter are labeled with a value of 1, while a value of 0 indicates pixels outside of a fire perimeter. Next, I uploaded these yearly rasters to EE and aggregated them using a raster calculation (i.e. add function) to create a raster of fire occurrence for 2000-2010. Any value above 0 indicated a fire at a given pixel sometime within 2000 to 2010 (according to the FRAP fire perimeters), with values ranging from 1 to a maximum value of 4. Approximately 96% of the pixels in the final fire occurrence raster had a value of 1 (meaning the pixel was only burned once in the study period), 4% had a

value of 2, and values 3 and 4 were negligible. A value of 0 indicated that the pixel was not included in any fire perimeter within the study period.

Additional Notes on Data for Environmental Conditions

This study used the National Elevation Dataset (m) for the United States, which is available in EE at a 10 m spatial resolution. The raster data for CWD (mm) was obtained from the California Climate Commons at a spatial resolution of 270 m. This study used the 2014 version of the 30-year summary for 1981-2010 as calculated by the California Basin Characterization Model (http://climate.calcommons.org/dataset/2014-CA-BCM). As CWD is a measure of how much potential evaporation exceeds actual evaporation, it was used in this study as a proxy for drought conditions. A raster of the 30-year normal conditions for mean temperature (degrees C) between 1981 and 2010 was obtained from the PRISM Climate Project at a spatial resolution of 800 m (http://prism.oregonstate.edu/normals/). The original spatial resolutions for CWD (270 m) and mean temperature (800 m) were coarser than the disturbance products and reference data (30 m) and the elevation data (10 m). Although it is typically encouraged to rescale analyses up to the resolution of the coarsest data, this would have greatly diminished the specificity of locations identified as disturbed in each product. A comparison of analyses of mean temperature at 30 and 800 m was also completed and did not indicate any major differences between the results (Appendix B).

Habitat Types Derived from FVEG data

For comparisons across vegetation types, this study used the dataset known as FVEG from CALFIRE FRAP, which collaborates with the California Department of Fish and Wildlife and USFS to compile the "'best available' land cover data available for California into a single comprehensive statewide data set" (as described in metadata found on http://frap.fire.ca.gov/data/frapgisdata-sw-fveg_download) . This study used version 15.1 for 1990-2014, within which the vegetation types are coded as wildlife-habitat relationships (WHR) for California at a 30 m spatial resolution. Before uploading to EE, I converted the ESRI File Geodatabase format of FVeg to a GeoTIFF raster using ESRI ArcGIS. Within EE, I relabeled the pixel values (i.e. WHR values) into four major vegetation cover types for analysis: forest, scrub/shrub, grass, and other (i.e. desert, agriculture, wetlands, barren, urban).

WHRNUM	WHRNAME	Class	Habitat Type
1	Alpine-Dwarf Shrub	Scrub_Shrub	Scrub_Shrub
3	Annual Grassland	Grass	Grass
4	Alkali Desert Scrub	Desert	Other
5	Aspen	Forest	Forest
6	Barren	Barren	Other
7	Bitterbrush	Scrub_Shrub	Scrub_Shrub
8	Blue Oak-Foothill Pine	Forest	Forest
9	Blue Oak Woodland	Forest	Forest
10	Coastal Oak Woodland	Forest	Forest
11	Closed-Cone Pine-Cypress	Forest	Forest
12	Chamise-Redshank Chaparral	Scrub_Shrub	Scrub_Shrub
13	Coastal Scrub	Scrub_Shrub	Scrub_Shrub
14	Douglas Fir	Forest	Forest
15	Desert Riparian	Desert	Other
17	Desert Scrub	Desert	Other
18	Desert Succulent Shrub	Desert	Other
19	Desert Wash	Desert	Other
20	Eastside Pine	Forest	Forest
21	Estuarine	Aquatic	Other
22	Fresh Emergent Wetland	Wetland	Other
24	Jeffrey Pine	Forest	Forest
25	Joshua Tree	Desert	Other
26	Juniper	Forest	Forest

27	Klamath Mixed Conifer	Forest	Forest
28	Lacustrine	Aquatic	Other
29	Lodgepole Pine	Forest	Forest
30	Low Sage	Scrub_Shrub	Scrub_Shrub
32	Mixed Chaparral	Scrub_Shrub	Scrub_Shrub
34	Montane Chaparral	Scrub_Shrub	Scrub_Shrub
35	Montane Hardwood-Conifer	Forest	Forest
36	Montane Hardwood	Forest	Forest
37	Montane Riparian	Riparian	Other
39	Perennial Grassland	Grass	Grass
40	Pinyon-Juniper	Forest	Forest
41	Palm Oasis	Agriculture	Other
42	Ponderosa Pine	Forest	Forest
43	Riverine	Aquatic	Other
44	Redwood	Forest	Forest
45	Red Fir	Forest	Forest
48	Subalpine Conifer	Forest	Forest
49	Saline Emergent Wetland	Wetland	Other
50	Sagebrush	Scrub_Shrub	Scrub_Shrub
51	Sierran Mixed Conifer	Forest	Forest
53	Urban	Urban	Other
55	Valley Oak Woodland	Forest	Forest
56	Valley Foothill Riparian	Riparian	Other
57	Water	Water	Other
58	White Fir	Forest	Forest

59	Wet Meadow	Wetland	Other
60	Cropland	Agriculture	Other
61	Orchard - Vineyard	Agriculture	Other
62	Undetermined Shrub	Scrub_Shrub	Scrub_Shrub
63	Undetermined Conifer	Forest	Forest
66	Dryland Grain Crops	Agriculture	Other
67	Deciduous Orchard	Agriculture	Other
68	Evergreen Orchard	Agriculture	Other
69	Irrigated Grain Crops	Agriculture	Other
70	Irrigated Row and Field Crops	Agriculture	Other
71	Irrigated Hayfield	Agriculture	Other
72	Pasture	Agriculture	Other
75	Vineyard	Agriculture	Other
76	Undetermined Hardwood	Forest	Forest
77	Eucalyptus	Forest	Forest
78	Rice	Agriculture	Other
79	Marsh	Wetland	Other

