UNIVERSITY OF CALIFORNIA,
IRVINE

Advancing Precipitation Prediction Using a Composite of Models and Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Civil Engineering

by

Baoxiang Pan

Committee Members:
Professor Kuolin Hsu, Chair
Professor Amir Aghakouchak
Professor Efi Foufoula-Georgiou
Professor Soroosh Sorooshian

2019

DEDICATION

To

Emma and Helen

Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

I would like to sincerely thank my advisors, Professor Kuolin Hsu, Professor Amir Aghakouchak, and Professor Soroosh Sorooshian for their consistent support, encouragement, and spur through the PhD program in the Center of Hydrometeorology and Remote Sensing at UCI. I was about to give up my academic career and switch to another road of life, until Kuolin kindly offered me the invaluable opportunity to study and conduct research here. I would always be grateful for his help at this critical turning point of my life. During the memorable three and half years, I went to him constantly for advice, guidance, and critics, and I was always glad that I did. Amir offered me priceless suggestions for surviving in the academic community. No matter how poorly I did, I was always encouraged to practice my presentations, rectify my writing flows, and seek optimal ways to convey the ideas, not only to the community, but also to myself. Professor Sorooshian is always supportive to my brain-storm ideas. His care and encouragement far exceed what I deserve, and his support strongly inspired me to review existing scientific achievements and explore new ideas.

Many thanks go to Professor Efi Foufoula-Georgiou, Proferssor Jasper Vrugt, Professor Jinyi Yu and Professor Xiaogang Gao from UCI, Dr. Wayne Higgins from NOAA, Professor Fuqing Zhang from Penn State University, and Professor Gabriel Katul from Duke University. I am consistently encouraged for what they have contributed to the academic community, as well as their passion for new understandings. They are my role models that direct my career and life.

My sincere gratitude goes to my dear friends at CHRS: Dianne, Dan, Hao, Tiantian, Yumeng, Hoang, Mohammed, Phu, Negar, Ata, Raied, Pooya, Negin, Matin, Vista, Ailing, Jenny, Eric, Mojtaba Felicia, Charlotte, Alex, Sara, Laurie, Iman, Omid, Hassan, Aneseh, Laurence, Hamed, Elisa, Simon, Qiaohong, Qian, Alireza, Carlos, Moj, Juan, Samaneh, Yasir, and many more. I have spent happy hours with you guys in bars, campus walking, and academic or gossip discussions.

I would not be here without the support of my family. To my parents, Yujiao Ying and Changchun Pan, thanks for your unconditional love and encouragement. Sincere gratitude to my parents-in-law, Jun Ma and Xiaofang Cui. I appreciate your trust and support through these years. My deepest love and gratitude to my dear wife, Emma, Yixuan Ma, for sharing the bitters and joys in our lives, for supporting each other, for offering me a lovely home.

The research works conducted in this thesis are supported by the National Aeronautics and Space Administration (NASA) grant NNX16AO56G, U.S. Department of Energy (DOE prime award DE-IA0000018), National Oceanic and Atmospheric Administration (NOAA) grant (NA 14OAR4310222), and California Energy Commission grants (500-15-005 and 300-15-005). The Quadro P5000 GPU is kindly donated by the Nvidia Corporation to support the deep learning-related research work for the Center for Hydrometeorology and Remote Sensing at UCI.

CURRICULUM VITAE

Baoxiang Pan

EDUCATION

Ph.D., Civil Engineering **2015-2019**
University of California, Irvine *California, U.S.A.*

Master, Hydrology and Water Resources **2012-2015**
Tsinghua University *Beijing, China*

Bachelor, Hydrology and Water Resources **2008-2012**
Wuhan university *Wuhan, China*

Publications

- **Pan B.**, Hsu K., AghaKouchak A., Sorooshian S., and Higgins W., 2018, Precipitation Prediction Skill for West Coast United States – From Short to Extended Range, *Journal of Climate*, 32.1 (2019): 161-182.
- **Pan B.**, Hsu K., AghaKouchak A., Sorooshian S., 2018, Improving Precipitation Estimation Using Convolutional Neural Network, *Water Resources Research*, 2019/1, doi: 10.1029/2018WR024090
- **Pan B.**, Cong Z., 2016, Information Analysis of Catchment Hydrologic Patterns across Temporal Scales, *Advances in Meteorology*, 2016 (2016), doi: 10.1155/2016/1891465.
- Miao Q., **Pan B.**, Wang H., Hsu K., Sorooshian S., 2019, Improving Monsoon Precipitation Prediction using Combined Convolutional and Long Short Term Memory Neural Network, *Water*, Water 11(5), doi: 10.3390/w11050977
- Ma Y., Zhang Z., Ihler A., and **Pan B.**, Estimating Warehouse Rental Price using Machine Learning Techniques. *International Journal of Computers, Communications & Control.* 13.2 (2018), doi: ijccc.2018.2.3034.
- **Pan B.**, Hsu K., AghaKouchak A., Sorooshian S., 2018, Improving Hourly Precipitation Forecast using a Neural1 Encoder-Decoder Model. Prepared to submit to *Water Resources Research*.

Presentations

- **Pan B.**, Hsu K., AghaKouchak A., Sorooshian S., 2017, A Deep Neural Network Model for Improving Hourly Precipitation Estimates from Numerical Models. *AGU Fall Meeting, Washington D.C.*, 2018.
- **Pan B.**, Hsu K., AghaKouchak A., Sorooshian S., 2017, The Use of Convolutional Neural Network in Relating Precipitation to Circulation. *AGU Fall Meeting, New Orleans*, 2017.

- Ma Y., Zhang Z., **Pan B.**, 2017, Attributing Crop Production in the United States Using Artificial Neural Network. *AGU Fall Meeting, New Orleans*, 2017.

- **Pan B.**, Cong Z., 2014, Monthly Hydrological Model Evaluation through Mapping the Hydrological Pattern to Information Space. *AGU Fall Meeting, San Francisco*, 2014.

ABSTRACT

Advancing Precipitation Prediction Using a Composite of Models and Data

By

Baoxiang Pan

University of California, Irvine, 2019

Advances in numerical weather forecasts have brought forward considerable societal benefits and raised expectations for higher resolution, more accurate, and longer predictions. Despite the consistent progresses achieved, the prediction of precipitation remains a less satisfyingly tackled task, with skills falling far behind those of other atmospheric variables. This dissertation serves as an inspection of prediction capacity and an exploration of predictability for the precipitation process, with a particular focus on the region of West Coast United States.

The sources of predictability, accuracy requirements, and optimal model configurations are distinct regarding the considered forecasting scales and ranges. To identify the successes and deficiencies in predictions and benchmark further advances, a seamless assessment of precipitation prediction skill for short range up to subseasonal scale range is conducted. The evaluation is based on the Subseasonal-to-Seasonal Prediction Project retrospective forecast database. The prediction skill–lead time relationship is evaluated, using multiple models, and measured by both deterministic and probabilistic skill scores. Results show advantageous deterministic skills for the evaluated models at Week-1. The best-performing models achieved $r \approx 0.6$ for Week-2 predictions.

The potential sources of predictability at extended range from some of the key climate variations are investigated based on a composite of statistical evidences and numerical predictions. Results show that periods of heavy precipitation associated with ENSO are more predictable at the extended range period. The excessive precipitation and improved extended-range prediction skill during ENSO periods are attributed to the meridional shift of baroclinic

systems as modulated by ENSO. Through examining precipitation anomalies conditioned on the MJO, I verified that active MJO events systematically modulate the area's precipitation distribution. Most of the evaluated models are still struggling to represent the MJO or its associated teleconnections, especially at phases 3–4. However, some models do exhibit enhanced extended-range prediction skills under active MJO conditions.

The advantageous precipitation prediction skill for short to medium range originates from a steady accumulation of scientific achievements in (1) inferring atmospheric initial states, (2) resolving atmospheric fluid dynamics, and (3) approximating unresolved atmospheric processes. Evaluation results suggest that we have not fully realized the potentials of these advances in fostering a corresponding improvement in precipitation prediction. Here, the old art of forecasting by reading weather chart and advances in deep learning for image recognition are combined to shed light on the precipitation prediction task from a top-down, data-driven viewpoint. A deep convolutional neural network (CNN) model is trained to learn precipitation-related dynamical features from the surrounding dynamical and moisture fields by optimizing a hierarchical set of spatial convolution kernels. The model applies an "end-to-end" learning strategy to automatically search, synthesize, and extract salient spatial features from the resolved high-dimensional atmospheric field for accurate precipitation estimation at daily scale. Experiments for different regions across the contiguous United States show that, provided with enough data, precipitation estimates from the CNN model outperform the reanalysis precipitation products, as well as the statistical downscaling products using linear regression, nearest neighbor, random forest, or fully-connected deep neural network.

The idea of "end-to-end" learning for inferring unresolved precipitation process based on resolved atmospheric field is further explored for hourly scale quantitative precipitation forecast. Hourly precipitation observations from various sources are collected, quality controlled, and concatenated to compose a unique long-term (1980/1/1- 2018/12/31) high temporal resolution precipitation observation dataset. A general framework for statistically modeling of spatiotemporal data and making use of inconsistently available observations is developed.

Hourly precipitation predictions using the deep neural network model give $r \approx 0.8$ at $2^\circ \times 2.5^\circ$ spatial scale, while the baseline numerical model achieved $r \approx 0.5$. The best performance at hourly, gauge-point scale reaches the order of $r \approx 0.6$ for some gauges. However, there is high skill variance in estimating precipitation at such a stringent spatiotemporal resolution. To further test the proposed model in practical forecasts, dynamical retrospective forecast experiments for two atmospheric river land-falling events are carried out using the Weather Research and Forecasting (WRF) model. The WRF dynamical simulations are used to force the trained neural network model for alternative precipitation process predictions. Simulation results verified the consistency and robustness of the proposed approach. It should be noted that the methods here are not intended to replace precipitation-related parameterization schemes using a "black box" model, rather, the target is to set a benchmark for precipitation prediction from a data-driven perspective, and offer directions for improving precipitation related parameterizations.

Overall, this work conducted a systematical evaluation of precipitation prediction skills across a spectrum of critical scales and ranges. Sources of predictability at subseasonal scale are explored based on a composite of statistical analysis and numerical prediction. The potential of deep learning for seeking evidences in improving precipitation prediction is explored by combining high quality observation data with numerical dynamical predictions.

Chapter 1

Introduction

1.1 Background

1.1.1 Numerical Weather Prediction

The steady accumulation in knowledge about the earth atmosphere dynamics starts from the recognition that a particular set of partial differential equations provides deterministic descriptions for the atmospheric dynamics [1, 18]. These equations, usually named as *primitive equations*, apply conservation laws and thermodynamics laws on the continuous control volume of the atmosphere to describe the evolution of the atmospheric status, as defined by its pressure (P), temperature (T), density (ρ), humidity (q), and the three components of the flow velocity vector (u, v, w).

To integrate the *primitive equations* or their variations over space and time from an initial estimate of the atmosphere status forms the foundation of dynamical weather forecast. This paradigm becomes practical with the establishment of earth observation networks and development of numerical analysis and computation techniques. The explosively growing computation power allows us to discretize these equations at increasingly finer computation grids, and approximate the weather evolution by integrating the discretized equations forward. Some of the key achievements in the development of dynamical weather forecast

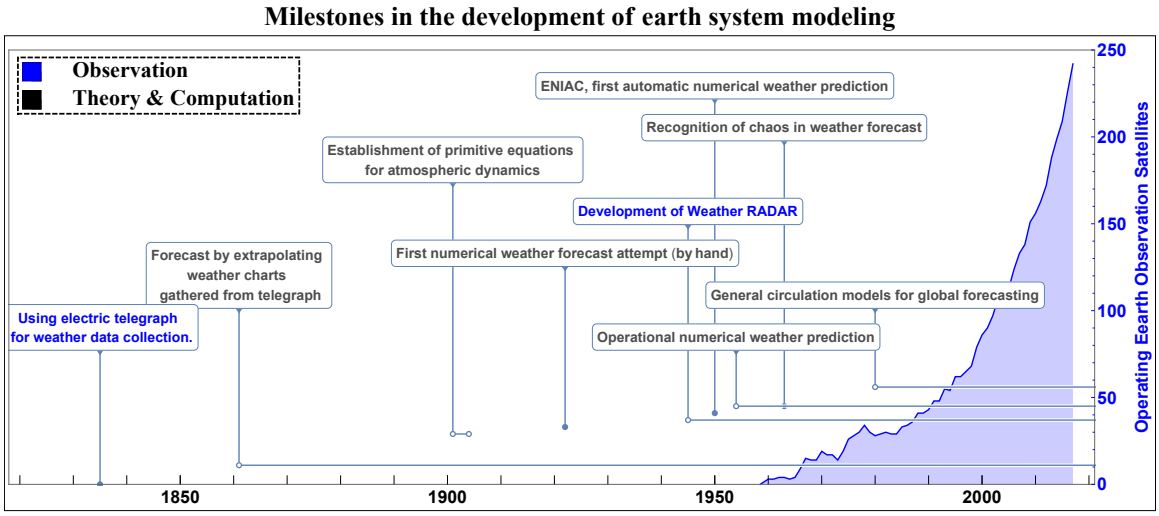


Figure 1.1: Key achievements in the development of atmosphere observation and modeling techniques. Blue denotations label the progresses in observation. Black denotations label the progresses in theory and modeling.

systems are summarized in Figure 1.1.

Modern numerical weather forecast systems have evolved to incorporate separate models around the modeling of the atmosphere [161], which facilitates the simulation of ocean, land/soil, cryosphere, and biogeochemical cycle, etc. Beside these modules, a data assimilation system is employed to merge observations with model predictions for optimal estimate of the system status. A suite of parameterization options are ready to be employed to account for the overall effects of models' unresolved processes, such as cumulus clouds, microphysics, radiation, planetary boundary layer, and land surface process. By integrating the components mentioned above, forecasts are nowadays regularly conducted by many of the operational forecast centers and research institutes across the world. The simulation results provide updated prediction information for short range up to climate range.

Within this dynamical forecast paradigm, prediction skill has been recognized to diminish along forecast lead time. This is because small status estimation errors would grow through model's iterative computations. For the sake of illustration, Figure 1.2 uses the Lorenz 1963 system to demonstrate the fact that two simulations from slightly pertubated initial states can yield widely diverging outcomes in a "chaotic" dynamical system.

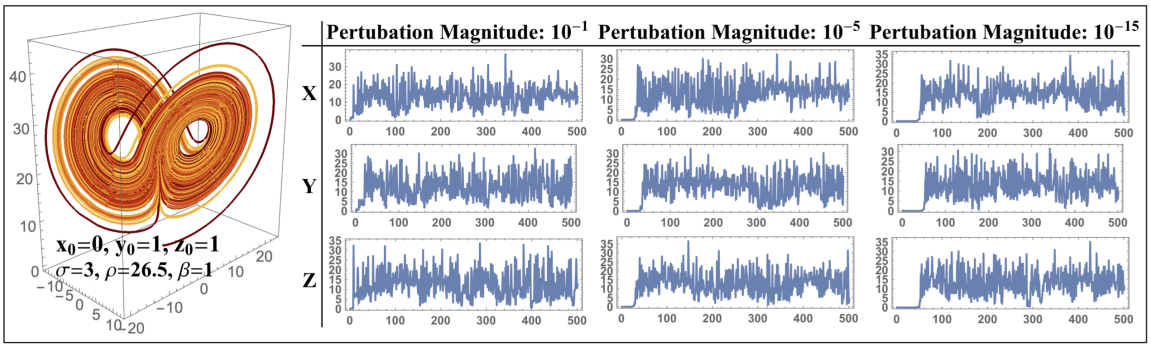


Figure 1.2: Numerical simulation of the Lorenz 1963 sysmtem: $\{\frac{dx}{dt} = \sigma(y - x), \frac{dy}{dt} = x(\rho - z) - y, \frac{dz}{dt} = xy - \beta z\}$. The left figure shows a trajectory simulation. The right figures show the trajectory discrepancy of two simulated projected on the x, y, and z axes. The perturbation magnitude for the initial state are 10^{-1}, 10^{-5}, and 10^{-15}.

Forecasting Range	Coverage	Source of Predictability	Primary Forecast Approach
Nowcast	0-2 hours	Initial State	Statistical Extrapolation
Very Short Range	Up to 12 hours	Initial State	Dynamical Model
Short Range	12-72 hours	Initial State	Dynamical Model
Medium Range	72 - 240 hours	Initial State	Dynamical Model
Extended Range	10-30 days	Initial&Boundary State	Dynamical Model
Long Range	30 days-2 years	Boundary State	Statistical Inference and Dynamical Model
Climate Range	beyond 2 years	Boundary State	Statistical Inference and Dynamical Model

Table 1.1: Meteorological Forecasting Ranges defined by World Meteorological Organization(WMO)

Results in Figure 1.2 highlight the crucial role of initial status estimation accuracy for short range forecasts and the difficulty for long-term deterministic forecasts. In practical weather forecast, it is imperative to distinguish critical forecasting ranges, clarify their sources of predictability, and figure out the optimal model configurations. Table 1.1 lists the key forecasting ranges defined by the World Meteorological Organization (WMO). At nowcast range, statistical extrapolation serves as the primary approach for inferring the evolution of the target atmospheric variable, such as precipitation. This is due to the fact that dynamical models typically takes certain spin-up time to include detailed observations in its state estimates. For short up to extended range, dynamical models have become arguably the only reliable tool for predictions. Forecasts beyond extended ranges generally show little deterministic skills, although we would expect models to capture low frequency signals from the boundary constraints to offer informative probabilistic forecasts. Statistical inferences

are often employed to seek evidences from observations, verifying model's capacities, and exploring prediction opportunities at long range to climate range.

The numerical dynamical forecasting paradigm introduced above have achieved consistent progress in predicting day-to-day weather variations, providing extreme weather warnings, informing climate variations, and supporting weather-related decisions. While progresses are still not constrained by limits on predictability yet [2], it is reasonable to expect further improvements to be achieved from advances in (1) inferring atmospheric initial states, (2) resolving atmospheric fluid dynamics, and (3) approximating unresolved atmospheric processes.

1.1.2 Precipitation Prediction using Numerical Weather Models

Precipitation plays fundamental role in the earth hydrological cycle. It is also crucial for agriculture, water resources allocation, emergency management, aviation, and many other aspects of human society. Despite its importance, precipitation remains to be less satisfyingly simulated and predicted in numerical weather prediction models. Precipitation prediction skills have been recognized to fall far behind those of other atmospheric variables, such as pressure, temperature, moisture, and wind speed. Evaluation studies suggest that models usually fail in revealing many critical aspects of precipitation, such as location, timing, intensity, or total accumulation [179, 184].

Compared to other atmospheric variables, precipitation demonstrates particularly high spatiotemporal variability. These irregular characteristics stem from the manifestation of individual formation and growth of precipitating clouds, which exhibit complex coupling with their surrounding atmospheric fluid dynamics [78]. To correctly simulate precipitation poses a stringent challenge on model's capacity to represent atmospheric dynamics and physics across a wide spectrum of scales: the model should first make realistic predictions of the atmospheric dynamics, so the air density, pressure, wind, and temperature are in the right place at the precise moment [184]; thereafter, the model should realistically infer the onset

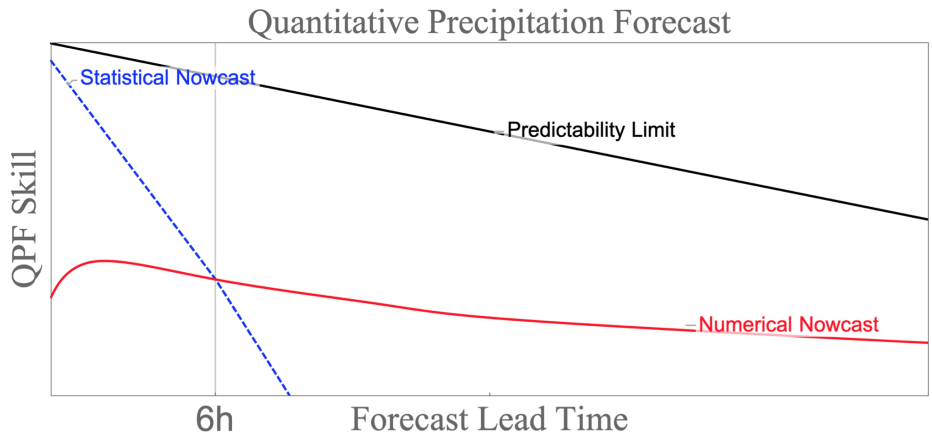


Figure 1.3: Schematic plot of statistical extrapolation (blue dashed line) and numerical prediction (red line) of precipitation at nowcast range.

and strength of the convections, approximate the evolution of cloud hydrometeors of different phases, and estimate the falling of the resulting precipitates. Errors from each of these aspects exhibit complex interactions through model's iterative computations, with all aspects quickly revealing themselves in model's precipitation outputs [184]. It is a daunting task to attribute, track, and rectify these error sources. Below I conduct a schematic analysis about how different error sources impact model's precipitation prediction at different prediction ranges.

Very Short Range to Medium Range Prediction

For very short range prediction of precipitation (0–6 hours), statistical extrapolation of current precipitation observations usually outperforms numerical precipitation predictions. A schematic comparison of prediction skill-forecast lead time relationship at very short range is drawn in Figure 1.3. The main reason that dynamical prediction skill falls behind statistical extrapolation is attributed to the errors in estimating the initial state of cloud hydrometeor distributions. In most prediction cases, we do not have direct and comprehensive observations for the initial field of cloud hydrometeors that come in liquid, solid, or mixed phases. The poor initial estimate results in poor predictions. It is widely accepted that statistical ex-

trapolation outperforms numerical predicting within a forecast lead time of 6 hours. Remote sensing for cloud and precipitation has shown promises in alleviating the initial estimation uncertainty for numerical modeling of cloud and precipitation. However, many peculiar issues should be tackled before fully realizing their potentials [172], such as the specification of model and observation error statistics, the formulation of the control vector, etc. For comprehensive reviews, see [43] and [107].

For short range up to medium range, numerical weather prediction models demonstrate particular advantage due to its comprehensive representation of atmospheric dynamics. However, deficiencies are found in the following aspects.

- Models' dynamical forcings are of limited resolution for making detailed representation of cloud dynamics and cloud microphysics. For instance, the overall effects of the subgrid embedding convective cells should be accounted by a cloud cumulus parameterization scheme. Since this approximation takes place at the "gray zone" between resolved and parameterized domain of the model, inevitable sampling variance error would be introduced.

- We usually cannot afford the computation cost to make detailed representation of the cloud microphysics. Cloud encapsulates particles of different phases, sizes, and physical properties. For physical comprehensiveness, bin-based microphysics parameterization schemes are preferred to bulk schemes, since the former explicitly represent the size distribution for each cloud particle species, while the latter apply few momentum parameters to approximate the size distribution based on predetermined distribution types. However, considering computation efficiency, bulk schemes are still preferred to bin-based schemes in most practices

- The equations and their related parameters in precipitation-related parameterization schemes are of inherent uncertainty, reflecting our insufficient understandings for the microscale phenomena [90], and the indeterministic nature of the parameterization task

[13].

The challenges above have been intensively investigated through:

- Increasing numerical model's resolution to resolve the "gray-zone" processes in coarse resolution models [139, 91].
- Updating our understanding and encoding for the unresolved processes based on targeted observations and high-resolution simulations.
- Calibrating parameters to reduce the mismatch between specific observations and model simulations [77].
- Shifting from deterministic to stochastic parameterization [13].
- Improving cloud and precipitation assimilation based on new observations and advanced algorithms.

These efforts are supposed to benefit from a clarification of how improvement from each single aspect contributes to model's QPF skill. Such a clarification offers insights into the sources of limits on prediction, thus directs further progresses to mitigate these limitations.

Extended to Long Range Prediction

Precipitation predictions for extended range up to long range subject to the limitation of the weather systems' predictability. As we try to predict precipitation at long lead time, the position, timing, and intensity of the precipitation forecasts start to diverge from what takes place in nature. Meanwhile, recent developments in forecast practices, statistical explorations, and theoretical analyses suggest the opportunity of predictability across a continuous spectrum of prediction ranges. Some of the potential directions for informative long term forecasts are highlighted below:

- **Ensemble Forecast:** Ensemble forecast gives indication of an envelope of possible future states of the atmosphere by running multiple simulations starting from pertubated initial estimations. The ensemble spread of forecasts can be applied to represent the prediction uncertainties. For extended range, predictions of high spatiotemporal resolutions generally hold little efficacy as influenced by the chaotic effect. However, ensemble predictions might offer crucial information about the probability of the distribution for the considered weather phenomenon. Ensembles involving multiple models can also help remedy the models' epistemic errors.

- **Sources of Predictability** The time and space discretization involved in numerical weather prediction models provides a discrete representation of the atmosphere, while in reality dynamical and physical processes operate on a continuous spectrum of spatial and temporal scales. The low frequency atmospheric signals, such as those from the sea and stratosphere, provide valuable sources of predictability at extended up to seasonal range. For instance, teleconnections between extratropical cyclone activities and tropical disturbances offer the potential for extending forecast lead time in the mid-latitude: Semi-periodic tropical variations, i.e., El Niño-Southern Oscillation (ENSO) and the Madden-Julian Oscillation (MJO), often trigger quasi-stationary Rossby wave-trains that propagate into mid-latitudes, which in turn influence the precipitation distribution [75, 158, 120, 74]. Such effects might exert different impacts for different regions. While much research focuses on explaining teleconnections [210, 126, 115], it is also imperative to keep investigating how forecasts of opportunities are expressed in numerical models. This is because statistical inference usually makes crude simplifications of the predictor-predictand relationship, while models makes more comprehensive representations of atmospheric inner-variability and boundary forcing information. Advances are more likely to be generated through going back and forth between statistical evidences and dynamical modelling [54].

1.1.3 Learning from Data

As is reviewed, precipitation prediction suffer from multiple sources of errors, with all of them entangling through iterative dynamical simulations. It is difficult to track, attribute, and rectify these intricate errors following a process-based approach. When a large number of variables are involved, the idea of learning from data offers a powerful alternative solution for model diagnosis.