MTBS Burn Severity

For a reference dataset on burn severity, this study used the raster data provided by the MTBS project from USGS EROS and USFS, which analyzes the LTS to map burn severity for all fires greater than 1,000 acres in size in the Western U.S (https://www.mtbs.gov/direct-download). To target burned areas, the MTBS project obtains fire perimeters from the National Interagency Fire Center, supplementing this data as needed from other land management and fire agencies. For these areas, burn severity is calculated using a dNBR analysis between pre- and post-fire images and pre-defined thresholds of difference that vary by ecological zone. After calculating burn severity over the targeted area, MTBS assigns a new fire perimeter based on their results. The resulting MTBS burn severity data are provided as annual rasters for 2000 to 2010 at the 30 m

spatial resolution of the other data in my study. The values of burn severity are classified as follows: (0) pixels not included in the data (i.e. background data or pixels not included in MTBS); (1) unburned to very low severity; (2) low severity; (3) medium severity; and (4) high severity. Approximately 21% of the pixels in my aggregated raster of maximum burn severity across 2000-2010 had a value of unburned to very low, while 33% had a value of low. 26% had a value of medium, and 20% had a value of high. A value of 5, indicating increased greenness after fire, was not included in my study. More information regarding the creation of the MTBS dataset can be found in Eidenshink et al., 2007.

Appendix 3-2. Detailed Results for Chapter 3

Annual and Total Reported Disturbance

Area Reported as Disturbed (m2/year)						
Year	**GFC**	**% of CA***	**NAFD**	**% of CA***	**LANDFIRE**	**% of CA***
2001	565,642,750.63	0.138%	1,240,159,716.43	0.303%	2,117,351,040.65	0.518%
burn **	264,608,477.85	0.065%	396,084,177.34	0.097%	1,029,627,458.31	0.252%
no burn	301,034,272.78	0.074%	844,075,539.09	0.207%	1,087,723,582.35	0.266%
2002	822,541,689.12	0.201%	1,455,070,286.99	0.356%	3,439,689,012.80	0.842%
burn **	508,177,644.35	0.124%	781,333,920.09	0.191%	2,183,726,250.84	0.534%
no burn	314,364,044.77	0.077%	673,736,366.90	0.165%	1,255,962,761.96	0.307%
2003	733,958,008.16	0.180%	1,132,966,059.06	0.277%	5,179,638,469.97	1.268%
burn **	448,610,269.33	0.110%	473,768,386.66	0.116%	3,911,508,620.81	0.957%
no burn	285,347,738.83	0.070%	659,197,672.40	0.161%	1,268,129,849.16	0.310%
2004	764,186,617.20	0.187%	1,851,571,965.49	0.453%	2,524,008,478.72	0.618%
burn **	404,518,936.43	0.099%	1,117,040,272.72	0.273%	1,123,632,290.37	0.275%
no burn	359,667,680.77	0.088%	734,531,692.77	0.180%	1,400,376,188.35	0.343%
2005	536,076,976.46	0.131%	858,378,650.41	0.210%	2,540,633,102.42	0.622%
burn **	157,232,454.53	0.038%	275,829,738.30	0.067%	1,015,485,477.23	0.249%
no burn	378,844,521.94	0.093%	582,548,912.12	0.143%	1,525,147,625.19	0.373%

2006	1,194,180,079.49	0.292%	1,001,729,100.62	0.245%	4,578,914,498.46	1.121%
burn **	779,857,647.69	0.191%	318,926,849.51	0.078%	3,124,010,128.34	0.764%
no burn	414,322,431.80	0.101%	682,802,251.11	0.167%	1,454,904,370.12	0.356%
2007	1,663,143,934.96	0.407%	2,502,284,298.95	0.612%	5,560,521,912.94	1.361%
burn **	1,282,671,446.25	0.314%	1,592,802,289.07	0.390%	4,198,234,134.34	1.027%
no burn	380,472,488.70	0.093%	909,482,009.88	0.223%	1,362,287,778.59	0.333%
2008	2,302,505,126.02	0.563%	2,696,203,753.39	0.660%	8,182,624,092.76	2.002%
burn **	1,927,956,263.08	0.472%	2,117,897,286.22	0.518%	5,834,299,166.56	1.428%
no burn	374,548,862.94	0.092%	578,306,467.17	0.142%	2,348,324,926.20	0.575%
2009	1,390,488,034.72	0.340%	1,515,451,902.32	0.371%	3,549,172,085.04	0.869%
burn **	1,093,282,788.94	0.268%	955,128,790.12	0.234%	2,076,418,733.38	0.508%
no burn	297,205,245.78	0.073%	560,323,112.19	0.137%	1,472,753,351.66	0.360%
2010	392,809,854.51	0.096%	1,143,055,586.16	0.280%	1,855,639,028.85	0.454%
burn **	117,449,984.09	0.029%	683,139,641.18	0.167%	884,212,228.89	0.216%
no burn	275,359,870.42	0.067%	459,915,944.98	0.113%	971,426,799.96	0.238%
Total Across Years	10,365,533,071.26	2.537%	15,396,871,319.82	3.768%	unique area = 34,380,090,846.33 multiple disturbances = 39,528,191,722.63	unique area = 8.413% multiple disturbances = 9.67%
Total Burned Across Years	6,984,365,912.54	1.71%	8,711,951,351.21	2.132%	unique area = 22,669,887,841.863 multiple disturbances = 25,381,154,489.074	unique area = 5.548% multiple disturbances = 6.21%

Total Unburned Across Years	3,381,167,158.72	0.83%	6,684,919,968.61	1.64%	unique area = 11,710,203,004.467 multiple disturbances = 14,147,037,233.555	unique area = 2.865% multiple disturbances = 3.46%

Disturbance by Habitat Type

Scrub_Shrub		
	Area (m2)	**Percent of FVEG Total Area for CA**
CA	60,500,772,046.28	-
GFC	3,879,083,801.63	6.41%
NAFD	6,240,823,898.91	10.32%
LANDFIRE	12,283,388,186.91	20.30%