The past decades have witnessed a booming of machine learning (ML) techniques accompanied by a deluge of data availability and exponential increase of computing capacity. From a nutshell viewpoint, a machine **M** is a statistical model. It is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P** if its performance at tasks in T, as measured by **P**, improves with experience **E** [122]. The key components a ML algorithm have been artfully summarized by Pedro Domingos [39]:

$$\text{Learning} = \text{Representation} + \text{Evaluation} + \text{Optimization} \tag{1.1}$$

Here the term *Representation* is applied to denote two aspects of meanings: the first aspect refers to the parametric form of the learning algorithm, which should be well defined and interpretable for a computer; the second aspect refers to the features we apply to quantify the predictors and predictand. As will be illustrated, a good feature representation of the predictors and predictands can very often alleviate the modeling difficulties and foster good performance in a ML task. *Evaluation* is typically expressed as a loss function, which tells the performance of the proposed learning algorithm. *Optimization* refers to the process of applying data to train the model for an optimal performance as measured by the loss function. To frame a ML problem requires a clear clarification of these three components and a careful formulation of these components regarding the characteristics of the learning task.

ML applications are preferred to offer evidences by combining data with prior knowledge

in order to generate novel knowledge [224]. In geoscience, where we have partially established principled solutions for the modeling of geophysical processes, numerical simulations are usually preferred to ML-based approaches. However, recent developments in the field of deep neural networks suggest that statistical models are capable of handling the complexities of the geophysical processes, and offering important implications for model improvements.

Deep neural networks (DNNs) are artificial neural network models (ANNs) with multiple hidden layers. An ANN approximates complicated functions through composing simple functions in hierarchical computing graphs. DNNs apply multiple hidden layers, with each layer transforming data representation at one level into a representation at a higher, slightly more abstract level [96]. Through the hierarchical transformation, DNNs can automatically learn customized feature representations for specific tasks. Besides being deeper, modern DNNs have developed efficient and effective architecture variations that scale well for high dimensional structured data. For instance, deep convolutional networks have demonstrated state-of-the-art performance in processing imagery data [93], deep networks with recurrent [69], attention [31], and memory [57] modules have brought about breakthroughs in sequential learning problems, such as natural language processing, speech and audio modeling [223]. A blending of the spatial/temporal modules have shown particular advantage in video and motion prediction [202], which have been recognized to share striking similarities to the modeling of geophysical processes [151]. While many recent research works have started to explore the applicability of DNNs for modeling geophysical processes [105, 149], it remains a question how DNN can translate the big data of observations and numerical simulations into precipitation estimation improvements [142].

1.2 Research Questions and Dissertation Organization

The fast development of numerical weather prediction techniques, accumulation in high quality observations and forecast experiment records, and new generation of statistical modeling

approaches offer unique opportunities for investigating precipitation prediction capacity, exploring their success and deficiencies at critical scale and ranges, and seeking the potential predictability that have not been realized by the modeling systems. To better understand and take use of these opportunities, the following research questions are proposed and addressed in this dissertation:

1. What is the status quo of current numerical weather prediction models in predicting precipitation at critical scales and ranges?

2. How do current numerical weather prediction systems capture the opportunity of predictability for precipitation prediction at extended ranges, compared to the evidences from statistical analysis?

3. How can deep learning be applied to shed light on improving precipitation prediction by combining data with our prior knowledge in dynamical atmospheric modeling?

To address these questions, the West Coast of United States is selected as the major study area. Data from various sources of observations, numerical retrospective forecast experiments, and numerical analysis are applied. Simple statistical analysis and complicated deep neural networks are explored. Results are verified against observations and targeted dynamical simulations.

The rest of this dissertation proceeds as follows: Chapter 2 conducts a seamless assessment of precipitation prediction skills for the West Coast of the United States. The potential sources of predictability at extended range from key climate variations are investigated, based on a comparison of statistical evidences and numerical predictions. Chapter 3 introduces the deep convolutional neural network to shed light on the old art of inferring precipitation from weather charts. The idea is further explored for hourly scale quantitative precipitation forecast in Chapter 4, where a general framework for statistically modeling of spatiotemporal data and making use of inconsistently available observations is developed. Conclusions are drawn in Chapter 5.

Chapter 2

Seamless Assessment of Precipitation Prediction Skills

2.1 Background

Precipitation is crucial for agriculture, water resources allocation, emergency management, aviation, and many other aspects of the society. The accuracy and extension of precipitation prediction is of consistent concern to many operational prediction and application communities. This chapter conducts an evaluation of precipitation prediction skills achieved by a number of state-of-the-art numerical weather prediction systems. The evaluated models have been extensively applied for operational forecasts by many forecast centers around the world. The primary objective is to offer a comprehensive inspection of precipitation prediction capacity at critical scales and ranges. Besides, the potential sources of predictability at extended range from some of the key climate variations are investigated, based on a comparison of statistical evidences and numerical predictions. Such an assessment helps to identify successes and shortcomings in the models [195], and sets the benchmark for further improvements.

The heavily populated West Coast of the United States is selected as the study area. This

region receives a majority of its precipitation during the cold season (October to March). This precipitation supports the water requirements of approximately 15.7% of the nation's population [137], generates approximately 52.6% of the domestic hydroelectricity [118], and waters approximately 21.7% of the country's irrigated farm land [192]. Occasional extended wet or dry periods, which are strongly linked to the presence or absence of winter storms, threaten the area's ecological and economic security. Additionally, extremes of droughts and floods can end or occur abruptly, posing challenges to public safety and many other aspects of the society. To better plan for and respond to the beneficial/destructive impacts of the precipitation variations, it is imperative to understand the accuracy and extent of the predictions.

There has been substantial progress in day-to-day precipitation predictions in the past six decades [9]. The skill improvements are largely due to 1) more realistic estimations of initial atmosphere conditions, and 2) improvements in the ability of numerical prediction models to simulate the dynamics and physics of the weather systems. While these advances have led to improved forecasts at longer lead time, it is also true that small-scale errors roughly double in 1–2 days, leading to a rapid loss of useful skill within about 2 weeks [108].

Provided with loosened requirements on spatial temporal resolutions, deterministic and ensemble-based forecasts occasionally provide useful prediction beyond the synoptic time range [204], which has the potential for significant economic value. Generally, the skill depends on 1) the existence of sources of predictability at corresponding temporal ranges, and 2) the model's ability to represent the dynamics associated with these modes of variability [130]. For regions with distinct dynamics and sources of predictability, the prediction skill at the extended range has been recognized to be different [226, 209].

Most winter precipitation events along the West Coast are driven by moisture convergence associated with passing extratropical cyclones [8, 33]. At short to medium ranges, due to the coherent life cycle of cyclone events, cyclogenesis is highly predictable. At the extended range, the prediction skill decreases rapidly [95, 226, 153], due to the chaotic nature of the baroclinic

systems. Teleconnections between extratropical cyclone activities and tropical disturbances offer the potential for extending forecast lead time. For instance, semi-periodic tropical variations, i.e., El Niño-Southern Oscillation (ENSO) and the Madden-Julian Oscillation (MJO), often trigger quasi-stationary Rossby wave-trains that propagate into mid-latitudes, which in turn influence the precipitation distribution [75, 158, 120, 74]. Such effects are expected to be more significant for the West Coast, given its proximity to the Pacific and the associated sources of potential predictability [204, 7, 126].

While much research focuses on explaining teleconnections [210, 126, 115], it is important to keep investigating how forecasts of opportunities are expressed in General Circulation Models (GCMs), since GCMs remain the most important tool for testing potential sources of predictability. Numerous studies have evaluated the ability of GCMs to predict intraseasonal variability at global and regional scales [132, 194, 98, 186, 205]. However, a systematic evaluation of the prediction skill for precipitation at short to extended range along the West Coast has not been reported.

The target of this chapter is to investigate short to extended range precipitation prediction skill for the West Coast during its rainy season. In particular, the impact of the leading modes of intraseasonal to seasonal variability on the distribution and prediction skill of precipitation is explored. The intention is to use the results as a baseline for follow-on investigations of seamless weather-climate prediction.

The evaluation is based on extended-range retrospective forecasts (hereafter referred to as hindcasts) experiments conducted by 11 operational centers and hosted by the World Weather Research Programme (WWRP)/World Climate Research Programme (WCRP) Subseasonal to Seasonal (S2S) Prediction Project science plan [195]. The abundance of hindcast cases and model diversity offer an unprecedented opportunity for investigation of the potential predictability and prediction skill of precipitation. The specific experiments are as follows:

1. Evaluate the prediction skill for West Coast precipitation during the cold season in

each GCM on time scales from short range to extended range.

2. Investigate the influence of intraseasonal and seasonal variability on precipitation prediction skill in the GCMs at extended range, with emphasis on ENSO and the MJO.

The rest of this chapter is organized as follows. Section 2 introduces the data used in this study. Section 3 describes the methodology, including evaluation strategy and skill scores. The evaluation results are presented in Section 4. Section 5 focuses on the impact of ENSO and MJO. Discussion and conclusions are given in Section 6.

2.2 Data and Materials

2.2.1 Study Area

The study area is restricted to the heavily populated coastal region of the Western United States, which includes California, Western Oregon, and Western Washington (to the west of 120°W as roughly divided by the Cascade Range). The west Cascade Range is considered separately from the east range due its distinct synoptic and precipitation regimes [19]. The proximity of this region to the neighboring Pacific Ocean suggests that it is likely to hold considerable predictability for extended range. Considering the climate variation within the study area, this region is further divided into four subdivisions for evaluation, namely Southern California (SCA), Northern California (NCA), Western Oregon (OR) and Western Washington State (WA). The study area is highlighted in Figure 2.1.

2.2.2 CPC Gauge based Daily Precipitation

The Climate Prediction Center (CPC) Unified Gauge-Based precipitation database [219] is used as "ground truth" for assessing performance of the GCMs. This database is constructed by merging various precipitation information sources, including gauge observations, satellite estimates, and numerical model predictions. It provides solid daily precipitation records

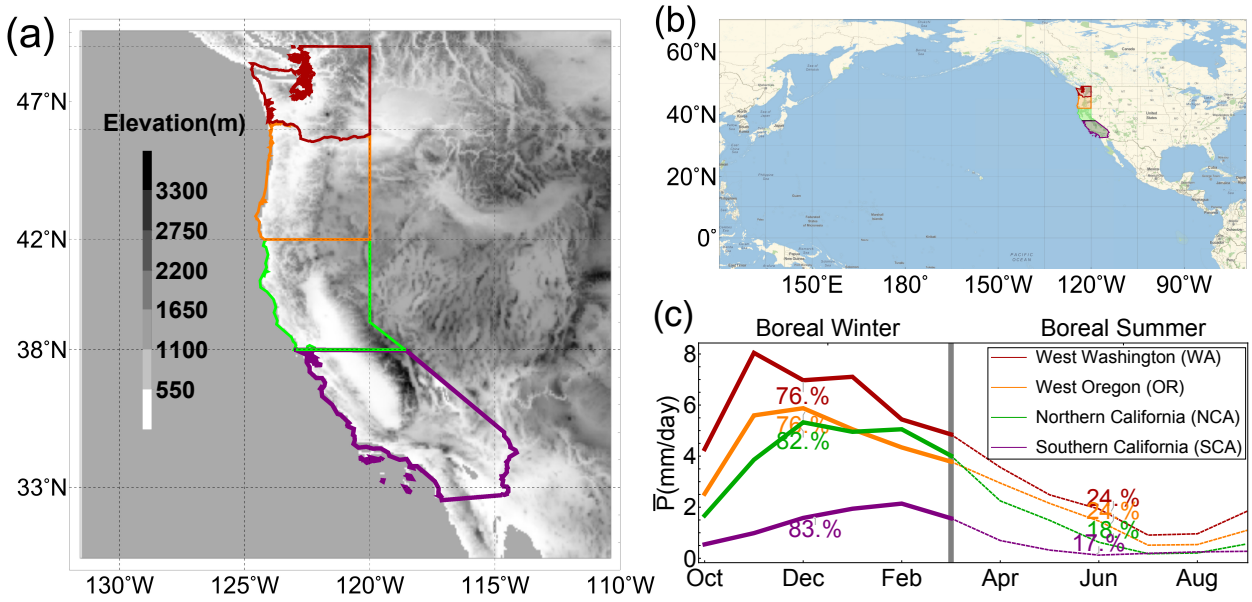


Figure 2.1: (a) The geographic map of the West Coast. The elevation data are provided by United States Geological Survey [53]. The four subdivisions, namely Southern California (SCA), Northern California (NCA), Western Oregon (OR), and Western Washington State (WA), are outlined with colored polygons. (b) Geoposition of the study area in a larger scale. (c) The monthly mean precipitation rate for the four subdivisions, based on the CPC precipitation dataset. The boreal winter (October to March) precipitation ratio are labeled.

covering the contiguous United States from 1948 to 2017 with spatial resolution of $0.25° \times 0.25°$. Data for the West Coast in cold season (October to March) are considered. The data are spatially-averaged to $(\text{lat}, \text{lon}) = 1.5° \times 1.5°$ grid to match the resolution of the GCM hindcast precipitation products.

2.2.3 Climate Indices

The leading patterns of intraseasonal to interannual variability considered in this study are ENSO and the MJO. ENSO is measured by the Niño 3.4 index [189], which is the mean monthly sea surface temperature anomalies (SSTA) averaged from 5°S–5°N and 170°W–120°W. The MJO is quantified by the real-time multivariate MJO Index (RMM), which consists of the two leading principal components (PCs) of the field that combines average outgoing long wave radiation, zonal wind at 850 hPa and 200 hPa from 15°S to 15°N [208].

2.2.4 Subseasonal to Seasonal (S2S) Hindcast Database

As a key component of the S2S Prediction Project, the S2S hindcast database offers a large number of hindcast cases to investigate the forecast skill and potential predictability at the extended time range. The database consists of extended-range hindcast cases implemented by 11 operational centers, namely:

- The Australian Bureau of Meteorology (BOM; [3])
- The China Meteorological Administration (CMA; [217])
- The European Centre for Medium-Range Weather Forecasts (ECMWF; [196])
- The Environment and Climate Change Canada (ECCC; [48])
- The Institute of Atmospheric Sciences and Climate of the National Research Council (ISAC-CNR; [114])
- The Hydrometeorological Centre of Russia (HMCR; [32])
- The Japan Meteorological Agency (JMA; [82])
- The Korea Meteorological Administration (KMA; [14])
- The Météo-France/Centre National de Recherche Meteorologiques (MétéoFrance; [199])
- The National Centers for Environmental Prediction (NCEP; [157])
- The U.K. Met Office (UKMO; [216])

Model configurations are listed in Table 2.1.

In each hindcast case, each model is initialized with realistic estimates of the atmosphere, land surface, and ocean states. After initialization, the model predicts the global weather evolution for a preset extent without any boundary constrains. It should be noted that there are large differences between the evaluated models when one considers the initialization

Table 2.1: Model configurations in the S2S hindcast database

Model	Time range(day)	Resolution	Hindcast frequency	Hindcast length	Ensemble Size
BoM	0-62	T47L17	1981-2013	6/month	33
CMA	0-60	T106L40	1994-2014	daily	4
ECCC	0-32	$0.45° \times 0.45°$ L40	1995-2014	weekly	4
ECMWF	0-46	Tco639/319 L91	past 20 years	2/week	11
HMCR	0-61	$1.1° \times 1.4°$ L28	1985-2010	weekly	10
ISAC-CNR	0-32	$0.75° \times 0.56°$ L54	1981-2010	every 5 days	1
JMA	0-33	Tl479/Tl319L100	1981-2010	3/month	5
KMA	0-60	N216L85	1991-2010	4/month	3
MétéoFrance	0-32	T255L91	1993-2014	2/month	15
NCEP	0-44	T126L64	1999-2010	day	4
UKMO	0-60	N216L85	1993-2015	4/month	3

strategy, dynamics core, parameterization schemes, resolution, ensemble generation scheme, hindcast extents, ocean and sea ice coupling, etc. This diversity may offer an opportunity to determine best practices for subseasonal predictions [195, 211]. In this study, daily precipitation hindcasts for the West Coast during the cold season (October to March) are used for the assessment.

2.3 Methodology

2.3.1 Evaluation Strategy

A basic fact in precipitation prediction is that the position, timing, and intensity of forecast diverge from reality as forecast lead time increases. A grid scale, day-to-day deterministic prediction generally holds little efficacy beyond the synoptic range, as state estimation errors accumulate through model's iterative computations. However, predictions might still have skill if assessed at regional scales or over a range of lead times.

Given this fact, the evaluations implemented here are carried out at both stringent and loosened spatiotemporal scales. The stringent scale evaluation refers to evaluating each grid's day-to-day prediction skill. In addition, the evaluation is also carried out for regional average predictions and predictions that span specific windows of lead time. The regional

average predictions are calculated by averaging predictions within the geographical divisions shown in Figure 2.1. Small deviations in predicting cyclone trajectories and their associated precipitation positions are likely to be averaged out in this way. Windows of lead time are defined following [226]: For a lead time of n days, the subsequent n days average precipitation prediction is evaluated. Thus, a *ndnd* evaluation refers to evaluating the prediction of the mean precipitation rate from $(n+1)$th day to $2n$th day. This strategy offers a fair comparison across a range of time scales from short range to extended range, since the deviation in predicting the timing of precipitation at longer lead time will be averaged out at wider evaluation windows. Using different spatial and temporal scales, the following four experiments are carried out:

1. Daily, grid-point scale evaluation: Evaluate the nth day prediction skill at each grid point, n ranges for the entire period of forecast. The overall skill for each climate division is calculated by averaging skill scores for all the grid points within this division.

2. Daily, regional scale evaluation: Evaluate daily regional average forecasts for each geographical division.

3. Variable temporal windows, grid-point scale evaluation: Evaluation is carried out at each grid point for various windows of lead time, following the strategy of [226].

4. Variable temporal windows, regional scale evaluation: For each geographical division, the regional average precipitation forecasts are evaluated for variable windows of lead time.

2.3.2 Skill Metrics

Deterministic Skill Metrics

Two deterministic skill metrics, namely the Pearson correlation coefficient (r) and the Nash-Sutcliffe model Efficiency Coefficient (NSE; [129]), are used to assess the performance of the

ensemble-mean forecasts. Their formulas are given as follows:

$$r = \frac{E[(P_{obser} - \overline{P_{obser}})(P_{simu} - \overline{P_{simu}})]}{\sigma_{P_{obser}} \sigma_{P_{simu}}} \tag{2.1}$$

$$\text{NSE} = 1 - \frac{\sum (P_{obser} - P_{simu})^2}{\sum (P_{obser} - \overline{P_{obser}})^2} \tag{2.2}$$

Here $\overline{P_{obser}}$ denotes mean value of the precipitation observations. P_{simu} denotes the ensemble mean prediction. Operator E denotes the expectation taken over all available samples, σ denotes standard deviation, r quantifies the linear correlation, NSE quantifies the relative magnitude of the mean square error compared to the climatology variance.

Probabilistic Skill Metrics

In general, forecasts beyond 10 days should no longer be considered to be deterministic [194]. Each ensemble member from the ensemble forecast system offers useful information in predicting the real-world weather evolution. To account for the information provided by each ensemble member, I also evaluate the probabilistic prediction skill based on all ensemble members. Here, the relative operating characteristics (ROC) score and the continuous ranked probability score (CRPS) are employed for probablistic evaluation.

The ROC score provides a complete summary of hit ratio and false alarm ratio for different observation intervals. To calculate the ROC score for each model, I construct a sample space that consists of all ensembles starting at different dates. For instance, for ECMWF, there are $1,482$ hindcast starts; each start has 11 ensemble members, so together there are $1,482 \times 11$ samples. For this sample space, the hit ratios and false alarm ratios for observation intervals of (x, ∞) (here x is set as 10 deciles of observation range) are calculated and scatter plotted (hit ratio on the vertical axis and false alarm ratio on the

horizontal axis) [44]. The points construct the ROC curve, which should be above the 1:1 line if the model has positive skill. The ROC score is defined as the area under the ROC curve. The closer the ROC score is to 1, the better. A no-skill forecast has an ROC score of 0.5 [193]. It should be noted that the ROC curve is constructed by sorting the elements of the joint distribution of observations and predictions. Thus, the actual numerical values are immaterial, and the final score is insensitive to prediction biases [214]. The score reflects the potential performance that can be achieved if the forecasts were correctly calibrated or bias corrected. It should be acknowledged that there are considerable biases that are introduce by the precipitation-related parameterization schemes. However, since the objective here is to investigate the potential precipitation prediction skills achieved by the dynamic modules of models, I believe applying the ROC metric score is justified.

The CRPS measures the ensemble forecast skill by comparing the probability distribution of the ensemble predictions and the observations [63]. As is shown in Figure 2.2, it is represented as the integrated squared difference between the cumulative probability distribution function (c.d.f.) of the forecasts and the observation.

To evaluate the general performance of the ensemble forecast systems, I apply the mean CRPS ($\overline{\text{CRPS}}$). The formulas are given as follows:

$$\begin{aligned} \overline{\text{CRPS}} &= \frac{1}{n}\text{CRPS} = \frac{1}{n}\int_R [F_{obser}(x) - F_{simu}(x)]^2 dx \\ &\approx \frac{1}{n}\hat{\text{CRPS}} = \frac{1}{n}\int_R [F_{obser}(x) - \hat{F}_{simu}(x)]^2 dx \end{aligned} \tag{2.3}$$

Here n represents the ensemble forecast case count, F_{obser}/F_{simu} is the c.d.f. of the precipitation observation/simulation as shown in Figure 2.2. $\hat{F}_{simu}$ could be easily estimated by assigning equal probability to each ensemble member.

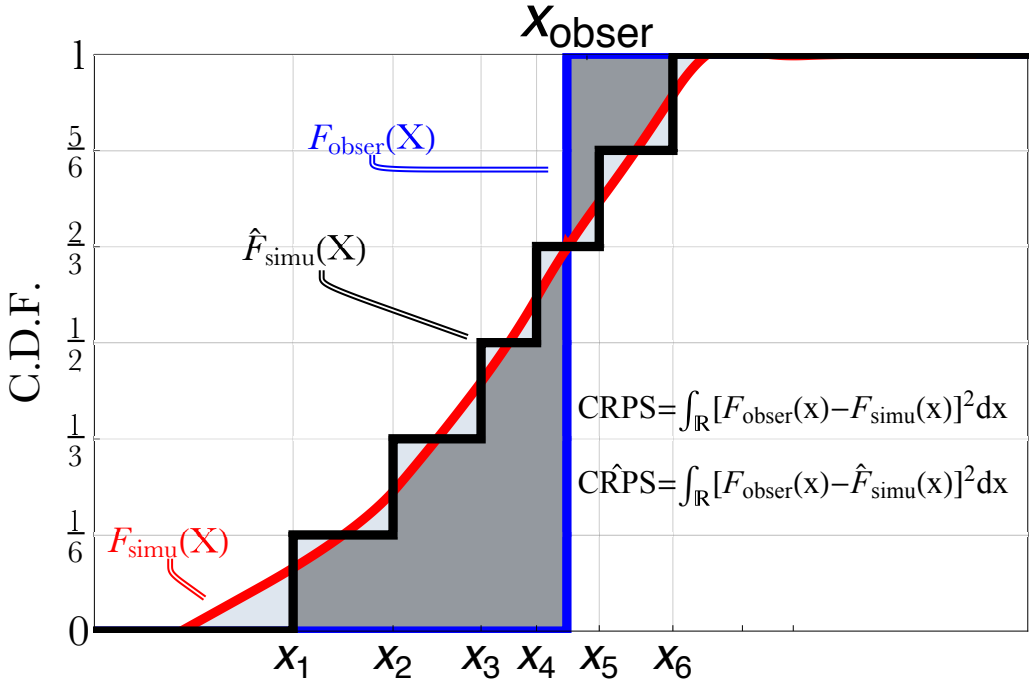


Figure 2.2: Describing CRPS using a 6-ensemble member forecast case (x_1, x_2, ..., x_6). The red/black line represents the theoretical/empirical c.d.f. of the precipitation forecast, which is denoted as $F_{simu}/\hat{F}_{simu}$; the blue line represents the c.d.f. of the observation, F_{obser}. Since the observation is deterministic, F_{obser} is in the Heaviside function form, i.e., if $x < x_{obser}$, $F_{obser}(x) = 0$, otherwise $F_{obser}(x) = 1$. The CRPS is defined as the integrated squared difference (shaded area) between the cumulative distribution functions of the forecasts and observations.

2.4 Evaluation Results

2.4.1 Deterministic Skills

Pearson Correlation Coefficient

The estimated Pearson correlation coefficient (r) between the ensemble mean predictions and the observations for the four experiments is presented in Figure 2.3.

For day-to-day evaluation (first two columns in Figure 2.3), as is expected, each model shows a rapid decrease of r skill with forecast lead time. The extent to which model's r skill is greater than 0.2 is labeled (this threshold is subjective and should be customized regarding specific application purposes). Generally, due to the model performance differences, r falls below 0.2 within 8 to 15 days (10 to 16 days) for Experiment 1 (Experiment 2). A comparison between Column 1 and Column 2 shows that with a lead time of as much as 2 weeks, regional average predictions generally have higher r skill compared to grid scale predictions. The skill improvements through spatial averaging are most obvious for SCA, which is attributable to the uneven precipitation distribution for this region.