Forest		
	Area (m2)	**Percent of FVEG Total Area for CA**
CA	122,343,395,591.28	-
GFC	5,148,225,444.01	4.21%
NAFD	7,327,788,601.76	5.99%
LANDFIRE	17,058,583,230.48	13.94%

Grass		
	Area (m2)	**Percent of FVEG Total Area for CA**
CA	43,719,015,098.11	-
GFC	426,423,162.99	0.98%
NAFD	651,917,757.83	1.49%
LANDFIRE	3,009,560,077.55	6.88%

Other (Ag, Desert, Urban, Wetland, etc)		
	Area (m2)	**Percent of FVEG Total Area for CA**
CA	182,079,305,596.65	-
GFC	911,800,662.63	0.50%
NAFD	1,176,341,061.33	0.65%
LANDFIRE	2,028,544,435.44	1.11%

Environmental Conditions

Habitat as Proportion of Map				
Map	**Scrub_Shrub**	**Forest**	**Grass**	**Other**
CA	14.81%	29.94%	10.70%	44.56%
GFC	37.42%	49.67%	4.11%	8.80%
NAFD	40.53%	47.59%	4.23%	7.64%
LANDFIRE	35.73%	49.62%	8.75%	5.90%

Elevation							
Map	**Mean**	**StdDev**	**Min**	**P_25**	**P_50**	**P_75**	**Max**
California	860.79	738.25	-83.03	239.88	719.90	1,328.26	4,407.22
GFC	1,113.75	566.25	-69.63	711.94	1,127.90	1,512.01	3,846.09
NAFD	1,161.34	596.50	-80.59	720.08	1,168.03	1,583.99	4,080.56
LANDFIRE	1,110.52	598.38	-67.79	615.97	1,096.13	1,560.12	3,935.54

Climate Water Deficit							
Map	**Mean**	**StdDev**	**Min**	**P_25**	**P_50**	**P_75**	**Max**
California	930.53	338.80	0.22	643.97	939.94	1,220.08	1,566.16
GFC	725.20	235.25	0.00	539.94	708.10	931.95	1,528.15
NAFD	742.26	234.45	0.39	556.06	740.11	940.01	1,545.86
LANDFIRE	756.78	246.15	20.50	556.03	740.05	980.04	1,525.99

Mean Temperature							
Map	**Mean**	**StdDev**	**Min**	**P_25**	**P_50**	**P_75**	**Max**
California	13.97	4.66	-7.00	11.07	15.00	17.00	24.00
GFC	12.15	3.04	-5.00	10.01	12.01	15.00	23.00
NAFD	11.98	3.33	-3.00	9.01	12.03	15.00	24.00
LANDFIRE	12.23	3.39	-1.00	10.02	13.01	15.00	23.00

Mean Temperature (800 m resolution)							
Map	**Mean**	**StdDev**	**Min**	**P_25**	**P_50**	**P_75**	**Max**
California	13.97	4.66	-7.00	10.97	15.00	17.14	24.00
GFC	12.08	3.03	2.00	10.00	11.47	14.00	18.00
NAFD	11.82	3.25	-1.00	9.00	11.48	14.00	23.00
LANDFIRE	12.07	3.39	0.00	9.51	11.94	14.67	23.00

FRAP Results

Detailed Summary of FRAP Occurrence Data					
	All California	**Scrub/Shrub (m2)**	**Forest (m2)**	**Grass (m2)**	**Other (m2)**
Unburned Area*	384,830,504,518.42	49,691,960,909.66	113,405,563,401.47	41,268,902,802.69	180,464,077,404.60
Fire Occurrence = 1	22,847,993,680.96	10,126,708,807.36	8,813,862,205.17	2,331,159,085.54	1,576,263,582.89
Fire Occurrence = 2	929,877,879.47	659,543,832.10	121,590,385.38	111,193,416.73	37,550,245.26
Fire Occurrence = 3	33,774,821.11	22,513,109.98	2,287,936.39	7,737,795.94	1,235,978.80
Fire Occurrence = 4	16,559.28	2,217.76	9,251.25	3,639.93	1,450.34
Total Burned Area	23,811,662,940.82	10,808,767,967.20	8,937,749,778.19	2,450,093,938.14	1,615,051,257.29
Total Habitat Area	408,642,167,459.24	60,500,728,876.86	122,343,313,179.66	43,718,996,740.83	182,079,128,661.89
Percent of Habitat Burned	**5.83%**	**17.87%**	**7.31%**	**5.60%**	**0.89%**

* no overlap between FVEG and FRAP

FRAP Fire Perimeters	
Year	**Count**
2000	174
2001	200
2002	238
2003	331
2004	268
2005	301
2006	309
2007	326
2008	425
2009	247
2010	204
Total	**3023**

Total count of FRAP Fire Perimeters by size class					
lt100	**gt100_500**	**gt500_1000**	**gt1000_10000**	**gt10000_90000**	**gt_90000**
1571	727	219	386	108	12

Mean percentage of pixels captured within FRAP Fire Perimeters for each size class						
	lt100	**gt100_500**	**gt500_1000**	**gt1000_ 10000**	**gt10000_ 90000**	**gt_90000**
GFC	22.32%	18.70%	16.17%	21.39%	29.44%	36.64%
NAFD	23.68%	21.43%	21.18%	28.12%	35.64%	44.70%
LANDFIRE	93.94%	96.20%	97.66%	95.67%	97.04%	98.66%

MTBS Results

Detailed Summary of MTBS data					
	All California	**Scrub/Shrub (m2)**	**Forest (m2)**	**Grass (m2)**	**Other (m2)**
Unburned Area*	385,056,609,933.27	49,549,087,184.92	113,423,096,108.55	41,539,319,494.38	180,545,107,145.42
Unburned to Low Burn Severity	5,003,826,045.20	1,577,259,365.15	2,335,995,955.39	548,382,893.16	542,187,831.50
Low Burn Severity	7,680,473,323.94	3,084,837,646.32	3,040,897,488.37	933,211,766.73	621,526,422.52
Medium Burn Severity	6,241,557,027.19	3,510,410,957.14	2,008,118,323.54	481,475,955.10	241,551,791.41
High Burn Severity	4,659,701,129.63	2,779,133,723.32	1,535,205,303.81	216,606,631.45	128,755,471.05
Total Burned Area	23,585,557,525.94	10,951,641,691.92	8,920,217,071.11	2,179,677,246.44	1,534,021,516.47
Total Habitat Area	408,642,167,459.21	60,500,728,876.84	122,343,313,179.66	43,718,996,740.82	182,079,128,661.89
Percent of Habitat Burned	**5.77%**	**18.10%**	**7.29%**	**4.99%**	**0.84%**

* no overlap between FVEG and MTBS

Results of Product Overlap with MTBS Severity by Habitat Type

Scrub/Shrub									
	MTBS Total Area (m2)	**Unburned_Low**	**% of CA Unb_L**	**Low**	**% of CA Low**	**Med**	**% of CA Med**	**High**	**% of CA High**
CA	10,951,641,691.92	1,577,259,365.15	-	3,084,837,646.32	-	3,510,410,957.14	-	2,779,133,723.32	-
GFC	3,435,616,024.03	61,856,245.56	3.92%	436,348,913.18	14.14%	1,300,564,710.30	37.05%	1,636,846,155.00	58.90%
NAFD	4,793,953,711.70	226,618,688.88	14.37%	900,424,701.45	29.19%	1,710,322,204.24	48.72%	1,956,588,117.14	70.40%
LF	10,609,740,794.76	1,522,896,176.51	96.55%	3,002,621,193.14	97.33%	3,402,635,598.20	96.93%	2,681,587,826.91	96.49%

Forest									
	MTBS Total Area (m2)	**Unburned_Low**	**% of CA Unb_L**	**Low**	**% of CA Low**	**Med**	**% of CA Med**	**High**	**% of CA High**
CA	8,920,217,071.11	2,335,995,955.39	-	3,040,897,488.37	-	2,008,118,323.54	-	1,535,205,303.81	-
GFC	3,117,796,899.05	178,542,117.17	7.64%	681,951,001.67	22.43%	1,016,578,791.64	50.62%	1,240,724,988.58	80.82%
NAFD	3,291,521,458.13	217,565,265.12	9.31%	750,105,619.93	24.67%	1,062,841,969.13	52.93%	1,261,008,603.94	82.14%
LF	8,658,381,148.51	2,243,380,627.21	96.04%	2,946,978,155.76	96.91%	1,955,515,833.95	97.38%	1,512,506,531.59	98.52%

Grass									
	MTBS Total Area (m2)	Unburned_Low	% of CA Unb_L	Low	% of CA Low	Med	% of CA Med	High	% of CA High
CA	2,179,677,246.44	548,382,893.16	-	933,211,766.73	-	481,475,955.10	-	216,606,631.45	-
GFC	222,516,350.47	5,939,172.58	1.08%	32,860,158.36	3.52%	75,753,211.49	15.73%	107,963,808.05	49.84%
NAFD	354,144,765.61	29,865,828.89	5.45%	93,126,339.30	9.98%	120,041,329.10	24.93%	111,111,268.31	51.30%
LF	2,028,278,512.08	517,594,067.85	94.39%	892,249,275.14	95.61%	438,230,703.29	91.02%	180,204,465.80	83.19%

Other Land Cover Types (Ag, Desert, Urban, Wetland, etc)									
	MTBS Total Area (m2)	Unburned_Low	% of CA Unb_L	Low	% of CA Low	Med	% of CA Med	High	% of CA High
CA	1,534,021,516.47	542,187,831.50	-	621,526,422.52	-	241,551,791.41	-	128,755,471.05	-
GFC	158,948,269.30	7,284,231.67	1.34%	24,001,480.99	3.86%	51,570,059.68	21.35%	76,092,496.96	59.10%
NAFD	277,445,683.26	27,678,907.97	5.11%	92,945,378.66	14.95%	75,063,191.61	31.08%	81,758,205.02	63.50%
LF	1,440,124,836.60	512,991,526.86	94.62%	585,463,431.29	94.20%	221,697,357.21	91.78%	119,972,521.23	93.18%

Appendix 4-1 Supplemental Material and Detailed Results for Chapter 4

California Bioregions from Jepson Herbarium data

"The Jepson eFlora divides California into 35 ecologically distinct "bioregions" for the purpose of indicating where plant taxa grow" (Jepson Flora Project, 2017). This biologically driven data reflects topographic, climatic, and vegetation variations across California, and thus provides more meaningful divisions for regions, as compared to administrative boundaries such as county boundaries. Additional details and a map of this dataset can be found on

http://ucjeps.berkeley.edu/eflora/filter_keys.html. In this analysis, we included all bioregions with the exception of the two island bioregions for the North and South Channel Islands. For the California Floristic Province, we analyzed and reported the data by bioregion and by the aggregated region level. For the Great Basin and Desert Provinces, we analyzed and reported the data by region.

Spatial Agreement

	Total Area (m2)	Percent of California*	Percent of Total Area Reported as Disturbed**
No products reported disturbance	**368,620,890,008.72**	**90.21%**	**-**

	Total Area (m2)	Percent of California*	Percent of Total Area Reported as Disturbed**
Only one product reported disturbance			
GFC only	**1,209,128,095.74**	**0.30%**	**3.02%**
overlap with FRAP fire occurrence	29,231,101.39	0.01%	0.07%
no overlap with FRAP fire occurrence	1,179,896,994.35	0.29%	2.95%
NAFD only	**3,970,937,453.95**	**0.97%**	**9.92%**
overlap with FRAP fire occurrence	233,571,095.32	0.06%	0.58%
no overlap with FRAP fire occurrence	3,737,366,358.63	0.91%	9.34%
LANDFIRE only	**20,726,999,429.38**	**5.07%**	**51.79%**
overlap with FRAP fire occurrence	12,144,400,588.40	2.97%	30.34%
no overlap with FRAP fire occurrence	8,582,598,840.98	2.10%	21.44%