For the temporal interval evaluation (Experiments 3 and 4 in the last two columns of Figure 2.3), the statistics of best and mean performances at different windows of lead time are given in Table 2.2. Within the synoptic range, the Day 2 ($1d1d$), Day 3−Day 4 ($2d2d$) and Day 5−Day 8 ($4d4d$) r skills are generally of the same order of magnitude (above 0.6 at grid scale and 0.7 at regional scale). This indicates that the decrease of prediction skill as lead time increases is compensated by the expanding of evaluation windows following the ndnd temporal averaging strategy. JMA, KMA, ECCC, and ECMWF models have the best performance at this temporal range. It is noteworthy that theses models are of higher resolution compared to the others. For Week 2 ($1w1w$), there is large variability in the models' r skills. The best performing model (ECMWF) achieves r skill of approximately 0.5 at grid scale and 0.6 at regional scales. The average performance for all models is of the order of 0.4 for both grid and regional scales. Beyond 2 weeks, the models generally show

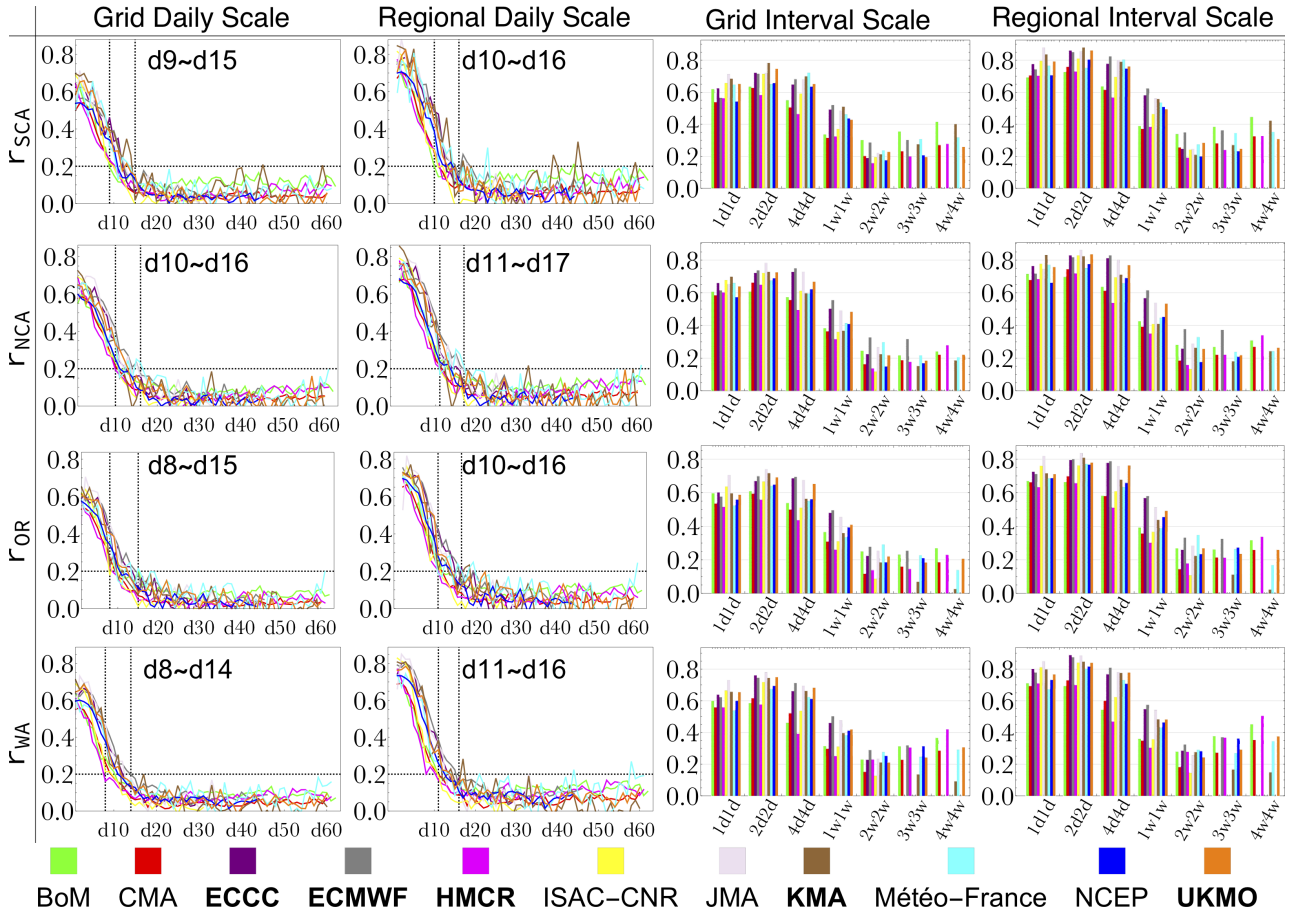


Figure 2.3: The Pearson correlation coefficient between the ensemble mean of precipitation predictions and the observations for the four experiments defined in Section 3. The evaluation results for the four divisions are shown in rows 1– 4. The columns represents different experiments. Column 1 shows the daily, grid-point scale evaluation results; column 2 shows the daily, regional scale evaluation results; column 3 shows the variable temporal windows, grid-point scale evaluation results; column 4 shows the variable temporal windows, regional scale evaluation results. For column 1 and 2, the extensions to which $r > 0.2$ for different models are labeled.

little usable skill. However, it is noteworthy that some models exhibit unexpectedly good performance at this time range, such as BoM for SCA and HMCR for WA.

Table 2.2: Correlation coefficient of precipitation predictions at temporal interval scales

Scale	SCA		NCA		OR		WA	
	Best	Mean	Best	Mean	Best	Mean	Best	Mean
Day2 Grid	0.71(JMA)	0.62	0.7(KMA)	0.63	0.71(JMA)	0.59	0.73(JMA)	0.62
Day2 Regional	0.88(JMA)	0.76	0.83(KMA)	0.74	0.82(JMA)	0.71	0.85(JMA)	0.76
Day3∼4 Grid	0.78(KMA)	0.69	0.78(JMA)	0.7	0.74(JMA)	0.66	0.78(JMA)	0.7
Day3∼4 Regional	0.88(KMA)	0.81	0.86(JMA)	0.79	0.84(JMA)	0.76	0.89(ECCC)	0.81
Day5∼8 Grid	0.72(MeteoFrance)	0.62	0.75(ECMWF)	0.63	0.7(ECMWF)	0.58	0.71(ECMWF)	0.6
Day5∼8 Regional	0.82(ECMWF)	0.73	0.83(ECMWF)	0.7	0.79(ECMWF)	0.67	0.81(ECMWF)	0.69
Week2 Grid	0.52(ECMWF)	0.43	0.56(ECMWF)	0.42	0.5(ECMWF)	0.38	0.5(ECMWF)	0.38
Week2 Regional	0.62(ECMWF)	0.5	0.61(ECMWF)	0.47	0.58(ECMWF)	0.44	0.58(ECMWF)	0.45
Week3∼4 Grid	0.3(BoM)	0.22	0.33(ECMWF)	0.22	0.29(MeteoFrance)	0.2	0.29(ECMWF)	0.22
Week3∼4 Regional	0.35(ECMWF)	0.26	0.38(ECMWF)	0.25	0.35(MeteoFrance)	0.24	0.32(ECMWF)	0.26
Week4∼6 Grid	0.35(BoM)	0.26	0.32(ECMWF)	0.2	0.25(ECMWF)	0.19	0.32(ECMWF)	0.26
Week4∼6 Regional	0.38(BoM)	0.29	0.37(ECMWF)	0.24	0.33(ECMWF)	0.24	0.38(BoM)	0.31
Week5∼8 Grid	0.41(BoM)	0.32	0.28(HMCR)	0.23	0.27(BoM)	0.18	0.42(HMCR)	0.29
Week5∼8 Regional	0.45(BoM)	0.36	0.34(HMCR)	0.28	0.34(HMCR)	0.23	0.51(HMCR)	0.36

The results above suggest that models' r skills are distinct regarding different regions and forecast lead time. For a same region and lead time, the informative predictable range may differ by up to 6−7 days due to the model performance differences. The huge sample size in the S2S dataset offers opportunity to test the significance of model performance differences at critical forecast lead time periods. The results would benefit model selections for practical forecasts and multi-model ensemble predictions. Below I carry out the significance test on models' r skill differences for the Day 2, Day 7, Week 2, and Week 3−4 period. These periods are selected since they represent critical lead time and scales in weather forecast. To perform the test, I first applied the Fisher r-to-z transformation [45] on the r estimations. Later, I applied significance test on the z statistics to assess the significance of the difference between models' r skills. Results are shown in Figure 2.4.

For the Day 2 forecast (row 1 in Figure 2.4), the ECCC, ECMWF, ISAC-CNR, JMA, and KMA generally show significant advantages over the other models; while the BoM and CMA model show significant lower skills. For the Day 7 forecast (row 2), the ECCC, ECMWF, JMA and KMA still lead the performance, while ISAC-CNR loses its advantage over most

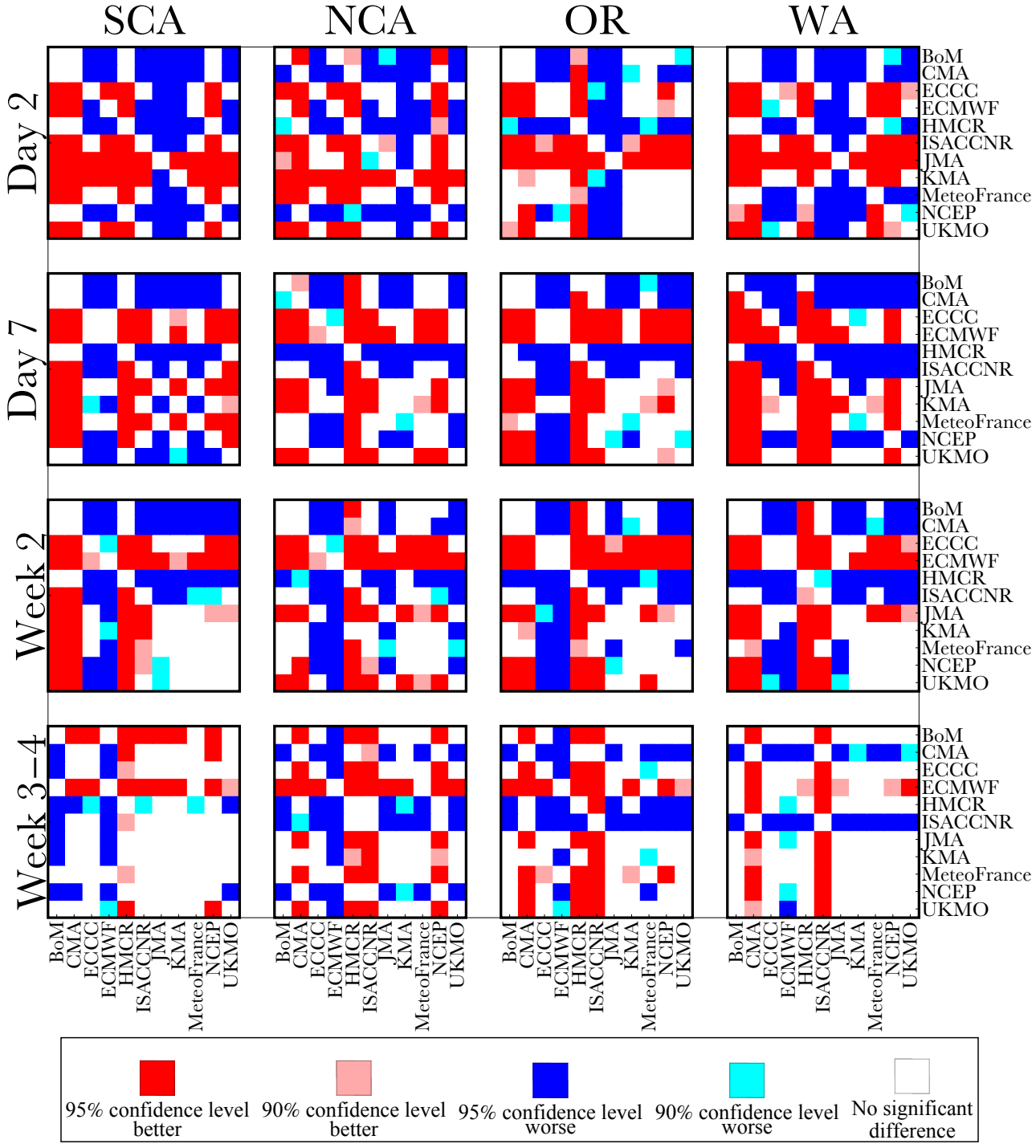


Figure 2.4: Significance test result for regional average predictions in the four divisions for Day 2, Day 7, Week 2, and Week 3–4. The rows represent the forecast period; the columns represent the geographic divisions. For each matrix, the grid at row m, column n is labeled red if the r skill for the model at the row m is better than the model at the column n at 95% confidence level, similar denotation for other colored labels.

models. This might due to the fact that the ISCA-CNR model is applying deterministic rather than ensemble forecast here. For Week 2 forecast (row 3), the best performing models are ECCC, ECMWF, and JMA. ECMWF shows significant advantage over all the other models except for the WA prediction when compared against JMA. For Week 3–4 forecast (row 4), although there is essentially no useful r skill, the ECMWF model still shows advantage over the rest models.

Nash Sutcliffe Model Efficiency

A large value of r generally results in a corresponding positive NSE, which indicates that the model outperforms the baseline of climatology. However, evaluation of model's precipitation ensemble mean predictions shows negative NSE in most experiments, indicating that models have skills in predicting precipitation variations but have difficulty in pinpointing the specific precipitation amount. Given this fact, a simple linear bias correction is carried out for each scale before evaluating with NSE. Results after the correction are shown in Figure 2.5. The best and mean performance are given in Table 2.3.

Table 2.3: Nash-Sutcliffe model efficiency of precipitation predictions (linearly bias corrected) at temporal interval scales

Scale	SCA		NCA		OR		WA	
	Best	Mean	Best	Mean	Best	Mean	Best	Mean
Day2 Grid	0.51(JMA)	0.39	0.49(KMA)	0.41	0.5(JMA)	0.35	0.54(JMA)	0.4
Day2 Regional	0.77(JMA)	0.59	0.69(KMA)	0.54	0.67(JMA)	0.5	0.72(JMA)	0.58
Day3∼4 Grid	0.61(KMA)	0.48	0.62(JMA)	0.5	0.55(JMA)	0.44	0.61(JMA)	0.49
Day3∼4 Regional	0.77(KMA)	0.65	0.74(JMA)	0.63	0.7(JMA)	0.58	0.79(ECCC)	0.66
Day5∼8 Grid	0.52(MeteoFrance)	0.4	0.57(ECMWF)	0.41	0.49(ECMWF)	0.35	0.51(ECMWF)	0.37
Day5∼8 Regional	0.68(ECMWF)	0.54	0.69(ECMWF)	0.5	0.62(ECMWF)	0.45	0.65(ECMWF)	0.49
Week2 Grid	0.27(ECMWF)	0.19	0.31(ECMWF)	0.19	0.25(ECMWF)	0.15	0.25(ECMWF)	0.16
Week2 Regional	0.39(ECMWF)	0.25	0.38(ECMWF)	0.23	0.34(ECMWF)	0.2	0.33(ECMWF)	0.21
Week3∼4 Grid	0.09(BoM)	0.05	0.11(ECMWF)	0.05	0.09(MeteoFrance)	0.05	0.09(ECMWF)	0.05
Week3∼4 Regional	0.12(ECMWF)	0.07	0.14(ECMWF)	0.07	0.12(MeteoFrance)	0.06	0.11(ECMWF)	0.07
Week4∼6 Grid	0.13(BoM)	0.07	0.11(ECMWF)	0.05	0.07(ECMWF)	0.04	0.1(ECMWF)	0.08
Week4∼6 Regional	0.15(BoM)	0.09	0.14(ECMWF)	0.06	0.11(ECMWF)	0.06	0.14(BoM)	0.1
Week5∼8 Grid	0.18(KMA)	0.11	0.09(HMCR)	0.06	0.08(BoM)	0.05	0.19(HMCR)	0.1
Week5∼8 Regional	0.2(BoM)	0.13	0.12(HMCR)	0.08	0.11(HMCR)	0.06	0.26(HMCR)	0.14

Overall, the NSE results offer similar indications as evaluation results based on r. Daily scale NSE reaches 0.2 within approximately 5 to 10 days. Models lose their advantage over

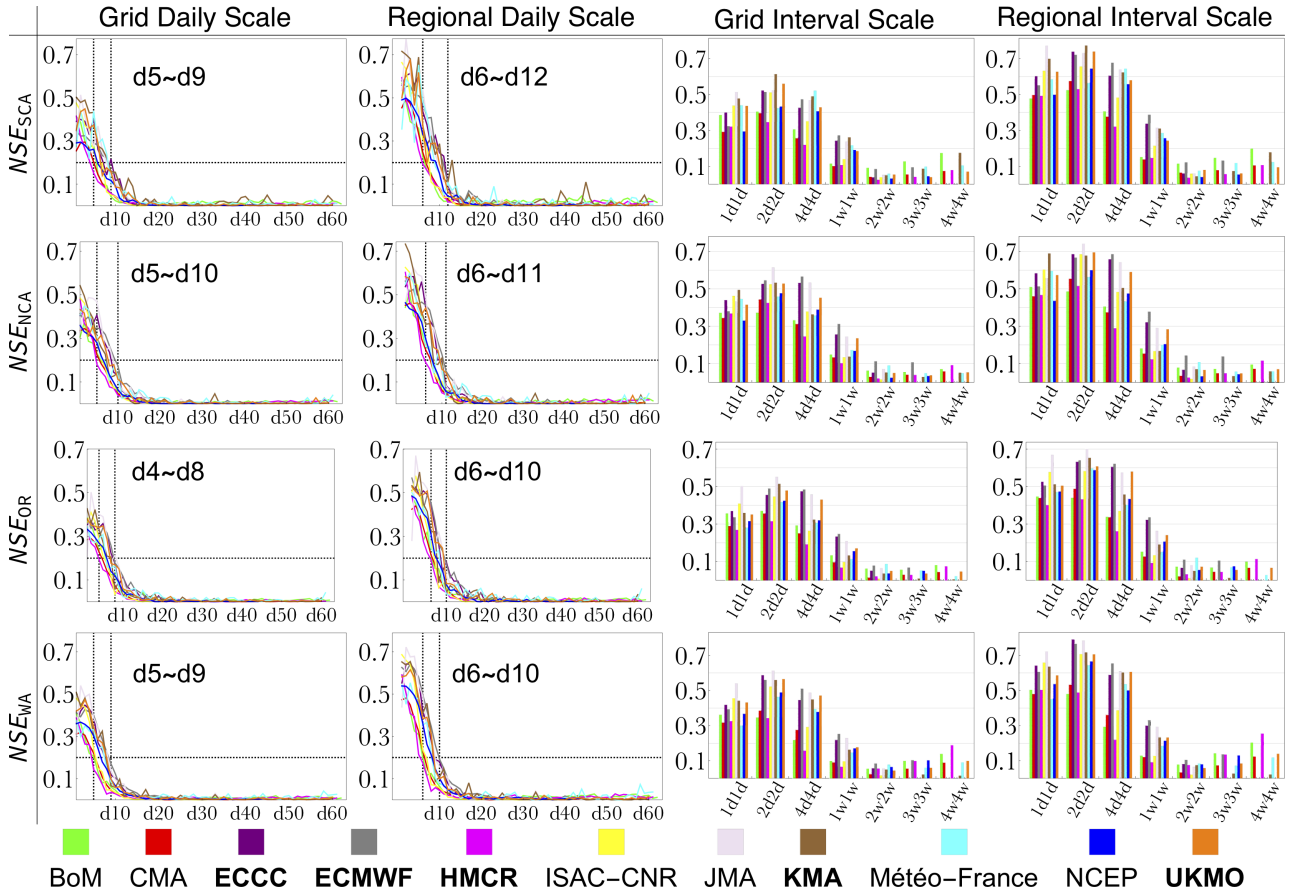


Figure 2.5: As in Figure 2.3, but using Nash-Sutcliffe Efficiency of ensemble mean predictions (linear bias corrected). For day-to-day evaluation, the extension to which models have NSE$>$ 0.2 is labeled.

.

climatology after approximately 2 weeks (NSE$\approx$ 0). Through spatial averaging, short range NSE could be improved by 0.2, and range of NSE$>$ 0.2 is extended by 1 day.

Considering evaluations at different windows of forecast lead time, for Day 2 and Day 3-4 predictions, JMA, KMA, and ECCC models still have the best performance. For the medium range, ECMWF achieves the largest NSE. Week 2 predictions are of considerable value, with NSE around 0.25 at grid scale and 0.35 at regional scale. Beyond two weeks, skills rapidly decrease.

2.4.2 Probabilistic Skill

ROC Score

This section evaluates the models' probabilistic skill using the ROC score. To illustrate the idea of ROC, I draw the ROC curves for ECMWF regional predictions in Figure 2.6. Each labeled point represents the hit ratio (HR) and false alarm ratio (FAR) for a corresponding interval of precipitation observations. For instance, the point labeled 0.9 represents the (HR,FAR) of prediction for the $(P_{90\%}, \infty)$ interval, where $P_{90\%}$ represents the 90% quantile of observed precipitation. The further a point is above the 1:1 line, the more likely a $P > P_{90\%}$ forecast is true. As is shown, points in all sub-figures here are above the $1:1$ line, showing a larger HR than FAR at different evaluating thresholds for all scales. Small precipitation cases generally appear on the top right part of the ROC curve (larger HR but also larger FAR), while large precipitation cases appear on the bottom left part of the ROC curve (smaller HR but also smaller FAR). For short to medium range, the ROC curves of Day 2 ($1d1d$), Day 3$-$Day 4 ($2d2d$) and Day 5$-$Day 8 ($4d4d$) show considerable overlap. The ROC curve for Week 2 ($1w1w$) prediction falls below the previous three cases. For Week 3$-$Week 4 ($2w2w$) and Week 4$-$Week 6 ($3w3w$), again the curves overlap and they fall below all the previous cases. The ROC score is defined as the area below the ROC curve, which summarizes the probabilistic prediction skill for different evaluating intervals. The scores are given at bottom right in each sub-figure. Day 3$-$Day 4 prediction achieves slightly better

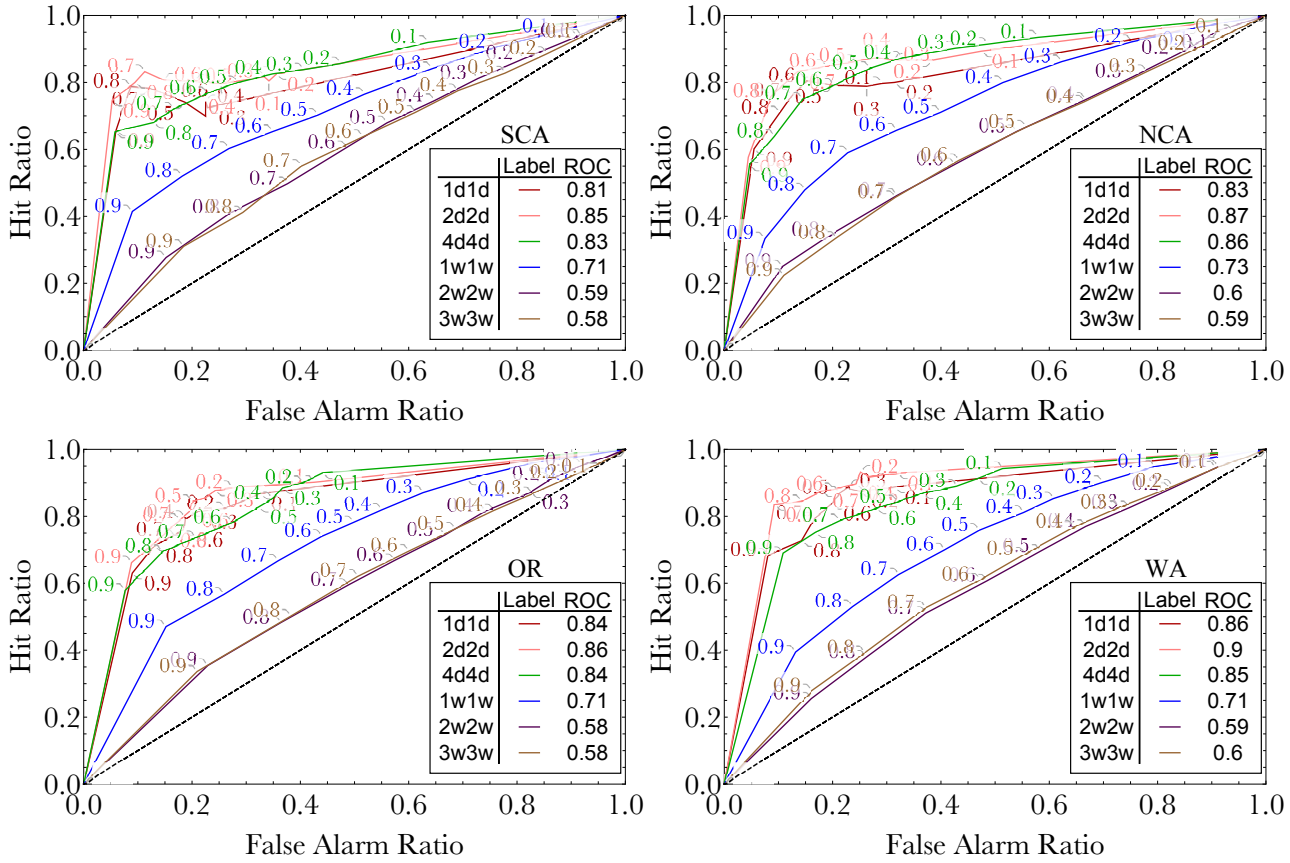


Figure 2.6: ROC curves for ECMWF regional precipitation predictions for the four regions at various windows of lead time. The points with labeled numbers show the hit ratio and false alarm ratio of a corresponding threshold. The ROC scores for different intervals are given in the tables.

ROC score compared to Day 2 and Day 5−Day 8. The Week 2 ($1w1w$) prediction achieves ROC score above 0.7 in all divisions. For Week 3−Week 4 ($2w2w$) and Week 4−Week 6 ($3w3w$), the score is around 0.6, showing better performance than random guess.

Based on this same approach, I calculated the ROC scores for all models in different evaluation experiments. Results are shown in Figure 2.7. For day-to-day evaluation (first two columns), I labeled the extent to which the ROC score is larger than 0.6 (again, this threshold is subjective and should be adjusted if necessary). Generally, the daily ROC scores begin to fall below 0.6 in the second week. They reach 0.5 at approximately 20 days, which means beyond 20 days, day-to-day estimations show no advantage over climatology.