	Total Area (m2)	Percent of California*	Percent of Total Area Reported as Disturbed**
Two products reported disturbance			
GFC and NAFD only	**461,135,970.42**	**0.11%**	**1.15%**
overlap with FRAP fire occurrence	26,879,704.40	0.01%	0.07%

no overlap with FRAP fire occurrence	434,256,266.02	0.11%	1.09%
GFC and LANDFIRE only	**2,688,288,961.70**	**0.66%**	**6.72%**
overlap with FRAP fire occurrence	2,073,986,701.97	0.51%	5.18%
no overlap with FRAP fire occurrence	614,302,259.73	0.15%	1.53%
NAFD and LANDFIRE only	**4,957,815,697.03**	**1.21%**	**12.39%**
overlap with FRAP fire occurrence	3,597,232,146.72	0.88%	8.99%
no overlap with FRAP fire occurrence	1,360,583,550.31	0.33%	3.40%

	Total Area (m2)	Percent of California*	Percent of Total Area Reported as Disturbed**
All three products reported disturbance	**6,006,971,842.28**	**1.47%**	**15.01%**
overlap with FRAP fire occurrence	4,854,268,404.77	1.19%	12.13%
no overlap with FRAP fire occurrence	1,152,703,437.51	0.28%	2.88%

* Total area of California: 408,642.17 square kilometers

** Total area identified as disturbed by at least one product: 40,021.28 square kilometers (9.79% of CA)

Uncertainty

Low Uncertainty of Disturbance			
	Total Area (m2)	**Percent of California***	**Percent of Total Area Reported as Disturbed****
Total Low Uncertainty	**6,006,971,842.28**	**1.47%**	**15.01%**
overlap with FRAP fire occurrence	4,854,268,404.77	1.19%	12.13%
no overlap with FRAP fire occurrence	1,152,703,437.51	0.28%	2.88%

Medium Uncertainty of Disturbance			
	Total Area (m2)	**Percent of California***	**Percent of Total Area Reported as**

			Disturbed**
Total Medium Uncertainty	**8,107,240,629.15**	**1.98%**	**20.26%**
overlap with FRAP fire occurrence	5,698,098,553.10	1.39%	14.24%
no overlap with FRAP fire occurrence	2,409,142,076.05	0.59%	6.02%

High Uncertainty of Disturbance			
	Total Area (m2)	**Percent of California***	**Percent of Total Area Reported as Disturbed****
Total High Uncertainty	**25,907,064,979.07**	**6.34%**	**64.73%**
overlap with FRAP fire occurrence	12,407,202,785.11	3.04%	31.00%
no overlap with FRAP fire occurrence	13,499,862,193.96	3.30%	33.73%

* Total area of California: 408,642.17 square kilometers

** Total area identified as disturbed by at least one product: 40,021.28 square kilometers (9.79% of CA)

Uncertainty by Habitat Type

Habitat Type	**Total Area (m2)**	**Total Area Not Reported as Disturbed by any product (m2)**	**Total Area Reported as Disturbed by at least one product (m2)**	**% of Habitat Area that has been reported as Disturbed**
Scrub	60,500,728,876.85	46,879,177,817.52	13,621,551,059.32	22.51%
Forest	122,343,313,179.66	102,642,440,747.05	19,700,872,432.61	16.10%
Grass	43,718,996,740.82	40,459,903,613.11	3,259,093,127.72	7.45%
Other	182,079,128,661.89	178,639,367,831.04	3,439,760,830.85	1.89%

Low Uncertainty of Disturbance			
Habitat Type	**Total Area (m2)**	**Percent of Total Habitat Area Across California***	**Percent of Total Habitat Area Reported as Disturbed****
Scrub/Shrub	**2,718,961,685.59**	**4.49%**	**19.96%**
overlap with FRAP fire occurrence	2,548,745,404.86	4.21%	18.71%
no overlap with FRAP fire occurrence	170,216,280.72	0.28%	1.25%
Forest	**2,875,488,410.41**	**2.35%**	**14.60%**
overlap with FRAP fire occurrence	2,040,564,723.19	1.67%	10.36%
no overlap with FRAP fire occurrence	834,923,687.22	0.68%	4.24%

Grass	**257,578,982.39**	**0.59%**	**7.90%**
overlap with FRAP fire occurrence	158,242,738.88	0.36%	4.86%
no overlap with FRAP fire occurrence	99,336,243.50	0.23%	3.05%
Other	**154,942,763.90**	**0.09%**	**4.50%**
overlap with FRAP fire occurrence	106,715,537.83	0.06%	3.10%
no overlap with FRAP fire occurrence	48,227,226.06	0.03%	1.40%

Medium Uncertainty of Disturbance			
Habitat Type	**Total Area (m2)**	**Percent of Total Habitat Area Across California***	**Percent of Total Habitat Area Reported as Disturbed****
Scrub/Shrub	**3,343,814,464.74**	**5.53%**	**24.55%**
overlap with FRAP fire occurrence	2,996,207,580.99	4.95%	22.00%
no overlap with FRAP fire occurrence	347,606,883.75	0.57%	2.55%
Forest	**4,082,737,709.96**	**3.34%**	**20.72%**
overlap with FRAP fire occurrence	2,281,105,299.36	1.86%	11.58%
no overlap with FRAP fire occurrence	1,801,632,410.61	1.47%	9.14%
Grass	**313,649,894.87**	**0.72%**	**9.62%**
overlap with FRAP fire occurrence	220,837,581.93	0.51%	6.78%
no overlap with FRAP fire occurrence	92,812,312.93	0.21%	2.85%
Other	**367,038,559.58**	**0.20%**	**10.67%**
overlap with FRAP fire occurrence	199,948,090.81	0.11%	5.81%
no overlap with FRAP fire occurrence	167,090,468.77	0.09%	4.86%

High Uncertainty of Disturbance			
Habitat Type	**Total Area (m2)**	**Percent of Total Habitat Area Across California***	**Percent of Total Habitat Area Reported as Disturbed****
Scrub/Shrub	**7,558,774,909.00**	**12.49%**	**55.49%**
overlap with FRAP fire occurrence	4,964,175,719.08	8.21%	36.44%
no overlap with FRAP fire occurrence	2,594,599,189.92	4.29%	19.05%
Forest	**12,742,646,312.24**	**10.42%**	**64.68%**
overlap with FRAP fire occurrence	4,359,131,640.04	3.56%	22.13%
no overlap with FRAP fire occurrence	8,383,514,672.20	6.85%	42.55%
Grass	**2,687,864,250.46**	**6.15%**	**82.47%**
overlap with FRAP fire occurrence	1,883,035,138.14	4.31%	57.78%
no overlap with FRAP fire occurrence	804,829,112.32	1.84%	24.69%

Other	**2,917,779,507.37**	**1.60%**	**84.83%**
overlap with FRAP fire occurrence	1,200,860,287.85	0.66%	34.91%
no overlap with FRAP fire occurrence	1,716,919,219.52	0.94%	49.91%

* as a percentage of the habitat reported in Total Area (m2)
** as a percentage of the disturbed habitat reported in Total Area Reported as Disturbed by at least one product (m2)

Uncertainty by Bioregion

	Total Pixels	Disturbed Pixels	Per_Disturbed
Cascade Ranges	**24,735,979.97**	**4,772,214.27**	**15.34%**
Cascade Range Foothills Subregion	4,482,779.15	410,138.93	9.15%
High Cascade Range Subregion	20,253,200.82	4,362,075.35	21.54%
Central Western California	**50,162,571.74**	**5,534,154.13**	**8.89%**
Central Coast Subregion	6,874,447.34	164,221.90	2.39%
Inner South Coast Ranges District	12,566,882.28	391,238.94	3.11%
Outer South Coast Ranges District	18,146,191.90	3,897,285.24	21.48%
San Francisco Bay Area Subregion	12,575,050.22	1,081,408.06	8.60%
Desert Province	**137,899,270.16**	**1,240,705.67**	**2.24%**
Desert Mountains Subregion	4,651,703.98	250,259.87	5.38%
Mojave Desert Region	96,637,929.48	811,156.26	0.84%
Sonoran Desert Region	36,609,636.69	179,289.54	0.49%
Great Basin Province	**51,146,205.35**	**3,659,755.78**	**6.83%**
East of the Sierra Nevada Region	12,708,498.76	757,182.08	5.96%
Modoc Plateau Region	32,735,554.26	2,603,664.56	7.95%
Warner Mountains Subregion	2,159,519.09	276,023.49	12.78%
White and Inyo Mountains Subregion	3,542,633.23	22,885.65	0.65%
Great Central Valley	**80,362,272.92**	**2,268,876.62**	**2.78%**
Sacramento Valley Subregion	21,493,765.18	575,607.32	2.68%
San Joaquin Valley Subregion	58,868,507.74	1,693,269.30	2.88%
Northwestern California	**85,674,705.64**	**13,009,789.87**	**14.36%**
High North Coast Ranges District	7,105,889.92	1,716,630.93	24.16%
Inner North Coast Ranges District	16,361,850.03	973,613.07	5.95%
Klamath Ranges Region	33,946,266.05	7,180,617.93	21.15%
North Coast Subregion	2,917,629.18	268,619.88	9.21%

Outer North Coast Ranges District	25,343,070.47	2,870,308.06	11.33%
Sierra Nevada	**92,409,848.34**	**12,041,661.45**	**10.55%**
Central High Sierra Nevada District	17,061,022.15	1,746,067.40	10.23%
Central Sierra Nevada Foothills District	6,464,352.79	546,955.45	8.46%
Northern High Sierra Nevada District	30,120,992.22	5,539,028.39	18.39%
Northern Sierra Nevada Foothills District	10,037,401.58	817,471.24	8.14%
Southern High Sierra Nevada District	17,626,177.11	2,774,743.69	15.74%
Southern Sierra Nevada Foothills District	8,562,777.80	414,207.46	4.84%
Tehachapi Mountain Area Subregion	2,537,124.68	203,187.81	8.01%
Southwestern California	**47,166,784.65**	**13,729,142.46**	**33.52%**
Peninsular Ranges Subregion	15,301,994.28	5,607,768.53	36.65%
San Bernardino Mountains District	3,469,702.55	1,249,746.84	36.02%
San Gabriel Mountains District	3,300,031.57	1,848,707.12	56.02%
San Jacinto Mountains District	1,182,269.57	382,428.59	32.35%
South Coast Subregion	12,484,817.14	687,752.03	5.51%
Western Transverse Ranges District	11,427,969.55	3,952,739.35	34.59%

Low Uncertainty by Bioregion	Pixels_Low_Unc	Pixels_Low_Unc_Burned	Per_TotalDisturbed_Low_Unc	Per_TotalDisturbed_Low_Unc_Burned
Cascade Ranges	**613,443.00**	**131,681.00**	**8.34%**	**2.25%**
Cascade Range Foothills Subregion	11,846.00	6,677.00	2.89%	1.63%
High Cascade Range Subregion	601,597.00	125,004.00	13.79%	2.87%
Central Western California	**1,574,186.00**	**1,542,748.00**	**16.64%**	**15.37%**
Central Coast Subregion	8,423.00	3,069.00	5.13%	1.87%
Inner South Coast Ranges District	41,403.00	37,361.00	10.58%	9.55%
Outer South Coast Ranges District	1,348,684.00	1,330,024.00	34.61%	34.13%
San Francisco Bay Area Subregion	175,676.00	172,294.00	16.25%	15.93%
Desert Province	**8,182.00**	**7,941.00**	**0.65%**	**0.62%**