Considering evaluations at different windows of lead time (last two columns in Figure

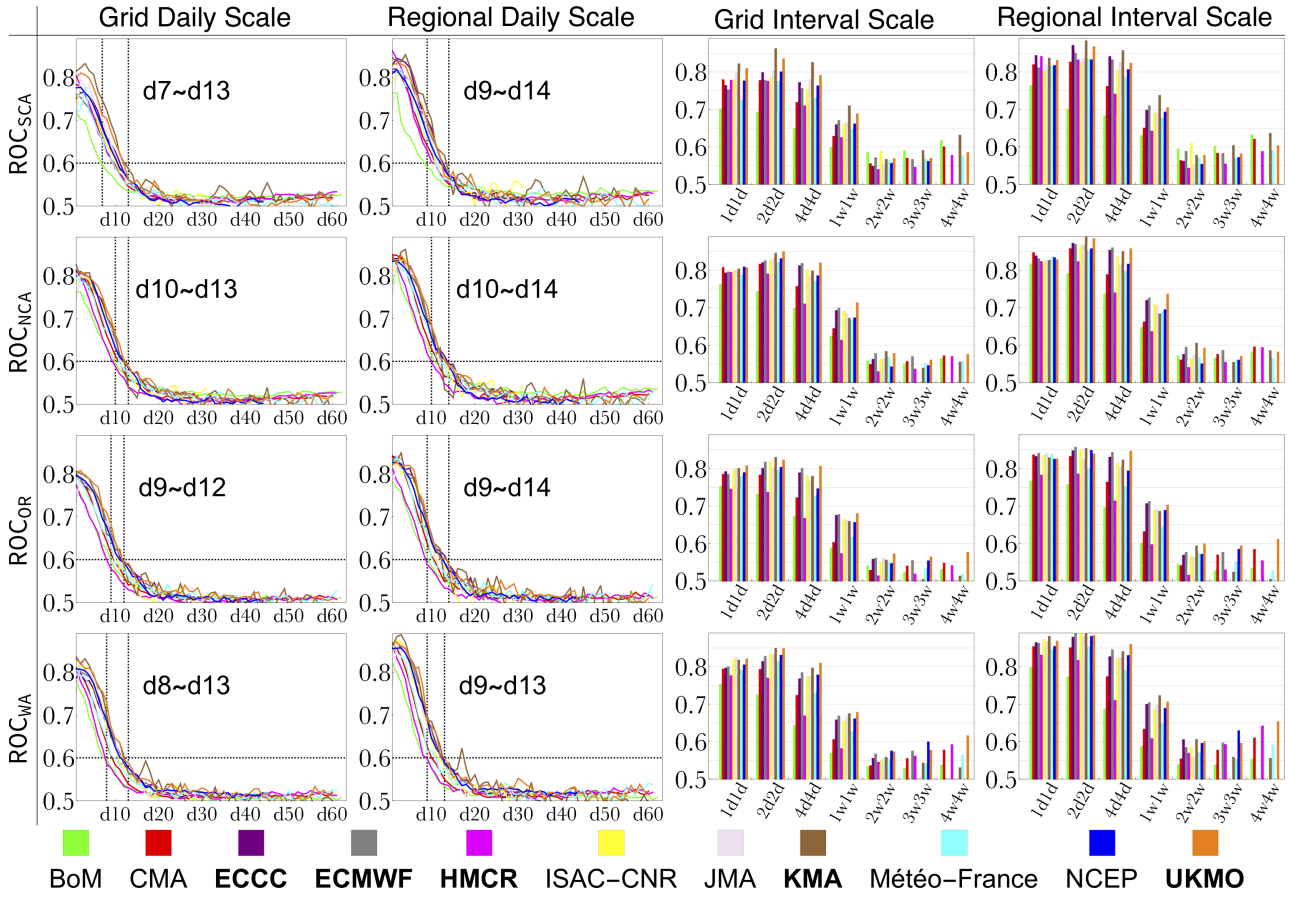


Figure 2.7: As in Figure 2.3, but using the ROC score of ensemble precipitation predictions. For day-to-day evaluation, the extension to which models have ROC score > 0.6 is labeled.

2.7), for short to medium range, best-performing models achieve ROC scores above 0.85 for Day 2 ($1d1d$), Day 3−Day 4 ($2d2d$) and Day 5−Day 8 ($4d4d$). For Week 2, the ROC score is around 0.65 for grid scale and 0.7 for regional scale. Models still hold positive probabilistic prediction skills beyond Week 2. The best and mean performance are summarized in Table 2.4.

Table 2.4: ROC score at temporal interval scales

Scale	SCA		NCA		OR		WA	
	Best	Mean	Best	Mean	Best	Mean	Best	Mean
Day2 Grid	0.82(KMA)	0.77	0.81(NCEP)	0.8	0.81(UKMO)	0.79	0.83(JMA)	0.8
Day2 Regional	0.84(ECCC)	0.82	0.85(CMA)	0.83	0.84(JMA)	0.82	0.88(KMA)	0.86
Day3∼4 Grid	0.86(KMA)	0.79	0.85(UKMO)	0.82	0.83(KMA)	0.8	0.85(KMA)	0.81
Day3∼4 Regional	0.88(KMA)	0.83	0.89(KMA)	0.86	0.86(ECMWF)	0.83	0.9(KMA)	0.87
Day5∼8 Grid	0.83(KMA)	0.75	0.82(UKMO)	0.78	0.81(UKMO)	0.75	0.81(UKMO)	0.75
Day5∼8 Regional	0.86(KMA)	0.8	0.86(ECMWF)	0.81	0.85(UKMO)	0.79	0.86(UKMO)	0.8
Week2 Grid	0.71(KMA)	0.66	0.71(UKMO)	0.67	0.68(UKMO)	0.64	0.68(UKMO)	0.64
Week2 Regional	0.74(KMA)	0.69	0.74(UKMO)	0.69	0.71(ECMWF)	0.67	0.72(KMA)	0.67
Week3∼4 Grid	0.59(ISACCNR)	0.56	0.58(KMA)	0.56	0.57(UKMO)	0.55	0.58(NCEP)	0.56
Week3∼4 Regional	0.61(ISACCNR)	0.57	0.61(KMA)	0.57	0.6(UKMO)	0.57	0.61(KMA)	0.58
Week4∼6 Grid	0.59(KMA)	0.57	0.57(ECMWF)	0.55	0.56(UKMO)	0.54	0.6(NCEP)	0.56
Week4∼6 Regional	0.6(KMA)	0.58	0.59(ECMWF)	0.57	0.59(UKMO)	0.56	0.63(NCEP)	0.58
Week5∼8 Grid	0.63(KMA)	0.6	0.58(UKMO)	0.57	0.58(UKMO)	0.54	0.62(UKMO)	0.57
Week5∼8 Regional	0.64(KMA)	0.61	0.6(CMA)	0.58	0.61(UKMO)	0.55	0.66(UKMO)	0.6

Continuous Ranked Probability Score

Compared to the previous skill metrics of r, NSE, and ROC score, which are dimensionless and roughly bounded within certain ranges, the CRPS is of the same dimension with the predictand [63, 198]. It has a lower bound of 0, but is not restricted by an upper bound. Due to these characteristics, evaluations using $\overline{\text{CRPS}}$ at different spatiotemporal scales are sensitive to the distribution variation of the predictand. Thus, the results for difference scales cannot be directly compared.

Figure 2.8 shows the $\overline{\text{CRPS}}$ evaluation results for the four experiments that are proposed in the paper.

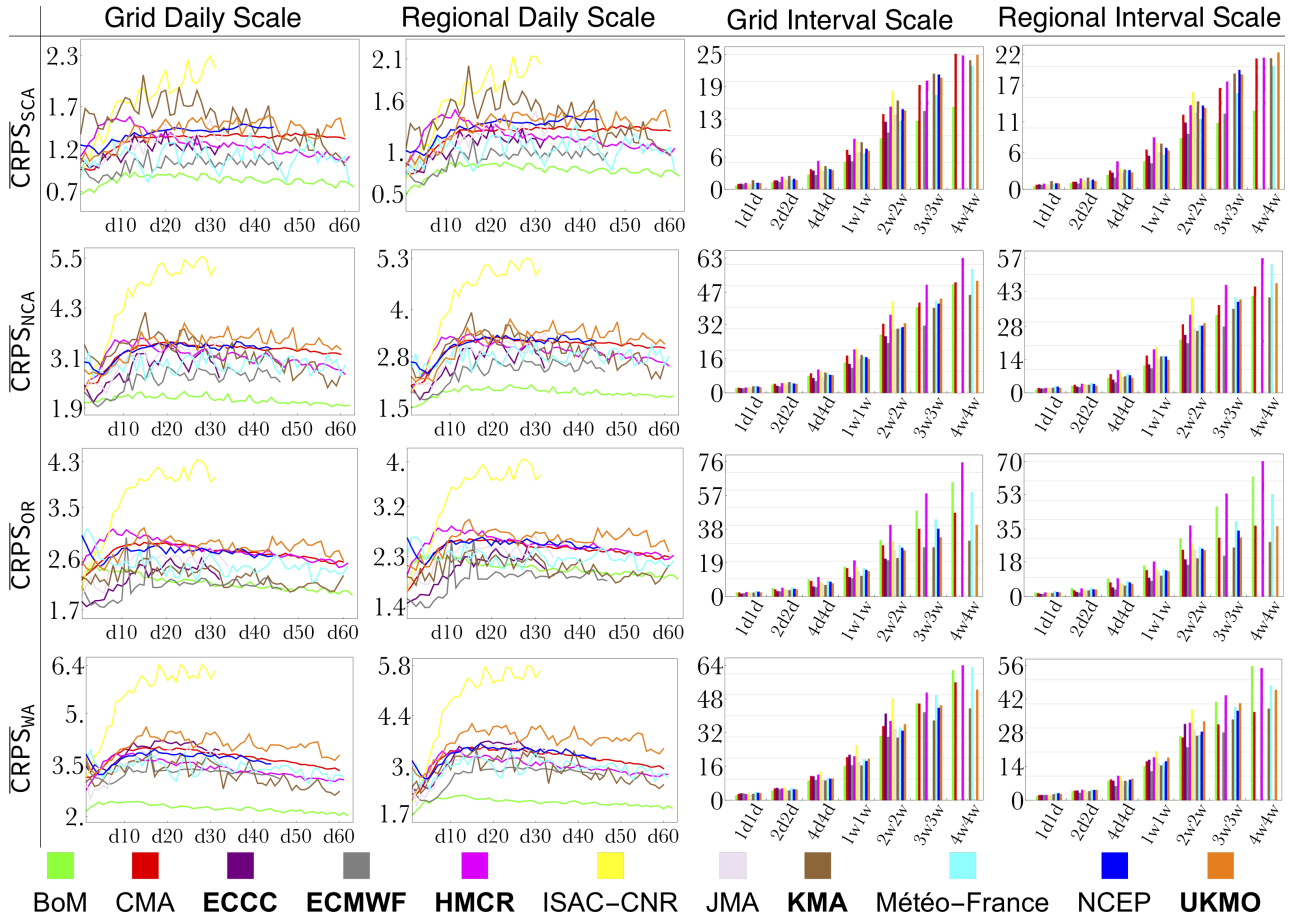


Figure 2.8: The $\overline{\text{CRPS}}$ for the four experiments defined in Section 3 of the paper. The evaluation results for the four divisions are shown in row 1– row 4. The columns represents different experiments. Column 1 shows the daily, grid-point scale evaluation results; column 2 shows the daily, regional scale evaluation results; column 3 shows the variable temporal windows, grid-point scale evaluation results; column 4 shows the variable temporal windows, regional scale evaluation results.

Considering the $\overline{\text{CRPS}}$ evaluation results at different windows of lead time, larger $\overline{\text{CRPS}}$s are achieved for larger evaluation time windows. I argued that this relationship is due to the fact that the $\overline{\text{CRPS}}$ shares the same dimension with the predictand and is scaled up as the aggregated precipitation amount increases with time window width. To illustrate this point, I re-scale the $\overline{\text{CRPS}}$ by the time window width, results are shown in Figure 2.9. I found no apparent skill variation for different evaluation windows.

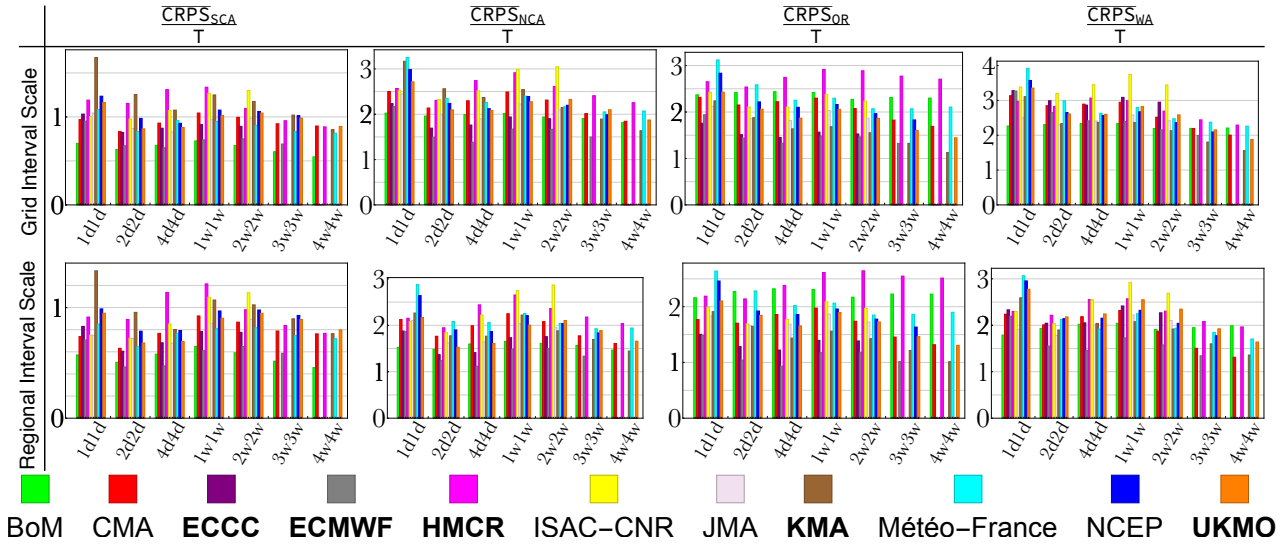


Figure 2.9: Re-scaled $\overline{\text{CRPS}}$ for the temporal interval evaluation. T represents the evaluation time window width. The $\overline{\text{CRPS}}$s are re-scaled by T to roughly account for the magnitude variation of the predictand in different evaluation experiments.

For the day-to-day evaluation, as can be expected, all models show increase of $\overline{\text{CRPS}}$ with forecast lead time until $\overline{\text{CRPS}}$ becomes relatively steady after approximately 2 weeks. Compared to the evaluation results using r, NSE, and ROC score, the $\overline{\text{CRPS}}$-forecast lead day curves show considerable oscillation. By examining the mean precipitation rate on different forecast lead days, I confirmed that the oscillations can be attributed to the variation of precipitation rates on these days. Considering the model performance differences, the BoM model shows best (lowest) $\overline{\text{CRPS}}$ in most daily evaluation cases; also, for longer forecast lead time, the $\overline{\text{CRPS}}$ for BoM does not increase as significantly as the other models. By examining the model configuration, I found that the BoM precipitation product is of lower spatial resolution compared to the rest models. Since evaluations using other skill scores suggest no significant advantage of BoM, I believe the low $\overline{\text{CRPS}}$ for BoM is due to its low spatial resolution rather than advantageous performance. On the other hand, the ISAC-CNR model shows significant worse (higher) $\overline{\text{CRPS}}$. This is because the ISAC-CNR model provides deterministic forecast with single ensemble member, while the other models have multiple ensemble members. Results here suggest the advantage of applying ensemble forecast rather than deterministic forecast, especially for the extended range period. For the rest 9 models,

whose precipitation products are of same spatial resolution, the $\overline{\text{CRPS}}$-forecast lead day curve show considerable overlap, with ECMWF model showing slightly better performance.

Considering the evaluation results at different windows of lead time, larger $\overline{\text{CRPS}}$s are achieved for larger evaluation time windows. This is because the $\overline{\text{CRPS}}$ shares the same dimension with the predictand and is scaled up as the aggregated precipitation amount increases with forecast time window width. In supplementary materials, I show the re-scaled $\overline{\text{CRPS}}$ by dividing the $\overline{\text{CRPS}}$ with corresponding time window width. Results suggest no apparent skill variation for different evaluation windows.

Overall, the $\overline{\text{CRPS}}$ evaluation results confirm the advantage of applying ensemble forecast rather than deterministic forecast, especially for the extended range. On the other hand, the scaling issue of the $\overline{\text{CRPS}}$ restricts us from comparing model performance differences at different spatiotemporal scales.

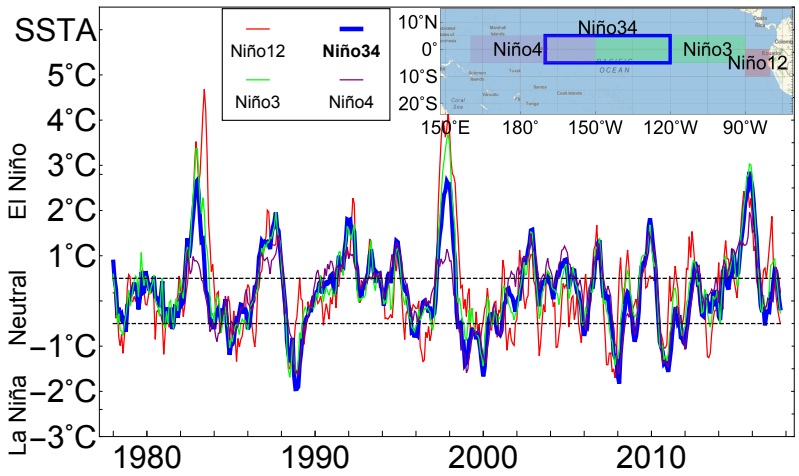
2.5 Impacts of ENSO and MJO

The evaluation results in the previous section show a sharp drop in prediction skill after Week 1. Beyond this time range, predictions rely heavily on the existence of sources of predictability and the model's ability to realize them for informative predictions. In this section, I explore the impact of key sources of intraseasonal to seasonal predictability on precipitation distribution and prediction skill at the extended range. The particular focus is put on ENSO and the MJO.

2.5.1 ENSO

ENSO is a semi-periodic variation in winds and sea surface temperatures over the tropical eastern Pacific Ocean. Typically, ENSO is quantified by specific regionsa's sea surface temperature anomaly, as shown in Figure 2.10.

ENSO influences the seasonal variability across the tropical Pacific and in much of the



Model	El Niño/La niña	El Niño	La niña	Neutral	All
BoM	744	360	384	444	1188
CMA	2488	1065	1423	1334	3822
ECCC	497	192	305	283	780
ECMWF	1004	440	564	478	1482
HMCR	1088	550	538	569	1657
ISACCNR	699	363	336	381	1080
JMA	348	180	168	192	540
KMA	152	76	76	88	240
MeteoFrance	172	78	94	92	264
NCEP	1579	640	939	605	2184
UKMO	368	180	188	184	552

Figure 2.10: Definition and time series of ENSO. ENSO is quantified based on the SSTA of certain Pacific tropical regions, as delineated on the top right. The time series plot for Niño indexes from 1978 to 2016 is shown. To investigate ENSO impact on extended range prediction skill, hindcasts for each model are clustered into different groups based on the ENSO phase at model start time. Case counts for each cluster are listed in the right table.

extratropics, including the North Pacific and North America. Precipitation anomalies in the regions along the West Coast are frequently related to ENSO through its influence on the Aleutian Low [17, 162] and subtropical jets [148, 188]. The connection has been investigated extensively using observations [162, 85], models [200, 35], and composite approaches [164, 111]. However, results suggest that the connections, such as the magnitude and sign of precipitation anomalies, are not robust [222].

Next I explore ENSO's impact on the precipitation distribution and extended-range prediction skill. I focus on evaluations at the weekly scale, as this is in accordance with the rough time scale of the life cycle for cyclone events.

Influence on Precipitation Distribution

In order to investigate the weekly precipitation statistics conditioned on the ENSO phases, I first constructed the winter time weekly precipitation time series by averaging precipitation records for consecutive 7-day windows (Day 1 to Day 7, Day 2 to Day 8, etc.). Next, I removed the impact of seasonal cycle by subtracting the leading two harmonics of the weekly

series, using Fast Fourier Transform [206]. Finally, I constructed the empirical distribution of weekly precipitation anomalies conditioned on different ENSO phases for early winter season (October, November, and December, denoted as OND) , late winter season (January, February, and March, denoted as JFM) , and the entire winter season (October to March, denoted as O−M). The distributions were estimated using the smooth kernel method, built on $5,000$ bootstrap samples. Samples were randomly selected (with replacement) from weekly precipitation anomaly time series for the corresponding season and ENSO phase. I applied the Kolmogorov-Smirnov test to determine whether the distributions differ significantly due to ENSO phase variations.

Figure 2.11 shows the empirical distributions of weekly precipitation anomalies. In early winter (OND), the El Niño phases tend to favor negative precipitation anomalies, while La Niña phases tend to favor positive precipitation anomalies. This pattern is more obvious for NCA, OR, and WA, as compared to SCA. SCA has a higher probability of receiving abnormally high precipitation events during La Niña, as shown in the tail of the distribution. For NCA, there tends to be more precipitation during neutral phases, on average. For OR and WA, there tends to be more precipitation during La Niña.

In late winter (JFM), with the onset of the rainy season in SCA, the influence of ENSO is flipped and strengthened. For El Niño (La Niña) phases, SCA receives 0.35 mm/day more (0.3 mm/day less) precipitation than climatology. For NCA, El Niño phases also receive more precipitation, but the negative influence of La Niña is not as obvious as it is for SCA. For OR, and WA, like early winter, La Niña phases receive more precipitation compared to climatology; however, El Niño is not accompanied by less precipitation on average.

The last column shows the weekly precipitation anomaly distribution for the entire winter season. As a summary of the two cases analyzed above, it is noted that during El Niño years, SCA tends to receive more precipitation on average, while NCA, OR, and WA tend to receive less precipitation. During La Niña years, SCA receives less precipitation while the others receive more. However, the variances are considerably large, making climatology a less robust

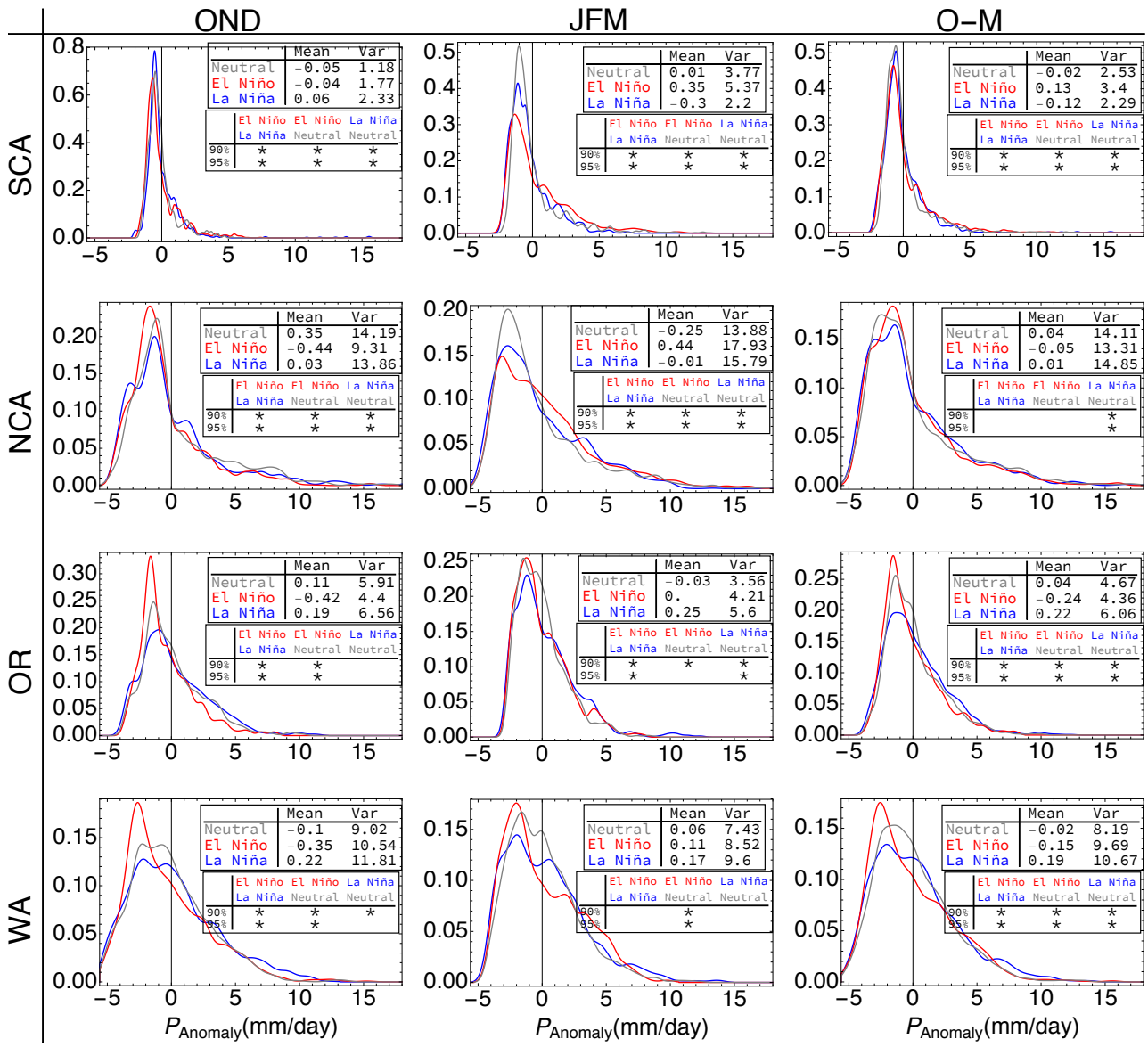


Figure 2.11: Distribution of weekly precipitation anomalies conditioned on ENSO phases. The first column is for early winter season (October, November, and December, OND); the second column is for late winter season (January, February, and March, JFM); the third row is for the entire winter season (October to March, O–M). The rows represent results for different geographic divisions. For each subfigure, I listed the mean and variance of the distribution conditioned on ENSO phases. The comparison between two distributions is labeled with asterisk if the Kolmogorov-Smirnov statistic lies out of the 90%(95%) confidence interval, indicating the two distributions are statistically significantly different.

reference for long-term prediction.