Desert Mountains Subregion	161.00	90.00	0.06%	0.04%
Mojave Desert Region	5,982.00	5,866.00	0.74%	0.72%
Sonoran Desert Region	2,039.00	1,985.00	1.14%	1.11%
Great Basin Province	**351,828.00**	**216,626.00**	**10.70%**	**8.83%**
East of the Sierra Nevada Region	70,785.00	66,204.00	9.35%	8.74%
Modoc Plateau Region	211,070.00	86,411.00	8.11%	3.32%
Warner Mountains Subregion	69,973.00	63,996.00	25.35%	23.18%
White and Inyo Mountains Subregion	0.00	15.00	0.00%	0.07%
Great Central Valley	**9,116.00**	**6,941.00**	**0.38%**	**0.26%**
Sacramento Valley Subregion	1,865.00	985.00	0.32%	0.17%
San Joaquin Valley Subregion	7,251.00	5,956.00	0.43%	0.35%
Northwestern California	**2,008,902.00**	**1,417,822.00**	**11.64%**	**6.95%**
High North Coast Ranges District	228,203.00	206,460.00	13.29%	12.03%
Inner North Coast Ranges District	46,547.00	38,022.00	4.78%	3.91%
Klamath Ranges Region	1,351,361.00	1,120,710.00	18.82%	15.61%
North Coast Subregion	23,593.00	4,049.00	8.78%	1.51%
Outer North Coast Ranges District	359,198.00	48,581.00	12.51%	1.69%
Sierra Nevada	**1,826,100.00**	**1,418,482.00**	**11.28%**	**9.43%**
Central High Sierra Nevada District	157,355.00	99,498.00	9.01%	5.70%
Central Sierra Nevada Foothills District	109,580.00	105,013.00	20.03%	19.20%

Northern High Sierra Nevada District	1,075,686.00	761,143.00	19.42%	13.74%
Northern Sierra Nevada Foothills District	87,371.00	66,020.00	10.69%	8.08%
Southern High Sierra Nevada District	383,105.00	374,332.00	13.81%	13.49%
Southern Sierra Nevada Foothills District	1,726.00	1,402.00	0.42%	0.34%
Tehachapi Mountain Area Subregion	11,277.00	11,074.00	5.55%	5.45%
Southwestern California	**2,063,272.00**	**2,034,928.00**	**14.04%**	**13.69%**
Peninsular Ranges Subregion	438,788.00	431,811.00	7.82%	7.70%
San Bernardino Mountains District	206,493.00	199,993.00	16.52%	16.00%
San Gabriel Mountains District	586,455.00	584,096.00	31.72%	31.59%
San Jacinto Mountains District	24,984.00	21,190.00	6.53%	5.54%
South Coast Subregion	10,298.00	9,092.00	1.50%	1.32%
Western Transverse Ranges District	796,254.00	788,746.00	20.14%	19.95%

Medium Uncertainty by Bioregion	Pixels_Med_Unc	Pixels_Med_Unc_Burned	Per_TotalDisturbed_Med_Unc	Per_TotalDisturbed_Med_Unc_Burned
Cascade Ranges	**996,987.34**	**157,711.26**	**15.83%**	**5.07%**
Cascade Range Foothills Subregion	39,887.90	29,558.47	9.73%	7.21%
High Cascade Range Subregion	957,099.44	128,152.79	21.94%	2.94%
Central Western California	**1,209,861.83**	**1,098,394.64**	**18.26%**	**14.76%**
Central Coast Subregion	16,235.70	6,479.83	9.89%	3.95%
Inner South Coast Ranges District	57,615.51	40,159.51	14.73%	10.26%

Outer South Coast Ranges District	847,202.94	784,672.31	21.74%	20.13%
San Francisco Bay Area Subregion	288,807.68	267,082.99	26.71%	24.70%
Desert Province	**111,533.39**	**94,105.23**	**10.83%**	**9.08%**
Desert Mountains Subregion	46,023.35	37,111.11	18.39%	14.83%
Mojave Desert Region	51,678.23	44,592.30	6.37%	5.50%
Sonoran Desert Region	13,831.81	12,401.83	7.71%	6.92%
Great Basin Province	**664,990.68**	**322,435.95**	**15.85%**	**11.29%**
East of the Sierra Nevada Region	204,872.31	184,114.51	27.06%	24.32%
Modoc Plateau Region	403,000.33	90,815.82	15.48%	3.49%
Warner Mountains Subregion	57,078.04	47,468.62	20.68%	17.20%
White and Inyo Mountains Subregion	40.00	37.00	0.17%	0.16%
Great Central Valley	**148,245.25**	**19,506.55**	**6.20%**	**0.67%**
Sacramento Valley Subregion	31,823.00	1,608.00	5.53%	0.28%
San Joaquin Valley Subregion	116,422.25	17,898.55	6.88%	1.06%
Northwestern California	**2,741,990.51**	**1,788,738.64**	**18.41%**	**10.48%**
High North Coast Ranges District	361,290.14	292,903.82	21.05%	17.06%
Inner North Coast Ranges District	158,845.08	125,559.47	16.32%	12.90%
Klamath Ranges Region	1,679,253.80	1,281,998.65	23.39%	17.85%
North Coast Subregion	36,741.76	4,503.04	13.68%	1.68%
Outer North Coast Ranges District	505,859.72	83,773.67	17.62%	2.92%

Sierra Nevada	**2,335,624.92**	**1,412,930.91**	**16.73%**	**11.42%**
Central High Sierra Nevada District	312,592.53	164,722.65	17.90%	9.43%
Central Sierra Nevada Foothills District	112,540.64	87,364.88	20.58%	15.97%
Northern High Sierra Nevada District	1,133,958.44	508,626.58	20.47%	9.18%
Northern Sierra Nevada Foothills District	140,652.93	78,174.05	17.21%	9.56%
Southern High Sierra Nevada District	582,583.15	530,383.25	21.00%	19.11%
Southern Sierra Nevada Foothills District	24,953.33	19,070.87	6.02%	4.60%
Tehachapi Mountain Area Subregion	28,343.90	24,588.62	13.95%	12.10%