Influence on Precipitation Prediction Skills

The previous section discussed ENSO's influence on weekly precipitation statistics. Ideally, we would like to see these effects to be well simulated in the GCMs. If so, then realistic representation of ENSO in models could facilitate useful boundary forcings in conditioning the precipitation distribution, and predictions at the extended range could be improved.

To examine the influence of ENSO on extended-range prediction skill, I first clustered the hindcast cases by the ENSO phase at the model's start time. Next, I computed and compared the prediction skills for the different clusters. Four clusters were adopted here, namely the ENSO active phase (El Niño/La Niña), El Niño phase, La niña phase, and neutral phase. Their union was also evaluated as a reference. Sample sizes are listed in Figure 2.10. Using the Fisher r-to-z transformation [45], the correlation skills are transformed to the z statistics, which are then applied to assess the significance of the difference between two r skills.

I focus on the regional average predictions for Week 2 and Week 3−4. Figure 2.12 shows the Week 2 and Week 3−4 r skills conditioned on different ENSO phases. The NSE skill score and ROC score show similar results and are not presented here.

For Week 2 forecast (column 1 of Figure 2.12), in SCA, all 11 models show improved r skill during El Niño phase compared to during La Niña or ENSO-neutral phase, with 8 of them showing 90% level significant better skill during El Niño. Particularly, for the better-performing models (i.e., ECCC and ECMWF), r could differ by up to 0.2 comparing El Niño prediction and La Niña prediction. Correspondingly, for OR, most models show better r skill during La Niña phase compared to during El Niño phase, with 7 of them showing 90% level significant better skill during La Niña. For ECCC and ECMWF, r could differ by up to 0.2 comparing La Niña prediction and El Niño prediction. For NCA and WA, the results for ENSO phase do not agree between the models, although for NCA, the better performing

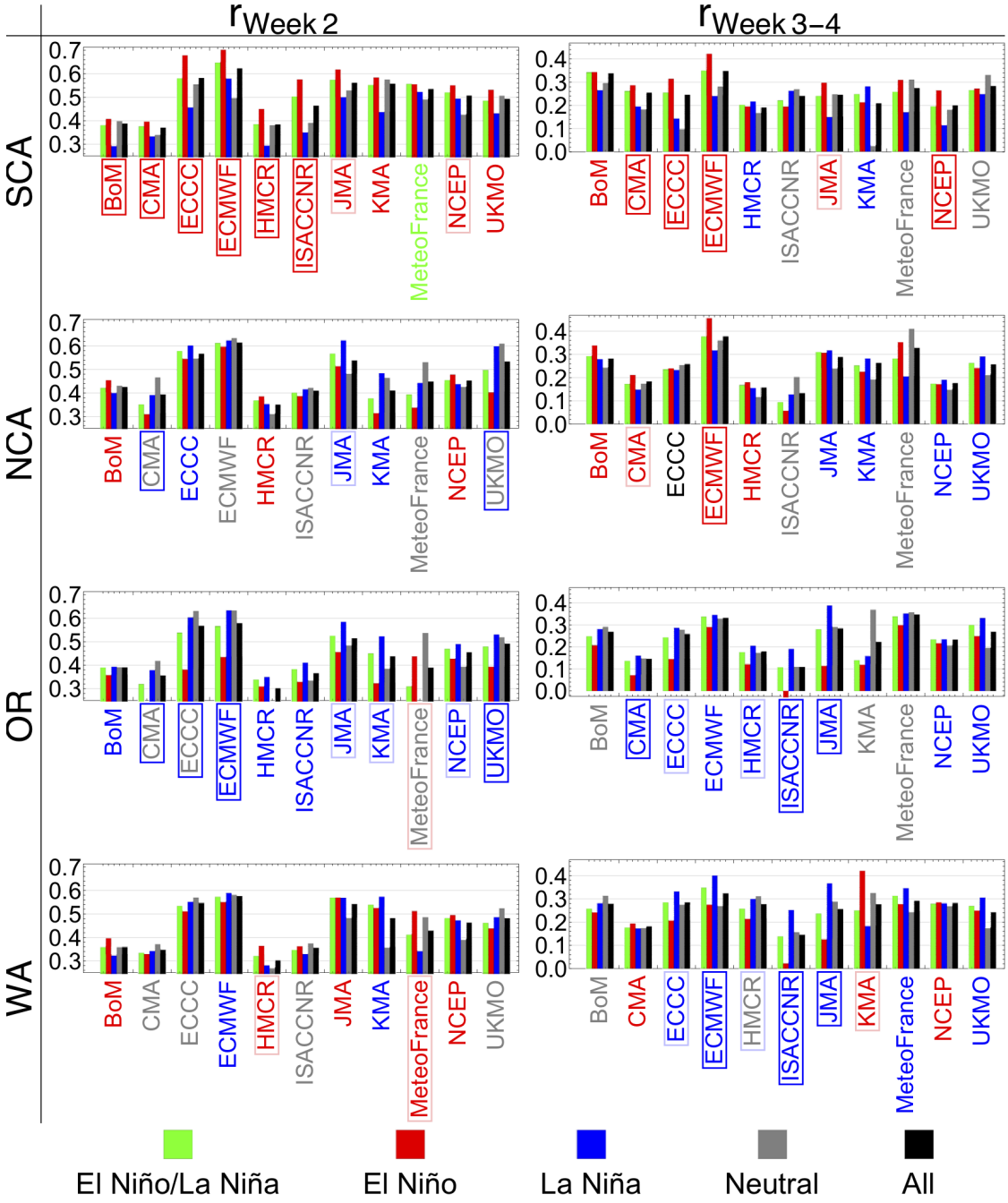


Figure 2.12: Week 2 (column 1) and Week 3−4 (column 2) precipitation prediction skills for different ENSO phases. The rows represent results for the four geographic divisions. For each sub-figure, 11 models are evaluated; each model is colored depending on the phase in which model has highest score. Models with significant r skill difference between El Niño and La Niña phase are framed : Red (Blue) frame indicates that model shows significantly better r skill for El Niño (La Niña) phase based on the z test. Light (Dark) frame indicates the difference is statistically significant at 90% (95%) confidence level.

models generally have higher r skill scores during La Niña phase.

For Week 3−4 forecast (column 2 of Figure 2.12), in SCA, more of the models have improved performance during El Niño than during La Niña, with 5 of them showing 90% level significant better r skill during El Niño. For ECMWF, r is 0.42 for El Niño, while for La Niña, r is 0.24. The advantage for ECMWF also occurs for NCA, with r being 0.46 during El Niño phase. This indicates that the El Niño phase might allow better extended-range prediction for California. For OR and WA, most of the models show improved r skill during La Niña when compared to El Niño, with 5 of them showing 90% level significant better r skill during La Niña.

2.5.2 MJO

The MJO is a traveling pattern characterized by a coherent eastward-propagating perturbation over the tropical Indian and Pacific Oceans [112]. Previous studies have found that MJO-related variability in the tropical Pacific convection modifies West Coast precipitation regimes through the propagation of extratropical wave trains [124, 65, 19]. Provided with realistic representation of the MJO and its teleconnections, this forecast opportunity could be realized as improvements in model's extended-range prediction skills. However, it is also noted that poor representation of the MJO results in systematically worsened forecast during the MJO active periods, as compared to predictions during quiescent periods [29, 62].

The prediction skill for the MJO and its teleconnections have been improved significantly in recent decades, reaching a useful forecast range beyond 20 days [88, 201, 100] and producing realistic teleconnections with large scale circulation [197]. Here, I make a dedicated examination on how MJO modifies precipitation distribution and extended-range prediction skills for the West Coast.

The RMM index is employed to quantify the MJO. As shown in Figure 2.13, the RMM index is composed of the two leading principal components (PCs) of the field that combines average outgoing long-wave radiation and zonal wind at 850 hPa and 200 hPa from 15°S to

15°N [208]. Active MJO events are defined based on the criteria in [19]:

1. There should be at least 30 days during which the amplitude ($\sqrt{PC_1^2 + PC_2^2}$) exceeds 0.5 (for [19], the threshold is 0.7 for U_{850hPa} field PCs).

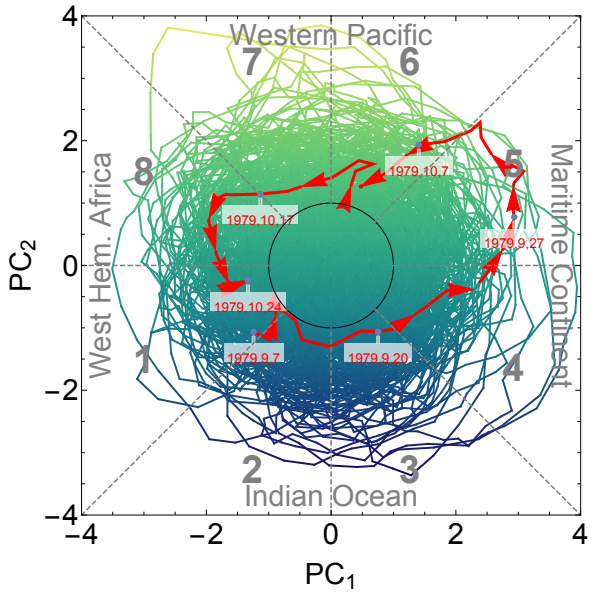
2. The MJO phase ($\tanh^{-1}[\frac{PC_1}{PC_2}]$) should move eastward for the entire period.

The detected active MJO events are displayed in Figure 2.13 as well. Based on the detected MJO phases, I explore the MJO's impact on precipitation distribution and extended range prediction skill below.

Influence on Precipitation Distribution

To investigate the MJO's influence on precipitation distribution, I follow the method in [126] to examine the average weekly precipitation anomalies conditioned on the MJO status. The methodology is briefly described as follows. First, the weekly precipitation anomalies are derived using same approach in the previous section. Next, for the early/late winter season and the entire winter season, I clustered the weekly precipitation anomalies based on the MJO status. Both the MJO phase and the lag days after the MJO phase are considered for constraining the distribution of precipitation anomalies, since it takes time for the MJO-related variability to exert their influences. Finally, I computed the mean value of the clusters and drew them in Figure 2.14.

The most obvious pattern in Figure 2.14 is the angled bands of precipitation anomalies, which generally stretch from top-right to bottom-left, following the MJO phase transitions. For instance, in late winter (JFM), abnormally high precipitation favors WA at Week 2 following onset of an active MJO event from Phase 1. The enhanced/suppressed precipitation bands are more distinctly separated for NCA, OR, and WA, as compared to SCA. A comparison between results for early winter season (column 1) and late winter season (column 2) shows that the phases of the MJO that promote enhanced/suppressed precipitation are



Model	MJO Active				MJO Non Active	All
	Phase 1-2	Phase 3-4	Phase 5-6	Phase 7-8		
BoM	95	123	130	103	737	1188
CMA	252	387	424	286	2473	3822
ECCC	57	70	86	55	512	780
ECMWF	100	142	178	110	952	1482
HMCR	137	183	197	151	989	1657
ISACCNR	83	112	121	93	671	1080
JMA	43	53	63	45	336	540
KMA	19	20	22	19	160	240
MeteoFrance	24	24	29	21	166	264
NCEP	150	232	253	173	1376	2184
UKMO	37	52	67	36	360	552

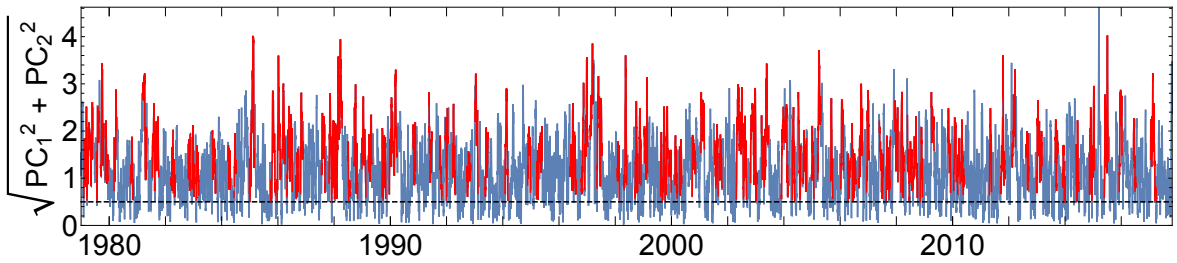


Figure 2.13: The top left figure shows the leading 2 Principal Components(PC) of the field that combines average outgoing long wave radiation, zonal wind at 850 hPa and 200 hPa from 15°S to 15°N. The phase and amplitude of MJO is defined based on the position of (PC_1,PC_2) and its distance to origin. For instance, the red arrowed line represents an MJO event that starts from September 7,1979, goes counterclockwise (eastward when reprojected to geographic map), and ends on October 24,1979. On most days, (PC_1,PC_2) lies out of the middle circle, whose radius is 1, indicating a strong MJO event. The bottom figure displays the time series of MJO amplitude, as represented by $\sqrt{PC_1^2 + PC_2^2}$. Active MJO events are labeled with red lines. Hindcasts for 11 GCMs are labeled as "MJO Active" and grouped into corresponding clusters if the start time is within an active MJO period, as shown in the right table.

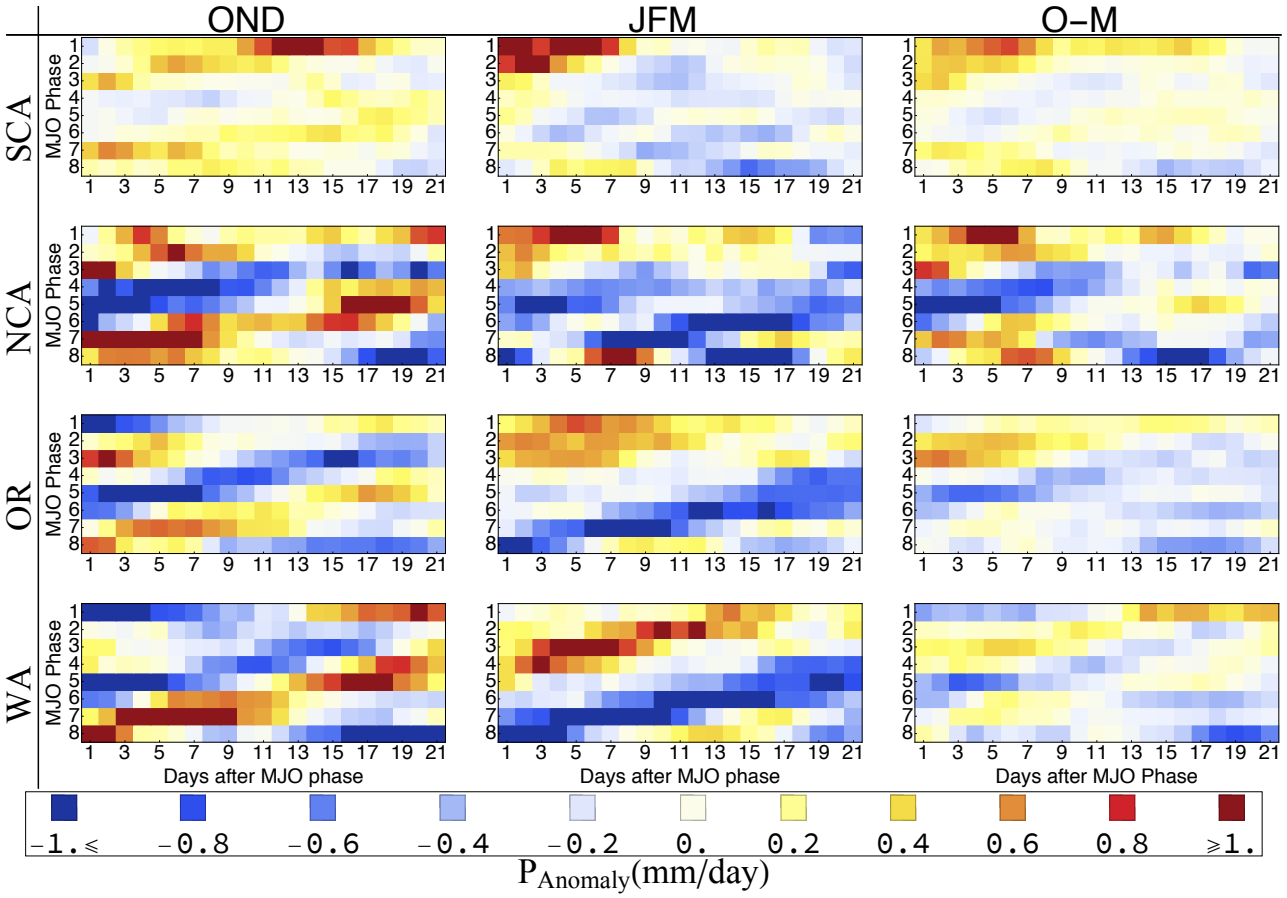


Figure 2.14: Mean value of weekly precipitation anomalies conditioned on active MJO events' phase and number of days after MJO phase onset. The four rows represent results for four geographic divisions, and the columns represent results for different seasons. In each sub-figure, the grid color represents the mean weekly precipitation anomalies for Day m after MJO phase n, here m ranges from Day 1 to Day 21, as labeled on the X axis, n ranges from Phase 1 to Phase 8, as labeled on the Y axis.

substantially different during these two seasons. This is in agreement with the findings of [19].

Generally, results confirmed the impact of the MJO on modulating precipitation regimes for the West Coast, especially for NCA, OR, and WA. The time lag for MJO to manifest its effects provides valuable potential for extending the range of skillful predictions.

Influence on Precipitation Prediction Skills

To investigate the impact of the MJO on precipitation prediction skills, I first grouped the hindcast cases for each GCM according to its start-time MJO status. Five groups are adopted here following [84]: Phase 1−2, Phase 3−4, Phase 5−6, and Phase 7−8 for the active MJO period, as well as the MJO-quiescent period. The sample size is listed in Figure 2.13. Next, I evaluated the Week 2 and Week 3−4 prediction skill for each group. The statistical significance of r skill differences between MJO-active groups and MJO-quiescent group were determined using the z test. Results for r skills and significant tests are shown in Figure 2.15.

While results confirmed the former findings that the prediction skill varies for different MJO groups and models [119], there are some common patterns here. Firstly, for hindcasts initialized during active MJO in Phase 3−4, most models show lower extended-range prediction skills as compared to MJO-quiescent cases, except for the Week 2 prediction in SCA. This skill drop was also found in [84] when studying the impact of the MJO on intraseasonal predictability in the mid-latitudes of the Northern Hemisphere. The skill drop might be attributed to the fact that many models can not represent well the propagation of the MJO across the Maritime Continent [101, 197, 201]. If so, models cannot produce the MJO-associated extratropical response revealed in Figure 2.14, resulting in systematic forecast biases. Secondly, for hindcasts initialized during active MJO in Phase 1−2, Phase 5−6, and Phase 7−8, forecasts from the better performing models (i.e., ECCC, ECMWF, MétéoFrance, and UKMO) are generally more skillful compared to MJO-quiescent cases. For

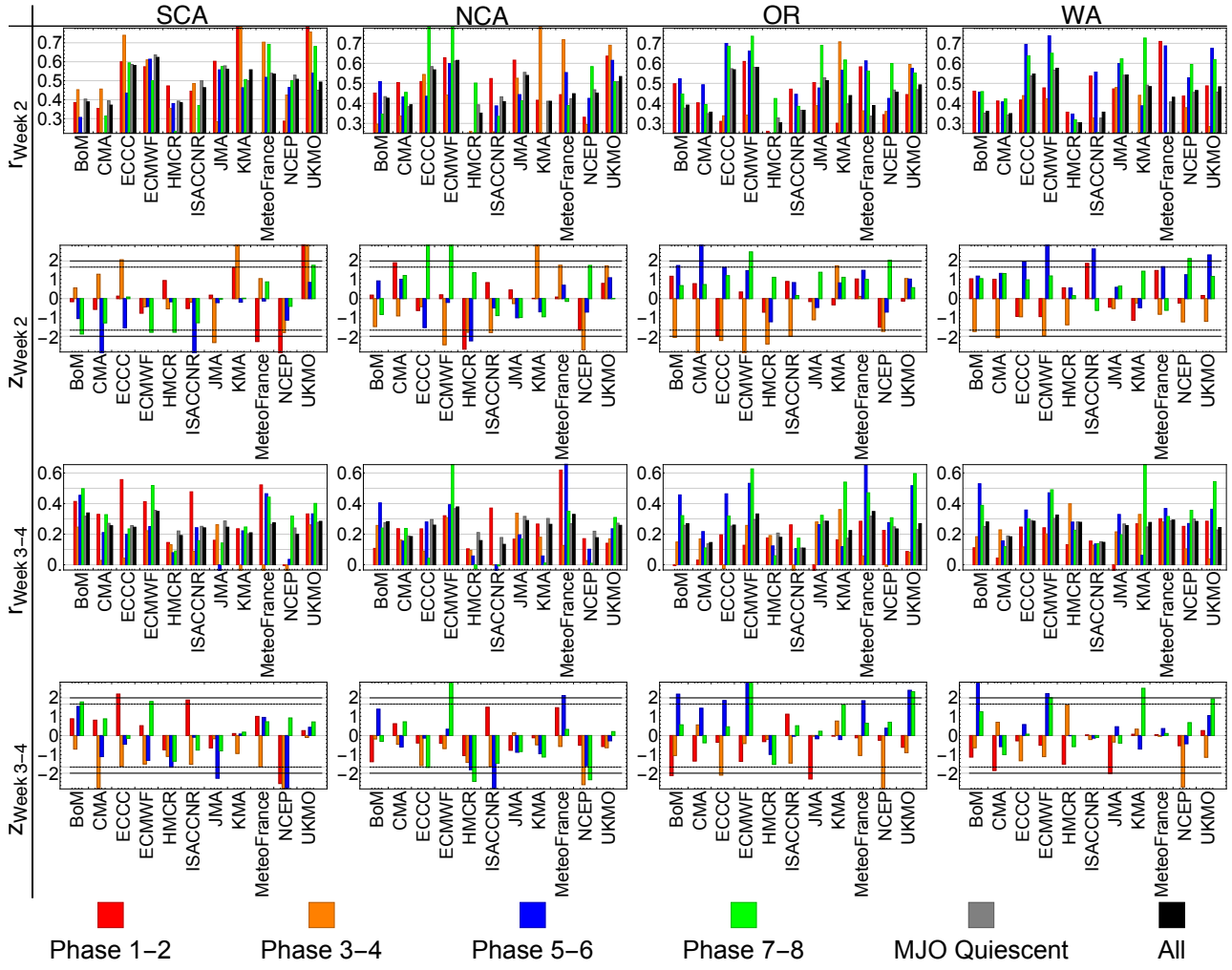


Figure 2.15: Row 1 (3) shows the Week 2 (Week 3-4) r skills for different MJO groups. Row 2 (4) shows the Week 2 (Week 3-4) z statistics of differences between r skills for MJO-active groups and MJO-quiescent group, the dashed (solid) grid line indicates statistical difference at 90% (95%) confidence level.

instance, for SCA and NCA, the prediction skill during active MJO in Phase 1−2 and Phase 7−8 is generally higher than in MJO-quiescent days; for OR and WA, the prediction skill during active MJO in Phase 5−6 and Phase 7−8 is generally higher than in MJO-quiescent days.

2.6 Discussion and Conclusions

This chapter evaluates the precipitation prediction skills at short to extended range for the West Coast, where precipitation variation significantly influences the local ecology and economy. The evaluation is based on the extended range hindcast dataset of the WWRP/WCRP S2S project. For the 11 models used here, the hindcast sample size ranges from 240 to $3,822$ across more than 20 winters, covering various climate circumstances categorized by intraseasonal to seasonal variability. This guarantees that the evaluation is less prone to biases from limited sample size or model diversity.

Since there is inevitable deviation of prediction along forecast lead time, I implemented the evaluation at both stringent and loosened spatial and temporal scales, measured by both deterministic scores (r and NSE) and probabilistic score (ROC and $\overline{\text{CRPS}}$). I further examined the impact of extended-range predictability sources, focusing on ENSO and the MJO. The key findings are listed as follows:

1. The S2S models' prediction skill-forecast lead time relationship is quantified for the four divisions in the West Coast.

    - For Week 1, the S2S models show advantageous precipitation prediction skills. The r/NSE/ROC score is approximately of the order of 0.8/0.7/0.8 for this period. By spatial averaging, the skill score can be further improved.
    - For Week 2, models show large variations regarding their performances. The Week-2 mean precipitation forecast from the best-performing model (i.e., ECMWF) is of considerable value, with $r > 0.6$, NSE> 0.35, and ROC score > 0.7.

- Beyond Week 2, predictions generally provide little deterministic skill. For this range period, probabilistic evaluation of ensemble forecasts using the $\overline{\text{CRPS}}$ shows significant advantage of ensemble forecast over deterministic forecast.
- Considering the performance difference of the S2S models, the informative predictable range may differ by up to $6-7$ days across different models. For the short range, models with higher resolution tend to have better performances (JMA, KMA, ECCC, and ECMWF). For medium to extended range, ensemble mean predictions show significant better performance compared to deterministic predictions. The best performing models for this range period are the ECCC, ECMWF, and JMA. For Week 3–4 forecast, although there is essentially no useful deterministic forecast skill, the ECMWF model still shows advantage over the rest models. Results here can benefit model selections for practical forecasts and multi-model ensemble predictions.

2. Through investigating the impact of ENSO on the West Coast precipitation distribution and models' prediction skill, I found a spatial see-saw effect for ENSO to modulate precipitation distribution and prediction skill:
   - During El Niño years, Southern California receives more precipitation in late winter on average, and most models show better extended-range prediction skills.
   - During La Niña years, Oregon receives more precipitation in winter season, with most models showing better extended-range prediction skills.

   For Northern California or Washington, ENSO influences the precipitation distribution, but specific models may either have higher or lower prediction skills depending on the ENSO phases. I assume the excessive precipitation and improved extended-range prediction skills accompany the meridional shift of baroclinic systems as modulated by ENSO. This predictability difference related to ENSO phase will be useful for extended-range prediction applications.

3. The impact of MJO on the West Coast precipitation distribution and models' prediction skill is explored.

    - To assess the impact of MJO on precipitation distribution, I examined the average precipitation anomalies conditioned on the MJO phases and days after MJO phases. Our results show that MJO systematically modulates the region's precipitation distribution. The time lag (here, up to three weeks) for MJO to manifest its effects provides valuable potential for skillful predictions at the extended range.
    - Regarding the impact of the MJO on GCM's extended-range precipitation prediction skills, I verified that for certain MJO phases (especially, Phases 5-6 and 7-8), some S2S models can well capture the MJO-associated teleconnections, improving Week 3–4 prediction skills. However, for hindcasts initialized during active MJO in Phase 3–4, most models show lower extended-range prediction skills as compared to MJO-quiescent cases, suggesting that the forecast opportunity may also be a curse if models have deficiencies in capturing the MJO or the related teleconnections.

Results here suggest the potential for predictability across a range of time scales [74, 226]. It is hopefully that the baseline provided here can foster practical subseasonal prediction applications and facilitate further research on improving mid-latitude subseasonal precipitation forecasts.

Chapter 3

Improving Precipitation Estimation with Convolutional Neural Network

3.1 Introduction

Evaluation results from the previous chapter show high r skill score for predictions at short to extended range. The considerable correlation indicates models' capacity in capturing the temporal variability of the precipitation processes. The skill is considered to be inherited from models' skills in predicting the underlying synoptic weather systems. On the other hand, a negative NSE score using GCMs' raw precipitation predictions indicates that models struggle in representing precipitation process at the right order of magnitudes. Here, ideas from the recent advances of deep learning techniques are applied for developing a statistical downscaling model that helps to alleviate the poor status of representing precipitation processes in numerical weather/climate models.

As has been discussed in the Introduction, the atmospheric *primitive equations*, which are derived by applying the conservation laws and thermodynamic laws on the continuous "control volume" of the atmosphere [18, 70], form the basic of numerical prediction of the atmosphere. With the rapid growth of computing power, we can discretize and resolve

these equations on increasingly finer computing grids. However, there remains many critical sub-grid scale processes that are not explicitly resolved.

Precipitation results from complex processes that remain mostly unresolved in current atmospheric models [184]. The modeling of precipitation involves explicit and implicit representations of the cloud physics, such as the water vapor convection, phase change, particle coalescence, etc. These processes take place at millimeter to molecule scales, which far surpass the resolution of current numerical models ($O(1\text{km})/O(10\text{km}) - O(100\text{km})$ for weather/climate models). Also, the assumptions of thermodynamic equilibrium and continuity lose their validity in describing some of the microscopic processes [178], making it necessary to adopt supplementary equations for physically solid simulations.

In numerical models, such unresolved processes are inferred from the resolved dynamics on the computational grid [87]. This process is known as parameterization. Specific to precipitation, the directly related parameterization schemes are cloud microphysics and subgrid convection. Given the intrinsic complexity of the cloud and precipitation process, the equations and their associated parameters in these parameterization schemes are generally of high structural and parametric uncertainties [40]. As a result, models' precipitation products are usually considered less reliable compared to the directly resolved variables, such as pressure and temperature [64, 15, 193, 24, 186].

Statistical Downscaling (SD) methods are also used for the purpose of inferring the poorly represented processes from the resolved dynamics and other data sources. However, SD has distinct objectives compared to the parameterization schemes. The main purpose of parameterization is to depict the sub-grid scale processes for realistic atmosphere modeling. The primary concern for SD, as indicated by the name, is to resolve the scale discrepancy between the existing model simulations and application requirements [116]. Accordingly, the model input/output, resolution, usage, and complexity of parameterization schemes and SD are different.

Besides the scaling issue, another aspect of SD is noted in practices: Compared to raw

outputs or the dynamically downscaled outputs from numerical models, SD occasionally provides more accurate estimates of the unresolved processes. This is because SD is customized for specific objective, region, and climate condition. The data-driven model with carefully designed model architecture and calibrated parameters may outperform the default parameterization schemes in relating the unresolved processes to resolved circulation. This phenomenon offers valuable implications for improving the relevant parameterization schemes and opportunities for enhancing the prediction of the parameterized processes [161, 149].

Here I focus on fostering this aspect of SD for weather scale precipitation forecast. Specifically, I propose to improve the accuracy of daily precipitation estimates through relating the precipitation process with the circulation data that are explicitly resolved in the atmospheric primitive equations. Compared to conventional SD applications, the task here poses much higher requirements on model resolution and accuracy. Recent developments in Machine Learning (ML) techniques, especially the branch of Deep Neural Networks (DNNs), offer an opportunity for describing and predicting such complicated physical processes using a compose of big data and advanced model architectures. Here I illustrate how a particular form of DNN, named Convolutional Neural Network (CNN) [97], can be adapted to address the precipitation estimation problem.

The rest of this chapter is organized as follows: I start with a brief review of relevant works. Then, I formulate the problem and illustrate the model requirements for this application. The model is described and tested thereafter. I show the model results and provide methods for analyzing and interpreting the models. I compare the model performance with some of the widely-adopted SD approaches. Conclusions are drawn at last.

3.2 Related Works

Many studies have been conducted on improving precipitation prediction accuracy with statistical approaches. Many of them share similar objectives and ideas compared to this

work. The relevant SD methods are reviewed. Also, I briefly review the basic concepts of DNNs, with a special emphasis on their applications in physical processes.

3.2.1 Statistical Downscaling

Following the survey in [116], SD approaches are classified into Perfect Prognosis (PP), Model Output Statistics (MOS), and Weather Generators (WGs). Since the objective for our study is deterministic precipitation prediction, I focus on SD approaches that make deterministic estimates of precipitation or its estimation biases. This includes PP and MOS. The WG models are not reviewed here.

PP models construct statistical relations between the large scale predictors and local scale predictands [46, 116]. Both the predictors and predictands are considered to be realistically simulated or observed, hence the name of "perfect". Along with the advancement of General Circulation Models (GCMs), many precipitation PP methods have been developed. The simplest form is linear regression, which estimates precipitation using an optimized linear combination of the local circulation features [128, 59, 99, 83]. The predictors usually consist of the raw variables or the leading Principal Components (PCs) of the moisture, pressure and wind field [213]. Besides the linear models, there are also approaches that utilize the non-linear features of relevant circulation field, such as Self-Organizing Map (SOM) [73], Support Vector Machine (SVM) [190], Nearest Neighbor [50], Random Forest [79], Artificial Neural Network (ANN) [163], etc.

MOS stands for the practice of using statistical approaches to enhance the model's prediction accuracy [55]. Compared to PP, MOS is more frequently used in Regional Circulation Models (RCMs) than in GCMs [116]. Also, the predictors of MOS are numerical models' raw outputs, which are not assumed to be perfectly estimated. For instance, a typical application of MOS is to correct the biases of the numerical model's raw precipitation estimates [81]. It should be noted that the validity and universality of precipitation MOS rely on the consistency of precipitation estimation biases, which is usually not guaranteed given the

continuous improvements of numerical models.

The performances of the above-mentioned SD approaches have been compared with dynamical downscaling results [127, 61, 160, 58, 5, 182]. For instance, an intercomparison of six SD models and five RCMs for Europe indicated that PP and MOS models achieved higher skill scores in estimating certain aspects of precipitation, such as the occurrence and intensity [5]. On the other hand, another comparison study showed clear advantage of RCMs for estimating precipitation over complex terrain [160]. Overall, the performance of SD depends on many factors, including the selection of predictors, the model and its implementation, the available data and the climate condition, etc.

3.2.2 Deep Neural Networks and their Applications for Physical Processes

Deep Neural Networks (DNNs) belong to the domain of Machine Learning (ML), which covers a general scope of computer aided statistical modeling. DNNs differ from traditional ML approaches in their modeling workflow. In a canonical ML modeling process, the raw-form data, which quantify certain attributes of the study object, should be transformed into a suitable feature vector before being effectively processed for the learning objective [96, 56]. The feature extraction process is typically performed in separation with the modeling process. Despite the expert knowledge and engineering works required for the feature extraction process, a pre-defined feature extractor captures little useful information beyond our prior knowledge. This issue is particularly severe for high dimensional problems, where it is difficult to have foresight in the intricate but important data structures. On the other hand, DNNs, together with a broader family of Representation Learning (RL) approaches [11], offer an "end-to-end" modeling workflow: the feature extraction process is integrated into the modeling process, which allows the model to learn customized features rather than subject to the pre-engineered features.

DNNs learn to customize features through building multiple levels of representation of

the data, which are achieved by composing simple but non-linear modules (named as neurons) that each transform the representation at one level into a representation at a higher, slightly more abstract level [96]. The differentiability of the hierarchical model allows applying the gradient descent algorithm to tune the neurons' parameters in order to make the model exhibit desired behavior. This process is widely known as backpropagation training [207, 155]. In addition to these basic concepts, modern DNNs involve numerous network architecture variations, training algorithms and tricks, regularization methods, etc. A comprehensive review is beyond the scope of this work and can be found in [96], [159], and [56]. A trans-disciplinary review of DNN relevant to water resources related research can be found in [167] and [168].

DNNs have dramatically improved the state-of-art in applications that can not be adequately solved with a deterministic rule-based solution, such as visual recognition [93], speech recognition [4], video prediction [109] and natural language processing [176], etc. For the modeling of the natural physical processes, where Ihave established principled solutions through analytic descriptions of the scientist's prior knowledge of the underlying processes [34], dynamical simulations are preferred to ML-based approaches. However, recent developments showed that provided with 1) big amount of data and 2) well-designed network architectures that encode the physical background knowledge, DNNs are competitive with numerical methods in simulating complex natural processes.

Generally two motivations are found for adopting a data-driven model besides the classical dynamical simulation. The first is computing efficiency. The computational demanding components in numerical simulations can be replaced by data-driven model counterparts to accelerate the simulation without significant loss of accuracy. Examples include using DNNs to simulate the Eulerian Fluid [187], and to predict the pressure field evolution in fluid flow [212]. The other concern is to represent the unresolved processes beyond the original numerical simulation. For instance, [52] trained a neural network to represent the sub-grid scale convection process in atmospheric modeling. The trained model was coupled in GCMs

and skillfully predicted many of the convective heating, moistening, and radiative features. [220] applied a a conditional Generative Adversarial Network to generate spatiotemporal coherent high resolution fluid flow based on its low resolution estimates.

For the applications mentioned above, a particular DNN architecture named Convolutional Neural Network (CNN) acts as a core building block. Compared to conventional neural networks, CNNs have significantly enhanced our capacity in processing structured high-dimensional data. This is achieved by utilizing the inner structure of the data to reduce the model structural redundancy and foster effective information extraction. Geophysical data are intrinsically structured in space and time. The huge geophysical datasets from remote sensing observations, numerical simulations, and their composite offer precious deposits for the application of DNNs [183]. CNNs have found applications in detecting extreme weather from the climate datasets [106] and precipitation nowcasting [221, 169]. More related to our objective, [191] developed a Super Resolution Convolutional Neural Network (SRCNN) for precipitation SD. The low resolution precipitation field (1°) and elevation field data were fed into the SRCNN to produce the high resolution precipitation field ($\frac{1}{8}^\circ$). Inoted that many of these geophysical CNN applications took little use of the atmospheric dynamical modeling products, which offer physically solid and comprehensive information of the atmosphere. While many recent research works have started to explore the applicability of DNN for parameterizing the unresolved processes in fluid and geofluid modeling [105, 149], it remains a question how DNN can translate the big data of observations and numerical simulations into precipitation estimation improvements [142].

3.3 Problem Formulation

To formulate the precipitation estimation problem, I first clarify the context by introducing a real-world precipitation scenario. Figure 3.1 shows a winter storm that hit the windward slope of the Cascade Range between Washington State and Oregon State of the U.S.A. The

case is selected for two reasons. Firstly, we have well-developed conceptual and theoretical models for describing such extratropical winter precipitation processes [154, 16]. The well established models offer accessible concepts for describing the circulation-precipitation connection. Secondly, the pronounced orographic effect [6] together with the lack of observations for mountainous region highlight the necessity for accurate precipitation estimation. While Ihave long-term observations for the target grid, Ihope the proposed model can translate the rich observations into new understandings for precipitation estimates in less-observed complex terrain areas.

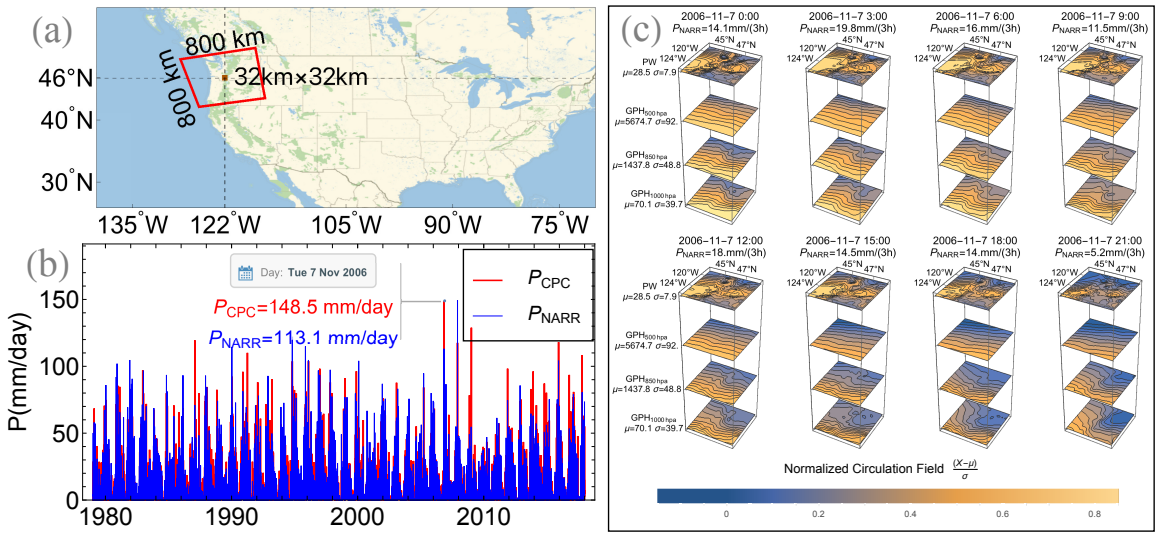


Figure 3.1: (a) The case study area of a 32km×32km geogrid centered at (46°N, 122°W). Its surrounding circulation field is delineated with the 800km×800km red polygon. (b) The geogrid's daily precipitation time series from 1979 to 2017. The red thick line represents the gauge-based precipitation records from the NOAA Climate Prediction Center (CPC); the blue slim line represents the model reanalysis records from the NCEP North American Regional Reanalysis Project (NARR). Data details are given in the Section 5.1. (c) The every 3 hour snapshots of the circulation profile for the storm event that happened on November 7th 2006. The geopotential height (GPH) at 1000hPa, 850hPa, and 500hPa, as well as the total column precipitable water (PW) are obtained form NARR. Data are normalized by subtracting the field mean (μ) and dividing by the field standard deviation (σ).

Dynamically, the precipitation process in Figure 3.1 is associated with the extratropical cyclone. The cyclone exists due to the baroclinic instability, which is caused by the strong meridional temperature gradient during winter time [166]. Aroused by the baroclinic dis-

turbances, the upper-level divergence leads the development of surface cyclone convergence and its associated cold and warm fronts, which in turn drive the moisture convection and precipitation formation [33].

The dynamical process exhibits characteristic synoptic appearances through the cyclogenesis process, and is accompanied by distinct precipitation distribution patterns. For instance, following the description of the classical Norwegian Cyclone Model [16], the *open wave* and *seclusion* stages for an idealized cyclone life cycle can be well distinguished in the circulation profile snapshots in Figure 3.1. These phenomenological understandings supported empirical precipitation forecast prior to the era of numerical weather prediction [131].

I follow the same phenomenological methodology but adopt statistical models rather than weather forecasters' expert knowledge for estimating precipitation from the resolved atmospheric dynamics. This is achieved by constructing an empirical mapping from the resolved circulation patterns to precipitation:

$$E(P|X, C) = f_C(X, \beta) \tag{3.1}$$

In Equation 3.1, E denotes the expected value, P denotes the precipitation estimates, X denotes the predictors, C denotes the local climate condition. f_C denotes the empirical function for the specific climate condition C, β denotes the parameters for f_C.

Specific to the objective here, P is defined as the total daily precipitation amount for the target grid. To estimate P with Equation 3.1, the predictor variables, their coverage and spatiotemporal resolution, the form of f_C and its training/validation process should be clarified.

To make deterministic precipitation estimation using Equation 3.1, X are assumed to be realistically simulated, which are often represented using reanalysis products that assimilate observations of various sources. The model is trained and tested by relating X with

realistically observed precipitation records.

Consider the 32km $\times$ 32km geogrid in Figure 3.1, the characteristic scale for the horizontal velocity of the atmosphere is 10m/s [70]. For a single day time, the dynamics within the coverage of 10m/s $\times$ 3600s/h $\times$ 24h $\approx$ 800km may either exert direct dynamical influence on the target geogrid, or provide important circulation context information. The circulation and moisture profile within this range constitute a four-dimensional physics field (three spatial dimensions and one temporal dimension). In typical SD applications, this field is firstly transformed into a feature vector before being applied to estimate precipitation. Ipoint out two common deficiencies when applying this conventional approach for weather-scale precipitation estimation. Firstly, the feature vector is often designed based on the characteristics of the predictors rather than the physical connection between the predictors and predictand. For instance, the leading Principal Components (PCs) of the circulation field are the widely-adopted features for precipitation SD. They compose a coarse picture of the circulation field at climate scale. For a specific cyclogenesis event, the precipitation is directly related to its corresponding cyclone geometries, which are disparate from event to event, not all of them could be well decomposed along the leading eigenvectors of the circulation field. Thus, although the leading PCs represent the coarse structure of the predictors, they may not provide comprehensive information for estimating precipitation. The second drawback is that the dominant factors which influence the precipitation distribution are not well disintegrated and represented. While SD models are preferred to recognize key circulation features of different appearances and locations, most existing statistical models make no explicit consideration of the cyclone depression intensity, coverage and its distance to the target geogrid.

To tackle these problems, I abide the "end-to-end" principle for extracting dynamical features and estimating precipitation. Specifically, the fine-scale circulation field resolved by numerical models is directly processed by the DNN to learn the representative features for precipitation estimation. Also, to disintegrate the impact of the cyclone geometric shape

and position, I adopt the convolution mechanism in the network modeling. The method is explained in the following section.

3.4 Methodology

3.4.1 Convolutional Neural Network

CNNs share many similarities with regular neural networks. For a regular neural network, a statistical connection between the inputs and the outputs is constructed through hierarchical connected layers of neurons. Each neuron is a computing unit that receives some inputs, performs a dot product and optionally follows a non-linear transformation. For supervised learning problems (i.e., classification and regression), a loss function is defined by comparing the network's output estimations with observations. The network parameters are typically trained through minimizing the loss function using gradient descent, which is known as backpropagation training.

CNNs differ from regular neural networks in their connection manner for neurons within and between layers. In regular neural networks, each neuron is fully connected to all neurons in the previous layer, while neurons within a single layer do not share connections. For data that come in the form of multiple arrays, such as images and videos, there is strong correlation within neighboring patches and less considerable correlation between remote components. The local patterns contribute to large-scale patterns when they are inspected from a broader point of view. To extract and utilize the local features in neural network modeling, the full connection of neurons between successive layers becomes redundant and local connection within a single layer becomes imperative. To tackle this, the CNN explicitly encodes local connection and prohibits remote connection; also, the extracted local patterns are down-sampled to compose large-scale patterns. These operations are achieved using the convolution operator and pooling operator.

Below I use the precipitation estimation example to explain the idea of these operators.

The storm event for the same geogrid in Figure 3.1 on 18:00 UTC, November 7^{th} 2006 is used as illustration here. In Figure 3.2, the circulation and moisture field (left part) are fed into a CNN for extracting useful features and estimating the precipitation of the target geogrid (right part). The blue color indicates low values, yellow indicates high values. A clear frontal system can be depicted as an occluded front forms around the mature low pressure area. This is accompanied by copious precipitation falling along the warm conveyor belt, as shown in the precipitation map. These characteristic circulation patterns could appear in different geometrical shapes and at different locations. I expect that an explicit encoding of the local connectivity could enhance the extraction of the circulation geometries for precipitation estimation.

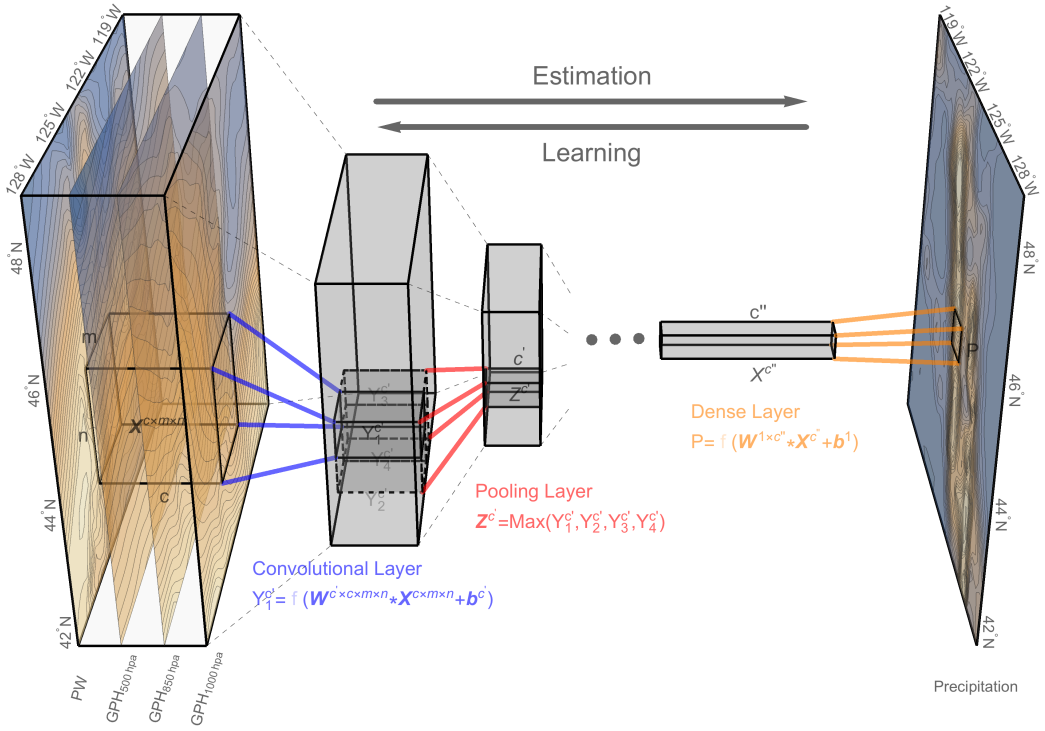


Figure 3.2: The CNN architecture for estimating precipitation using the numerical model resolved geopotential height and moisture field. The data are obtained from NARR dataset. The stacked frames on the left side show the PW, GPH at 500hPa, 850hPa and 1000hPa for the delineated 800km×800km region in Figure 3.1. The blue lines indicate a convolution operation applied on the circulation field. The red lines indicate the pooling operation that down-samples the local features. Several stages of convolution and pooling layers are stacked, followed by the fully connected dense layers (orange lines). The dense layer applies all the extracted features to estimate precipitation for the target geogrid, which is labeled on the precipitation map on the right part. The convolution and dense layers are optionally followed by a non-linear transformation f, which are represented with semi-translucent fonts.

To extract the spatial salient features, the convolutional layer applies a $c' \times c \times m \times n$ tensor to go through the input with a pre-defined stride. Each convolution operation is performed by computing the element-wise dot product between the tensor and different patches of the input, which is represented as a $c \times x \times y$ array. Here c is the category of the predictors. Following previous works [213], the predictors consist of the circulation constraint and moisture constraint. The circulation constraint is represented by geopotential height (GPH) at different pressure levels. The moisture constraint is represented by the total column precipitable water (PW). The spatial coverage of the dynamical field is defined by x and y.

The $c' \times c \times m \times n$ tensor is named as kernel of the convolution layer, where c' is the output channel number, $m \times n$ represent the *receptive field* of kernel. Each multiplication result is optionally followed by a non-linear transformation f, such as the ReLU function: $f(x) = max(0, x)$. The convolution operation can be interpreted as using c' filters to transform the input into a more representative form for the learning objective. Preferably, the network will learn filters that activate when they see critical local dynamical patterns on the first layer. Fostered by spatial down-sampling, which is achieved by the pooling operation, the network may eventually learn a synoptic atmospheric pattern that promotes precipitation on higher layers of the network.

The pooling layers act to coarsely grain local semantically similar features into one [97]. Through down-sampling, the higher layer convolutions work on extracted local features, which enables learning higher level abstractions on the expanded receptive field [96]. A typical pooling unit computes the maximum of a local patch of units in one feature map. In Figure 3.1, I apply a 2×2 maximum pooling. For a typical CNN, several stages of convolution, non-linearity and pooling layers are stacked, followed by dense layers that apply all the extracted features for the learning target.

To estimate the total daily precipitation, we usually have several snapshots of its surrounding dynamical field at different hours through the day. Conventionally, these snapshots are averaged to reach a single picture of the general dynamical pattern for the daily precipitation estimation. Here, considering the fact that each circulation snapshot provides its information for a specific time of the day, I map the same CNN model on each of the dynamical field snapshot. Results from all the network computations are summed as the total daily precipitation estimation.

3.4.2 Regularization, Loss Function and Training

Regularization

DNNs usually have much more complicated structures and more parameters than conventional ML algorithms, which make it possible for models to perform exceptionally well on the training data, but predict the test data poorly. This phenomenon is called overfitting. Regularization refers to the strategies to avoid overfitting and make the model generalize better to unseen data. The dropout [177] and batchnormalization [80] modules are adopted to enhance the model's performance.