Southwestern California	**3,184,274.14**	**3,021,228.59**	**21.60%**	**19.84%**
Peninsular Ranges Subregion	1,176,365.23	1,108,427.67	20.98%	19.77%
San Bernardino Mountains District	340,370.38	317,132.82	27.24%	25.38%
San Gabriel Mountains District	551,867.98	542,276.18	29.85%	29.33%
San Jacinto Mountains District	75,342.05	58,551.05	19.70%	15.31%
South Coast Subregion	45,891.06	34,020.94	6.67%	4.95%
Western Transverse Ranges District	994,437.44	960,819.93	25.16%	24.31%

High Uncertainty by Bioregion	Pixels_High_Unc	Pixels_High_Unc_Burned	Per_TotalDisturbed_High_Unc	Per_TotalDisturbed_High_Unc_Burned
Cascade Ranges	**3,161,783.93**	**463,466.57**	**75.83%**	**34.16%**
Cascade Range Foothills Subregion	358,405.02	261,209.15	87.39%	63.69%
High Cascade Range Subregion	2,803,378.91	202,257.43	64.27%	4.64%

Central Western California	**2,750,106.30**	**1,577,551.33**	**65.09%**	**31.80%**
Central Coast Subregion	139,563.20	41,245.38	84.98%	25.12%
Inner South Coast Ranges District	292,220.43	156,641.62	74.69%	40.04%
Outer South Coast Ranges District	1,701,398.29	980,629.96	43.66%	25.16%
San Francisco Bay Area Subregion	616,924.38	399,034.38	57.05%	36.90%
Desert Province	**1,120,990.28**	**756,162.88**	**88.53%**	**49.52%**
Desert Mountains Subregion	204,075.52	142,935.64	81.55%	57.11%
Mojave Desert Region	753,496.04	576,740.53	92.89%	71.10%
Sonoran Desert Region	163,418.73	36,486.72	91.15%	20.35%
Great Basin Province	**2,642,937.11**	**958,826.85**	**73.45%**	**24.59%**
East of the Sierra Nevada Region	481,524.77	323,306.13	63.59%	42.70%
Modoc Plateau Region	1,989,594.23	541,004.06	76.42%	20.78%
Warner Mountains Subregion	148,972.45	94,358.67	53.97%	34.19%
White and Inyo Mountains Subregion	22,845.65	158.00	99.83%	0.69%
Great Central Valley	**2,111,515.37**	**776,560.36**	**93.42%**	**33.02%**
Sacramento Valley Subregion	541,919.32	175,973.16	94.15%	30.57%
San Joaquin Valley Subregion	1,569,596.05	600,587.20	92.70%	35.47%
Northwestern California	**8,258,897.37**	**3,413,453.42**	**69.95%**	**24.52%**
High North Coast Ranges District	1,127,137.79	569,098.92	65.66%	33.15%

Inner North Coast Ranges District	768,220.99	458,329.16	78.90%	47.08%
Klamath Ranges Region	4,150,003.13	2,050,567.49	57.79%	28.56%
North Coast Subregion	208,285.12	6,275.56	77.54%	2.34%
Outer North Coast Ranges District	2,005,250.34	329,182.29	69.86%	11.47%
Sierra Nevada	**7,879,936.54**	**2,754,259.49**	**71.99%**	**32.95%**
Central High Sierra Nevada District	1,276,119.88	321,773.91	73.09%	18.43%
Central Sierra Nevada Foothills District	324,834.81	173,473.45	59.39%	31.72%
Northern High Sierra Nevada District	3,329,383.95	653,798.56	60.11%	11.80%
Northern Sierra Nevada Foothills District	589,447.31	165,075.44	72.11%	20.19%
Southern High Sierra Nevada District	1,809,055.55	1,104,256.88	65.20%	39.80%
Southern Sierra Nevada Foothills District	387,528.13	225,771.19	93.56%	54.51%
Tehachapi Mountain Area Subregion	163,566.91	110,110.06	80.50%	54.19%
Southwestern California	**8,481,596.33**	**6,489,209.40**	**64.36%**	**46.70%**
Peninsular Ranges Subregion	3,992,615.31	3,003,775.55	71.20%	53.56%
San Bernardino Mountains District	702,883.46	542,751.95	56.24%	43.43%
San Gabriel Mountains District	710,384.14	589,073.29	38.43%	31.86%
San Jacinto Mountains District	282,102.54	191,175.58	73.77%	49.99%
South Coast Subregion	631,562.97	388,261.76	91.83%	56.45%
Western Transverse Ranges District	2,162,047.91	1,774,171.27	54.70%	44.88%

Uncertainty by burn condition

Uncertainty by FRAP Fire Perimeter Size Class						
Certainty of Disturbance	**fire_lt100**	**fire_gt100_500**	**fire_gt500_1000**	**fire_gt1000_10000**	**fire_gt10000_90000**	**fire_gt90000**
No Reported Disturbance	10.94%	7.63%	5.52%	4.35%	2.12%	1.24%
Low Uncertainty	4.90%	6.61%	7.19%	12.06%	18.71%	27.59%
Medium Uncertainty	11.28%	12.71%	13.23%	19.56%	26.50%	26.06%
High Uncertainty	72.87%	73.05%	74.06%	64.03%	52.68%	45.12%

Uncertainty by MTBS Burn Severity				
	Unburned Low Severity	**Low Severity**	**Medium Severity**	**High Severity**
No Reported Disturbance				
total area	184,954,756.14	187,905,805.15	126,177,256.16	75,306,279.98
% of severity class	3.70%	2.45%	2.02%	1.62%
Low Uncertainty				
total area	70,763,900.19	542,325,585.08	1,642,537,876.36	2,564,917,281.06
% of severity class	1.41%	7.06%	26.32%	55.04%
Medium Uncertainty				
total area	591,813,766.83	1,861,856,959.92	2,030,359,436.09	1,252,135,576.76
% of severity class	11.83%	24.24%	32.53%	26.87%
High Uncertainty				
total area	4,156,293,622.03	5,088,384,973.79	2,442,482,458.58	767,341,991.81
% of severity class	83.06%	66.25%	39.13%	16.47%
Total Area in Severity	5,003,826,045.19	7,680,473,323.94	6,241,557,027.19	4,659,701,129.62