The idea of dropout is to assign a *probability of existence* to the neurons and their associated connections. Thus, to train a DNN with n neurons is equivalent to train an ensemble of 2^n "thinned" networks. During testing and applications, the weights are multiplied by the same *probability of existence* to offset the dropout effect. This prevents neurons from co-adapting and has shown significant improvements in reducing overfitting [177].

Batchnormalization addresses the problem of *internal covariate shift* in training DNNs. Specifically, the distribution of each layer's inputs changes during the training process, which requires continuously adaption and hinders the training process. Batchnormalization performs normalization of the output in the hidden layers [80]. It has shown good performance in accelerating the training and regularizing the model in various DNN applications.

Loss Function and Skill Metrics

The Root Mean Square Error (RMSE) between the precipitation simulations and observations is used as loss function here:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum(P_{\mathrm{obser}} - P_{\mathrm{simu}})^2} \tag{3.2}$$

Here P_{obser} denotes the observed daily precipitation records and P_{simu} denotes the simulated daily precipitation records.

The Pearson Correlation Coefficient (r) between simulated and observed daily precipitation is also used as supplementary skill metric for measuring model performance:

$$r = \frac{cov(P_{\text{obser}}, P_{\text{simu}})}{\sigma_{P_{\text{obser}}} \sigma_{P_{\text{simu}}}} \tag{3.3}$$

Here *cov* denotes covariance and σ denotes standard deviation.

Training

The backpropagation method is applied to train the network model. The method requires estimating the partial derivative of the loss function with respect to each parameter in the network, including those from both the convolutional layers and dense layers. The parameters are then adjusted along the gradient descent direction by a predefined stride, which is named "learning rate". A detailed derivation for backpropagation in CNN can be found in [21].

Considering the fact that a same CNN is mapped to several snapshots of the dynamical field in a single day time, and their outputs are summed up as the final estimate of the daily precipitation, the gradient of the loss function is equally attributed to each snapshot before applying the backpropagation training.

3.4.3 Model Implementation

I implement the network using the Wolfram Mathematica V11.3 Deep Learning Platform [215]. I use the Nvidia Quadro P5000 GPU (Graphics Processing Unit) to accelerate model training.

3.5 Experiments

3.5.1 Data

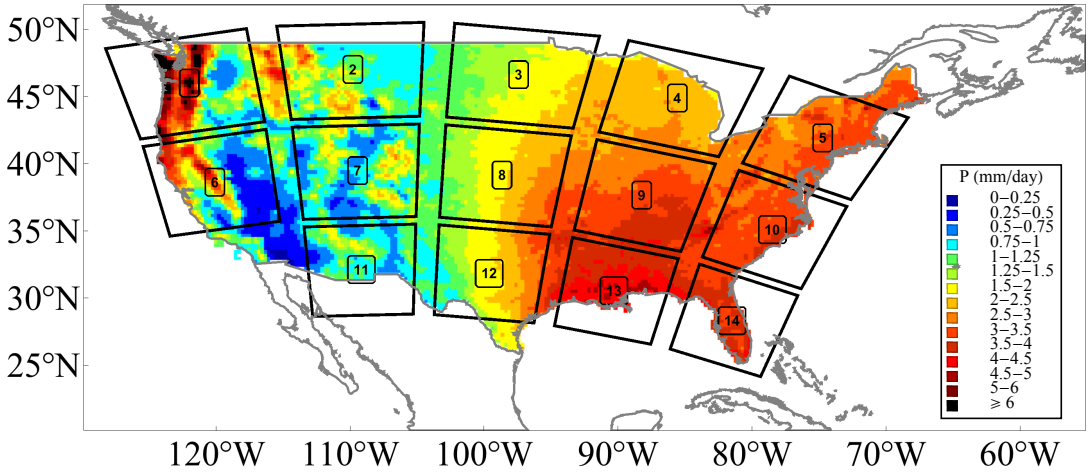
The predictors used for building the network models are the GPH and PW field data from the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR) dataset [121]. The dataset is generated by regional downscaling of the NCEP Global Reanalysis for the North America region, using the NCEP Eta Model and the 3D Variational Data Assimilation System. Also, the updated Noah Land-surface model and numerous datasets additional to the global reanalysis were applied to improve the data quality. The dataset covers 1979 to near present and is provided every three hours, with spatial resolution of 32km/45 vertical layers. The 3h total column PW, GPH at 500hPa, 850hPa and 1000hPa from 1979 to 2017 were downloaded for use.

Besides the pressure and moisture data, the precipitation product from the NARR is used as baseline here. It should be noted that the NARR precipitation product is not raw output from the numerical models but is achieved by assimilating precipitation observations as latent heating profiles [103]. Thus, the data quality is superior to the raw numerical precipitation estimates or conventional reanalysis precipitation products that do not assimilate precipitation [24, 10]. It poses a high challenge for the DNN model to provide comparable precipitation estimates.

I use the gauge-based daily precipitation dataset from the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) [218] as the "realistic" precipitation records for training and test the CNN. The dataset provides high-quality unified precipitation records by combining different information sources. It covers 1948 to 2017 with resolution of $0.25° \times 0.25°$. The data are re-sampled to 32km using nearest neighbour method to match the resolution of NARR.

3.5.2 Experiments Design

To test the applicability of the model for different climate conditions, I selected 14 sample grids that roughly cover the characteristic climate divisions of the contiguous United States. The samples together with their associated circulation field domains are shown in Figure 3.3. The circulation field for each gird is composed of a $8 \times 4 \times 25 \times 25$ tensor. 8 indicates that there are 8 3h circulation snapshots per day; 4 indicates the feature numbers, which includes the PW, $\text{GPH}_{1000\text{hPa}}$, $\text{GPH}_{850\text{hPa}}$ and $\text{GPH}_{500\text{hPa}}$; 25×25 represents that there are 25×25 32km grids within the considered dynamical field coverage.



	(lat,lon)	$\overline{P}$(mm/day)	r	RMSE(mm/day)		(lat,lon)	$\overline{P}$(mm/day)	r	RMSE(mm/day)
Point 1	(46°N, 122°W)	6.36	0.77	8.2	Point 8	(39°N, 99°W)	1.92	0.77	3.8
Point 2	(47°N, 110°W)	1.17	0.73	2.4	Point 9	(38°N, 88°W)	3.29	0.69	6.
Point 3	(47°N, 97°W)	1.53	0.75	3.2	Point 10	(35°N, 78°W)	3.33	0.57	7.6
Point 4	(45°N, 86°W)	2.19	0.72	3.7	Point 11	(32°N, 109°W)	1.01	0.74	2.2
Point 5	(42°N, 75°W)	3.45	0.63	6.6	Point 12	(32°N, 100°W)	1.9	0.76	4.5
Point 6	(39°N, 120°W)	3.01	0.82	5.4	Point 13	(31°N, 90°W)	4.39	0.65	8.8
Point 7	(39°N, 109°W)	1.05	0.53	2.6	Point 14	(28°N, 81°W)	3.45	0.63	6.8

Figure 3.3: The sample grids used in the experiment. For each grid, the surrounding 800km×800km dynamical field is delineated. The color indicates the mean daily precipitation rate, which is calculated by averaging the CPC daily precipitation records from 1979 to 2017. The table shows the samples' coordinates, mean precipitation rate, the r and Root Mean Square Error (RMSE) between NARR and CPC precipitation for the grids.

For each sample grid, I carry out the following steps to build the model:

1. Data Normalization. Each feature field is normalized by subtracting the mean (μ) and dividing by the standard deviation (σ). Here μ and σ are scalar values that are calculated based on the flattened circulation field for the entire dataset. The grids are

not normalized individually in order to maintain the circulation structure.

2. Divide the data into training, validation and test sets. To guarantee each set covers different climate conditions, for each of the twelve months, I randomly select 23 years of that particular months data into the training set, 6 years into validation set and 10 years into test set. Thus, I have 23/6/10 years' recomposed data in the training/validation/test set. The training and validation sets are used to calibrate the model parameters and prevent overfitting. The test set is kept unseen through the training process. It is used to provide an unbiased evaluation of the constructed model.

3. Determine the network hyperparameters. Hyperparameters are the variables that determine the network structure (i.e. layer type, neuron size) and the variables that determine how the network is trained (i.e., the learning rate) [56]. I adapt the architecture of a classical CNN implementation named LeNet [97] in determining the hyperparameters. The specific network structure is shown in Figure 3.4. Based on this basic model architecture, I focus on one sample grid and implement a series of architecture variations to figure out the best network structure configuration and attribute the contribution to the adjusted modules.

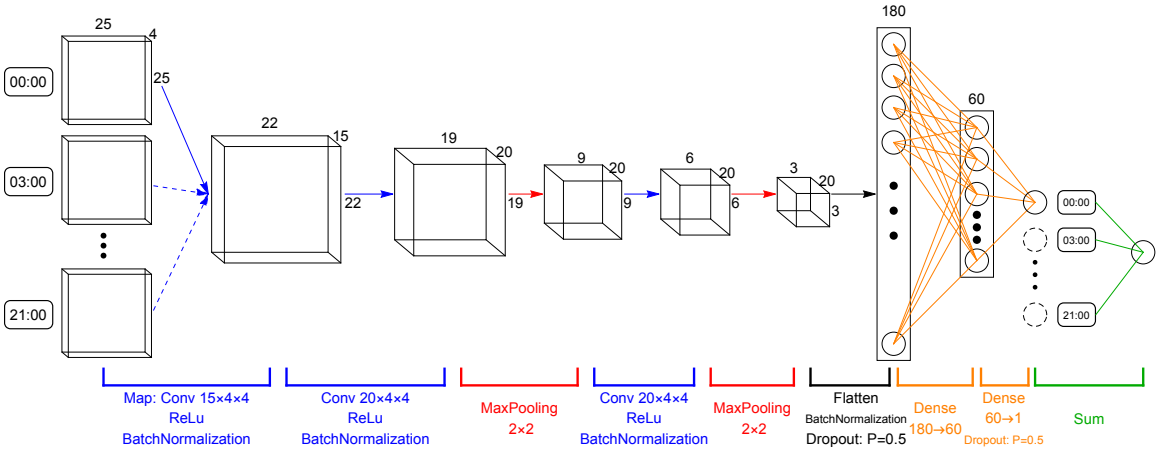


Figure 3.4: The network structure for precipitation estimation. Each 3h snapshot of the dynamical field, which is represented by a $4\times25\times25$ tensor, is sequentially processes through the convolutional layers and pooling layers. The extracted features are flattened and processes by two consecutive dense layers. The dimension of each layer's output is labeled out. Different layers/operators are denoted with corresponding colors. Results for eight 3h snapshots are summed as the total daily precipitation estimate. In total, this network consists of 2,4076 parameters to be trained.

4. Train the network. The parameters in the network are firstly initialized based on standard normal distribution. To guarantee model's robustness with respect to parameter initialization, I carry out several implementations with different parameter initializations. After initialization, the parameters are further trained using Stochastic Gradient Descent (SGD) [20] to minimize the MSE for precipitation estimate. The SGD method applies a stochastic approximation of gradient descent approach to alleviate the high computing cost in evaluating the derivatives for the global loss function. The considered learning rate are 10^{-2}, 10^{-4} and 10^{-6}. I adopt the early stopping strategy to regularize the model: the training process is terminated when further training improves performance only for the training set but not for the validation set.

5. Model evaluation. The network simulation results are evaluated against the CPC precipitation records, using skill metrics of RMSE and r. The performance are compared against the original NARR precipitation products.

3.6 Results

The CNN estimated precipitation ($\overline{\mathrm{P_{CNN}}}$) and the NARR estimated precipitation ($\mathrm{P_{NARR}}$) are compared against the CPC precipitation records. Figure 3.5 shows the comparison results for the test set. Here $\overline{\mathrm{P_{CNN}}}$ is the mean estimation from three CNN implementations with different parameter initializations.

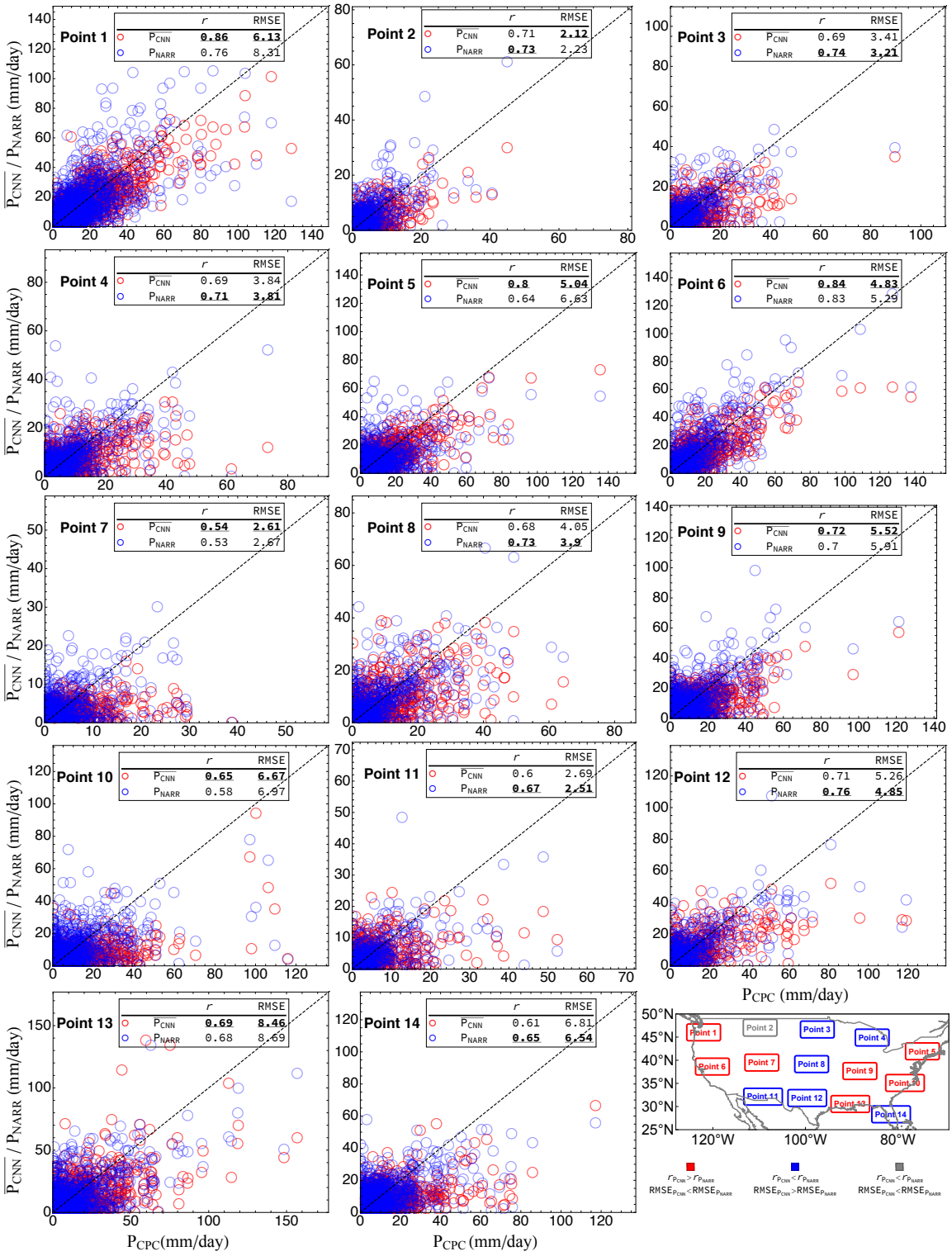


Figure 3.5: The scatter plots compare the $\overline{\mathrm{P_{CNN}}}$ (red circles) /$\mathrm{P_{NARR}}$ (blue circles) against the CPC precipitation records ($\mathrm{P_{obser}}$) for the 14 sample points. Results are for the test set only. The skill scores of r and RMSE for each point are given in corresponding sub-figures. The bold and underlined value indicates the better statistics of the two estimates. The bottom right geographic map shows the geoposition of the 14 points. The point is labeled red/blue if both skill scores indicate that $\overline{\mathrm{P_{CNN}}}$ / $\mathrm{P_{NARR}}$ performs better. It is labeled gray if the two skill scores show disagreement.

Without a careful tuning of hyperparameters, the CNN models perform relatively well compared to the NARR precipitation product. Considering the fact that the NARR precipitation product has already assimilated observations, the results are quite satisfactory.

As indicated by the two skill scores, $\overline{\mathrm{P_{CNN}}}$ outperforms $\mathrm{P_{NARR}}$ for most sample points from the west and east coast, where precipitation is more copious than the other areas. The skill improvement is impressive for some of the sample points. For instance, for Point 5, r/RMSE improves from 0.64/6.63 to 0.80/5.04 comparing $\overline{\mathrm{P_{CNN}}}$ with $\mathrm{P_{NARR}}$. For the rest sample points from the middle part of the continent, $\overline{\mathrm{P_{CNN}}}$ performs slightly worse compared to $\mathrm{P_{NARR}}$. Particularly, the CNN models show systematic underestimation for the large precipitation events.

Table 3.1 shows the skill scores of r and RMSE for the training, validation and test set for each of the three CNN implementations carried out here. The skill scores of CNN ensemble prediction and NARR precipitation product are also included for comparison purpose.

Table 3.1: Precipitation Estimation Skills of CNN and NARR for the Training/Validation/Test Set

		Training					Validation					Test				
		CNN_{R1}	CNN_{R2}	CNN_{R3}	$CNN_{\overline{R}}$	NARR	CNN_{R1}	CNN_{R2}	CNN_{R3}	$CNN_{\overline{R}}$	NARR	CNN_{R1}	CNN_{R2}	CNN_{R3}	$CNN_{\overline{R}}$	NARR
Point 1	r	0.9	0.91	0.9	**0.91**	0.78	0.85	0.85	0.85	**0.86**	0.74	0.85	0.85	0.85	**0.86**	0.76
	RMSE	5.39	5.1	5.47	**5.15**	8.17	5.97	6.	5.99	**5.81**	8.18	6.23	6.35	6.3	**6.13**	8.31
Point 2	r	0.79	0.82	0.79	**0.82**	0.74	0.63	0.64	0.65	0.66	**0.71**	0.67	0.7	0.68	0.71	**0.73**
	RMSE	2.14	**2.04**	2.13	2.05	2.38	2.52	2.49	2.48	2.44	**2.43**	2.21	2.15	2.2	**2.12**	2.23
Point 3	r	0.78	0.78	0.8	**0.8**	0.76	0.68	0.66	0.68	0.69	**0.71**	0.67	0.68	0.67	0.69	**0.74**
	RMSE	3.03	3.1	2.94	**2.94**	3.26	3.29	3.35	3.29	**3.23**	3.26	3.48	3.48	3.53	3.41	**3.21**
Point 4	r	0.78	0.77	0.78	**0.79**	0.72	0.73	0.71	0.68	0.72	**0.76**	0.68	0.66	0.66	0.69	**0.71**
	RMSE	3.23	3.25	3.26	**3.14**	3.61	3.72	3.81	3.97	3.73	**3.55**	3.89	3.98	3.96	3.84	**3.81**
Point 5	r	0.82	0.8	0.8	**0.82**	0.63	0.76	0.75	0.74	**0.76**	0.62	0.78	0.79	0.79	**0.8**	0.64
	RMSE	4.74	4.92	5.	**4.74**	6.58	5.36	5.43	5.56	**5.3**	6.56	5.24	5.17	5.17	**5.04**	6.63
Point 6	r	0.91	0.92	0.93	**0.93**	0.82	0.82	0.82	0.83	**0.83**	0.82	0.84	0.84	0.82	**0.84**	0.83
	RMSE	3.76	3.64	**3.29**	3.41	5.47	4.88	4.89	4.78	**4.7**	5.24	4.96	4.89	5.19	**4.83**	5.29
Point 7	r	0.59	0.58	0.59	**0.61**	0.53	0.53	0.54	0.52	**0.55**	0.52	0.53	0.51	0.51	**0.54**	0.53
	RMSE	2.32	2.36	2.37	**2.31**	2.54	2.8	2.79	2.86	**2.78**	2.98	**2.6**	2.64	2.68	2.61	2.67
Point 8	r	**0.79**	0.75	0.75	0.78	0.77	0.71	0.71	0.71	0.73	**0.86**	0.67	0.66	0.66	0.68	**0.73**
	RMSE	**3.77**	3.99	4.01	3.82	3.89	4.66	4.56	4.57	4.51	**3.31**	4.12	4.18	4.16	4.05	**3.9**
Point 9	r	0.74	**0.79**	0.76	0.78	0.7	0.72	0.72	0.7	**0.73**	0.65	0.7	0.71	0.71	**0.72**	0.7
	RMSE	5.6	**5.23**	5.44	5.31	6.11	5.27	5.29	5.41	**5.19**	6.	5.65	5.64	5.61	**5.52**	5.91
Point 10	r	**0.75**	0.65	0.66	0.71	0.6	0.54	0.54	0.54	**0.57**	0.47	0.57	0.63	0.62	**0.65**	0.58
	RMSE	7.79	**7.11**	7.23	7.12	7.56	8.95	**8.13**	8.24	8.27	8.76	7.52	**6.47**	6.67	6.67	6.97
Point 11	r	0.72	0.71	0.71	0.73	**0.77**	0.66	0.62	0.61	0.65	**0.75**	0.6	0.58	0.59	0.6	**0.67**
	RMSE	2.2	2.25	2.25	2.18	**2.04**	2.74	2.86	2.88	2.77	**2.4**	2.7	2.79	2.74	2.69	**2.51**
Point 12	r	0.8	0.75	0.79	**0.8**	0.74	0.74	0.71	0.7	0.74	**0.8**	0.7	0.67	0.69	0.71	**0.76**
	RMSE	**3.76**	4.2	3.9	3.8	4.36	5.07	5.29	5.32	5.07	**4.49**	5.37	5.58	5.45	5.26	**4.85**
Point 13	r	0.82	0.82	0.77	**0.82**	0.65	0.69	0.72	0.7	**0.72**	0.59	0.67	0.66	0.68	**0.69**	0.68
	RMSE	6.52	6.5	7.17	**6.44**	8.77	8.15	7.8	7.98	**7.76**	9.39	8.68	8.86	8.69	**8.46**	8.69
Point 14	r	**0.69**	0.64	0.67	0.68	0.64	0.62	0.62	0.62	**0.63**	0.59	0.61	0.59	0.58	0.61	**0.65**
	RMSE	**6.08**	6.48	6.29	6.15	6.68	6.97	7.	7.03	**6.87**	7.36	6.79	6.96	7.02	6.81	**6.54**

R1, R2 and R3 indicate 3 implementations of CNN with different parameterization initializations. $\overline{R}$ represents the mean estimation.

The bold and underlined values indicate the best statistics for corresponding dataset.

Compared to individual implementations of CNN, the ensemble estimation of CNN ($CNN_{\overline{R}}$) improves the skill scores in most cases . However, the improvement is generally not significant. Different implementations of CNN show similar skills. This indicates that the parameter initialization carried out here does not significantly influence the modeling performance, in other words, the model is robust with respect to different parameter initializations.

Considering the performance difference for the training, validation and test sets, as can be

expected, all points show better performance of CNN models for the training set compared to the validation and test sets. The overfitting phenomena are assumed to be responsible for the relatively poorer performance for the CNN models in the middle part of the continent. For instance, for Point 3, $r_{\overline{\mathrm{P_{CNN}}}}$ is 0.80/0.69/0.69 for the training/validation/test set, while $r_{\mathrm{P_{NRAA}}}$ is 0.76/0.71/0.74. The overfitting may due to the fact that for this sub-arid region, there are much less samples of precipitation events compared to the rest areas. The limited informative data can not effectively support the construction of a complicated deep neural network model. Despite the overfitting, the CNN models have relatively similar performance for the validation and test set, which guarantees the trained model can generalize well to the unseen data.

Another important aspect Inoticed is that the CNN models frequently underestimate large precipitation values. Ibelieve the underestimation might be caused by the following reasons: first, Ido not have enough large precipitation samples due to the uneven distribution of daily precipitation; second, the convective storms, which are common for the east and southeast of the continent, might require finer dynamical field for accurate estimation.

3.7 Discussion

3.7.1 Network Architecture

The results above are achieved using a same default network architecture as presented in Figure 3.4. To 1) attribute the credits to the introduced modules, 2) figure out their optimal configurations, and 3) relate our models to the classical ANN SD approaches, Iimplement a series of network architecture variations based on the default network structure.

Given the complexity of DNN structures and huge computing cost for model training, it is impractical to enumerate all the possible network architecture compositions. Here Ifocus on two dominant configurations in CNN design, namely the *receptive field* and the *network depth*. For processing convenience, Iuse a single geogrid to carry out the experiments. The

geogrid of Point 1 is selected since Ihave made detailed descriptions for it.

Receptive Field: Extensive Or Exclusive?

The *receptive field* for a convolutional layer refers to the patch size on which the kernel convolve with. In Figure 3.2, it is denoted as $m \times n$. It constrains the spatial scale of the features that Iexpect to extract through convolution. Large scale spatial features can be achieved either by adopting a big receptive field on the initial layers or assembling local features in deeper layers.

Consider the extreme condition of applying the most extensive receptive field, i.e. $m = x$ and $n = y$ for the case in Figure 3.2, all the pixels in the input layer are thus fully connected. The CNN degenerates to a regular fully connected neural network, which has been extensively applied in ANN SD. Considering another extreme condition of applying the most exclusive receptive field, i.e. $m = 1$ and $n = 1$, the one-by-one convolution performs a coordinate-dependent transformation in the filter space, which has been used to modify the dimensionality in the channel dimension [102, 181].

I carry out the experiments by modifying the receptive field for all the convolutional layers. Imaintain the other network configurations the same as the default setting. As is shown in Figure 3.4, originally, a 4×4 receptive field is adopted for all of the three convolutional layers. Here, Iconsider the receptive field of 1×1, 3×3, 5×5, 6×6 and full input size (25×25 for the first convolutional layer and 1×1 for the rest). The results are shown in Table 3.2.

Table 3.2: Model Performances for Different Receptive Fields and Convolution Depth

		Receptive Field						Convolution Depth			
		1×1	3×3	$4 \times 4^*$	5×5	6×6	Full	1	3^*	5	7
	Training	0.869	0.885	0.902	0.922	**0.939**	0.864	0.837	0.902	**0.921**	0.887
r	Validation	0.836	0.848	0.851	**0.853**	0.836	0.824	0.811	**0.851**	0.850	0.837
	Test	0.832	0.847	0.851	**0.854**	0.847	0.822	0.808	0.851	**0.853**	0.838
	Training	6.11	5.78	5.32	4.79	**4.59**	6.29	7.01	5.32	**4.82**	5.73
RMSE	Validation	6.25	6.05	5.99	**5.95**	6.36	6.55	6.94	**5.99**	6.02	6.27
	Test	6.6	6.37	6.29	**6.25**	6.49	6.85	7.32	6.29	**6.27**	6.58

*: The skill scores for the default settings (receptive field: 4×4, convolution depth: 3) are the averaged scores for the three implementations in the previous section.

For Point 1, the best training set performance is achieved when using a receptive field of 6×6; while the best validation/test set performance is achieved when using a receptive field of 5×5. The two extreme condition experiments achieve poorer performance than the others. The one-by-one convolution network works better than the full receptive field network, or in other words, the fully connected neural network. It should be noted that the optimal receptive field size might be different for areas with different precipitation mechanisms.

Physically, precipitation is highly variant in space. Its occurrence and intensity are closely related to local circulation patterns. The above experiments verified that an explicit encoding of local spatial circulation structures enhances the estimation of precipitation.

Network Depth: Shallow Or Deep?

The *network depth* can be roughly represented as how many layers there are in the neuron network. These layers learn representations of the data with multiple levels of abstraction [96]. Despite the simplicity of the transformation in each layer, the stacking of many layers allows learning intricate structures for complicated applications.

Here Iapply a relatively shallower CNN model and two deeper CNNs to examine the impact of network depth. The shallower CNN model is constructed by removing the latter

two convolutional layers and the last pooling layer from the default network in Figure 3.4. Thus, its convolution depth is 1. The deeper CNN models are constructed by adding two/four extra convolutional layers before the first pooling layer for the default network architecture in Figure 3.4. The kernel size of the included convolutional layers is set to $20 \times c \times 4 \times 4$, where c is the channel number of the previous layer. The deeper networks are thus of convolution depth of 5/7. The performance of the models are also presented in Table 3.2.

Compared to the deeper network models, the model with single convolutional layer achieves significant lower skill scores in estimating precipitation. The default network with 3 convolutional layers achieves optimal performance for the validation set. The model with 5 convolutional layers achieves optimal performance for the training and test set. Overall, results indicate that the shallow network is not as effective as the deeper networks in extracting useful dynamical features for precipitation estimation. Ideally, the deeper networks hold more potential in estimating the intricate features. However, results here show that the network with 7 convolutional layers achieve lower skill score than the networks with 3/5 convolutional layers. This might due to inefficient backpropagation training for such a deep network model.

3.7.2 Model Interpretations

The network models applied here involve much more complicated structures and more parameters compared to the existing SD approaches. It is imperative to explain what is learned through adopting these network components, and how the network can learn better. In response to this requirement, many approaches for understanding CNNs have been developed in recent years [42, 173, 225]. Here Iapply two commonly used visualization and analysis methods to interpret the models and their results.

Layer Activation

Layer Activation refers to setting break points in the middle layers of the network and visualize the activated outputs at these break points. [225] offered an excellent example in illustrating how layer activation can be used for interpreting and diagnosing CNNs. Ivisualize the layer activations of the storm event in Figure 3.1 as well as a light rain event on December 16th, 2017. Results are shown in Figure 3.6.

The input fields for the two events in Figure 3.6 show different spatial structures. However, it is difficult to tell how these patterns are related to precipitation. The outputs from the first convolutional layer (Conv 1) provide a sharper distinction for the heavy/light precipitation events: for certain channels, the outputs for one event are activated while the outputs for the other event are not. For instance, the light precipitation event show high spatial variance in the channels of C_3, C_6, C_{10} and C_{11}; for the storm event, there are little spatial variance in these channels; on the other hand, the light precipitation event show little spatial variance in the channels of C_5, C_7, C_9, C_{13}, C_{14} and C_{15}; for the storm event, there are high spatial variance in these channels. The results in Conv 1 are further processed through deeper layers. Similar distinctions within same channel for two events can be depicted in Conv 2. Overall, the results here provide supportive evidences that the CNN models enhance the extraction of characteristic features by filtering the data with the learned kernels.

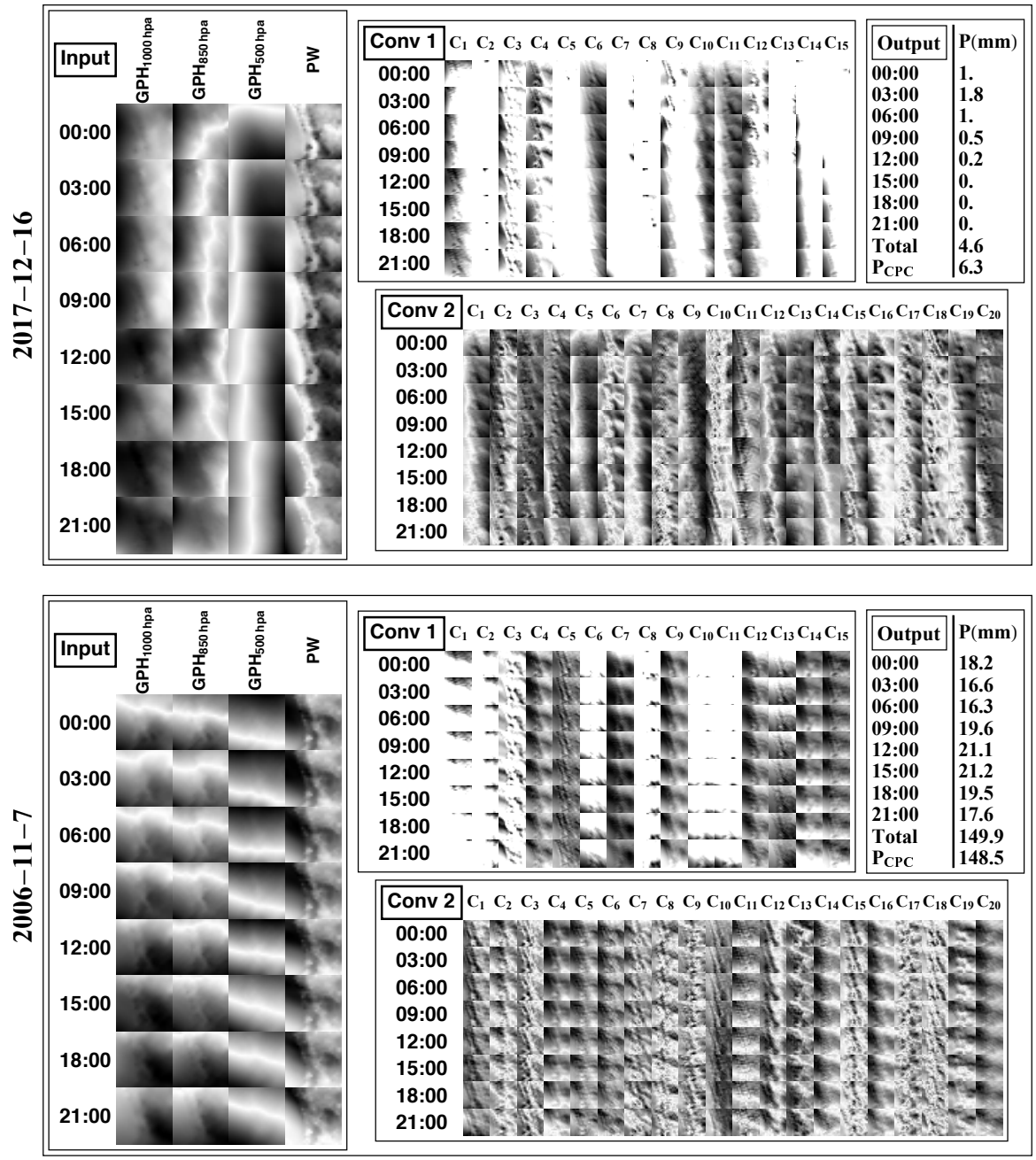


Figure 3.6: Layer activations for the December 16, 2017 light precipitation event (top) and November 7, 2006 storm event (bottom). The dark color represents low values and bright color represents high values. The left part shows the eight 3h snapshots of the dynamical field ($\text{GPH}_{1000\text{hPa}}$, $\text{GPH}_{850\text{hPa}}$, $\text{GPH}_{500\text{hPa}}$ and PW) through the day. Conv 1/2 shows the activated output for the first/second convolutional layer. The Conv 1 result is composed of 8×15 sub-figures. 8 indicates that there are eight 3h dynamical field snapshots; 15 indicates that the output is of 15 channels, which are labeled as from C_1 to C_{15}. Similar denotations for Conv 2. The Output panel shows the results by mapping the CNN to each 3h snap shot of the dynamical field. The sum of them consists the total daily precipitation estimate, which is compared against CPC records.

Perturbation Sensitivity

For image classification problems, the occlusion sensitivity analysis (OSA) tells the impact of different portions of the image on the classification result. It is performed by systematically occluding different portions of the input image with a grey square, and monitoring the output of the classifier [225]. The results of OSA illustrate if the model is effectively localizing the target object in the image, or just using the surrounding context.

For the problem here, I apply a similar method to quantify the precipitation-related impact of different portions of the circulation field. Rather than applying a grey square to occlude the input, Isystematically perturb the input field with a rescaling matrix. The dimension of the rescaling matrix is set to be the same as the receptive field for the first convolutional layer. All of its elements are set to $1+\epsilon$, where ϵ is the perturbation magnitude. The rescaling matrix is multiplied to different portions of the input. For each perturbation, Imonitor the model output change. Mathematically, this operation is roughly equivalent to estimating the partial derivatives of the precipitation estimate with respect to different portions of the dynamical field. The relation between perturbation location and model output change is visualized in Figure 3.7.

In Figure 3.7, as Islide the scaling matrix over different geolocations in different species of the dynamical fields, the model estimated precipitation amount changes correspondingly. For instance, for the 2007-11-7 storm event, the CNN model will produce larger precipitation estimate if the 1000hPa GPH for the central region is lower and the 1000hPa GPH for the surrounding area is higher. This is in accordance with our prior knowledge that heavy precipitations are related to intensive surface depressions. It is interesting to note that the perturbation sensitivity map occasionally present the characteristic appearances of cyclones. Overall, the precipitation estimations are highly sensitive to the dynamics from the central region of the field. This is the area where the target geogrid point lies. The surrounding dynamics also provide important context for inferring precipitation.

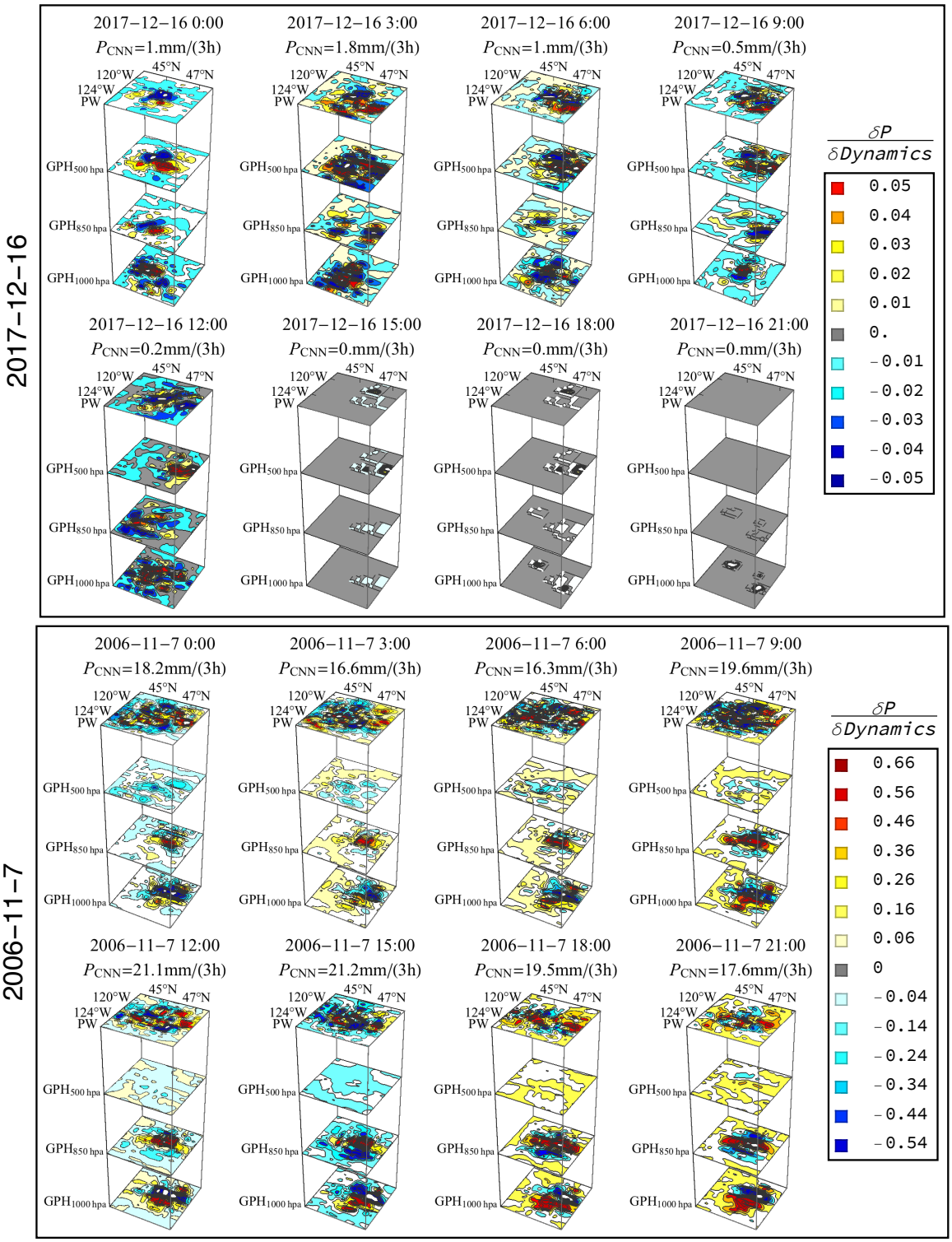


Figure 3.7: Perturbation sensitivity analysis for the December 16, 2017 light precipitation event (top) and November 7, 2006 storm event (bottom). For each case, Ivisualize the model output changes by systematically perturbing different portions of the scene with a rescaling matrix that is of same dimension as the first convolutional layer receptive field. The perturbation magnitude is set to 5%. The results are denoted as $\frac{\partial P}{\partial Dynamics}$. Iprovide clear 2D projections of these figures in the supplementary material.

3.7.3 Comparison Experiments

Previous sections have compared the CNN precipitation estimates with 1) NARR precipitation product, and 2) precipitation estimates using fully connected deep neural network. Results show that the convolution module enhances deep neural networks for precipitation estimate, achieving advanced performance compared to observation-adjusted numerical precipitation product.

In Section 3.3, I made a critical review on existing SD approaches, which motivates us to turn to CNN for explicit encoding of precipitation-related circulation geometric patterns. To justify the critics and the motivation, it is imperative to compare the CNN performance with classical SD approaches. Here Icarry out a series of comparison experiments using some of the widely-adopted SD methods.

The following SD models are selected as baselines : 1) linear regression, 2) nearest neighbor, and 3) random forest, all of which have been extensively applied and verified for SD tasks [99, 50, 79]. For each of the model, I adopt same input variables as for CNN, with optional feature extractions before feeding the input to the model. The data normalization, partition of training/validation/test set are the same as in the CNN experiment. The optional feature extraction is done using Principal Component Analysis (PCA). I carry out simulations using input composed of the leading 2, 8, 16, 64, and 256 PCs of the circulation field data, as well as simulations using the raw circulation field data. Details of the models and feature extraction are given in the appendix at the end of this chapter [145, 171, 110, 22, 215]. The precipitation estimation results for the test set are shown in Table 3.3.

Table 3.3: Precipitation estimation performance for the test set using 1) linear regression, 2) nearest neighbor, and 3) random forest model. For each model, Icarry out simulations using input composed of the leading 2, 8, 16, 64, and 256 principal components (PCs) of the circulation field data, as well as simulations using the raw circulation field data. The dimension for the input variable is labeled, for instance, 4×2 indicates that the leading 2 PCs for the $\text{GPH}_{1000\text{hPa}}$, $\text{GPH}_{850\text{hPa}}$, $\text{GPH}_{500\text{hPa}}$, and PW field are used as input. The r and RMSE score are used to measure model performance.

Input	Skill Score	Linear Regression	Nearest Neighbor	Random Forest
2 PCs	r	0.74	0.80	0.80
(4×2)	RMSE	8.01	7.35	7.23
8 PCs	r	0.79	0.81	0.81
(4×8)	RMSE	7.33	7.34	7.05
16 PCs	r	0.81	0.81	0.81
(4×16)	RMSE	6.98	7.50	7.09
64 PCs	r	0.81	0.58	0.80
(4×64)	RMSE	7.03	11.2	7.19
256 PCs	r	0.76	0.67	0.80
(4×256)	RMSE	8.86	12.47	7.19
Raw Input	r	0.52	0.79	0.81
$(4 \times 25 \times 25)$	RMSE	10.35	7.41	7.04

The best performance in the comparison experiments is achieved by the linear regression model using input of the leading 16 PCs of the circulation field ($r = 0.81$, RMSE = 6.98). The non-linear models outperform linear model when the input dimension is relatively low. As Iinclude more PCs as input, the skills for the models decrease (linear regression and nearest neighbor) or saturate (random forest).

The performance of the models here are comparable, or slightly better than the NARR precipitation estimates ($r = 0.76$, RMSE = 8.31). DNN with fully connected computation graph achieves better performance ($r = 0.82$, RMSE = 6.85). The skill can be further improved if Iapply the convolution and pooling modules to explicitly extract the spatial information from the high dimensional dynamical field ($r = 0.86$, RMSE = 6.13). To sum up, the comparison experiments empirically suggest that CNN is competitive in making precipitation estimations based on the resolved surrounding atmospheric dynamics.

3.8 Conclusion

Precipitation estimation provides fundamental information to better understand the land-atmosphere water budget, improve water resources management, and aid in preparation for increasingly extreme hydrometeorological events. However, the precipitation process is generally considered to be poorly represented in current numerical weather/climate models.

Statistical Downscaling (SD) approaches often provide more accurate precipitation estimates compared to the raw precipitation products in numerical models. However, Ipoint out two common deficiencies in adopting the existing precipitation SD approaches for daily precipitation estimation: Firstly, most existing SD approaches rely on human-engineered features to extract information from the raw form high-dimensional predictors, such as the principal components, however, the engineered features are often designed based on the characteristics of the predictors rather than the connections between the predictors and the precipitation process; secondly, the circulation geometries and positions that dominant precipitation distribution are not well disintegrated and represented.

The Convolutional Neural Network (CNN) model is introduced to overcome these two deficiencies in improving precipitation estimation. The CNN model stacks several convolution and pooling operators to extract the intricate but important circulation features for precipitation estimation. Instead of applying pre-engineered feature extractors, the model applies "end-to-end" learning. Specifically, the kernels that are used to extract the salient features from the resolved dynamical field are optimized by backpropagating the precipitation estimation error through the convolutional layers. Thus, the learned features are determined based on the relation between the predictors and the predictand for the exact learning target. Also, through hierarchical convolution, we can well disintegrate dominant circulation features of different geometric properties and from different locations.

The model is tested for 14 geogrid points that roughly cover different characteristic climate divisions of the contiguous United States. I use the every 3h geopotential height (GPH) and precipitable water (PW) field from the NARR dataset to provide realistic and

fine-scale dynamical forcing. The CNN models are implemented by connecting the dynamical forcings with the CPC gauge-based daily precipitation records. Considering the fact that each 3h snapshot of the dynamical field provides specific information for the particular time of a day, I map the CNN network to each 3h snapshot and sum up the results as the daily precipitation estimate. The parameters of the network are trained by minimizing the estimation error using backpropagation.

Results show that the CNN model outperform original NARR precipitation estimates for the west and east coast, where precipitation is more copious compared to the other areas. For the middle part of the continent, the CNN model show slightly worse performance, which can be attributed to model overfitting when there is limited precipitation samples for training the model.

Focusing on a single geogrid point, the influence of the network architecture on model performance is examined. Specifically, I focus on the receptive field and network depth. By varying the receptive field of the convolutional layers, I verify that the CNN model outperforms conventional fully connected ANN SD in estimating precipitation through explicit encoding of local spatial circulation structures. By varying the network depth, I found that deep networks generally have better performance compared to shallow networks. However, I also noticed the difficulty for training very deep networks.

To interpret the model, I visualize the activation of the middle layers of the network using a storm event and a light precipitation event. Results show that different channels are activated for the two cases of different dynamical condition and precipitation amount. I also implement the perturbation sensitivity analysis to quantify the precipitation-related impact of different portions of the dynamical field.

The model performance is compared with some of the widely-adopted SD methods, including linear regression, nearest neighbor, and random forest. Results show that the CNN model outperforms the baseline SD approaches for accurate precipitation estimates.

Overall, the CNN model shows impressive performance in estimating precipitation. The

CNN model applies hierarchical spatial convolution kernels to explicitly search the surround circulation field for precipitation-dominant dynamical patterns, followed by dense layers that relate the extracted dynamical patterns to the precipitation of a target grid.

Appendix

Perturbation Sensitivity Analysis

This section provides high resolution 2D figures of the perturbation sensitivity analysis results in Figure 7. Specifically, I systematically perturb the input fields with a rescaling matrix. The dimension of the rescaling matrix is set to be the same as the receptive field for the first convolutional layer. All of its elements are set to $1+\epsilon$, where ϵ is the perturbation magnitude (set as 5% here). The rescaling matrix is multiplied to different portions of the input. For each perturbation, Imonitor the model output change. Mathematically, this operation is roughly equivalent to estimating the partial derivatives of the precipitation estimate with respect to different portions of the dynamical field.

The rows in Figure 3.8 and 3.9 represent the hours of a day. The columns represent the different dynamical field. As I slide the scaling matrix over different geolocations in different species of the dynamical fields, the model estimated precipitation amount changes correspondingly.

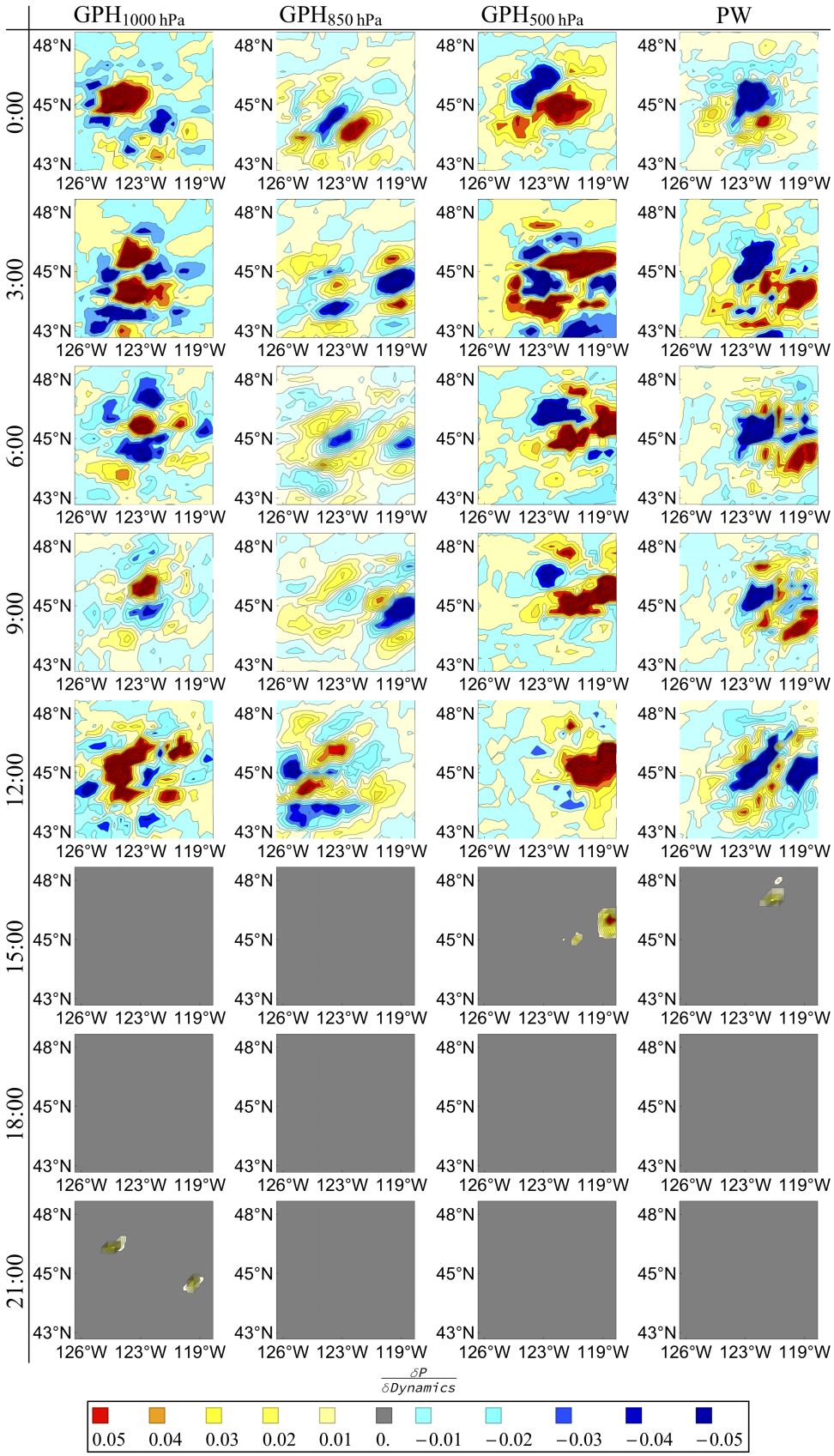


Figure 3.8: Perturbation sensitivity analysis for the December 16th 2017 light precipitation event.

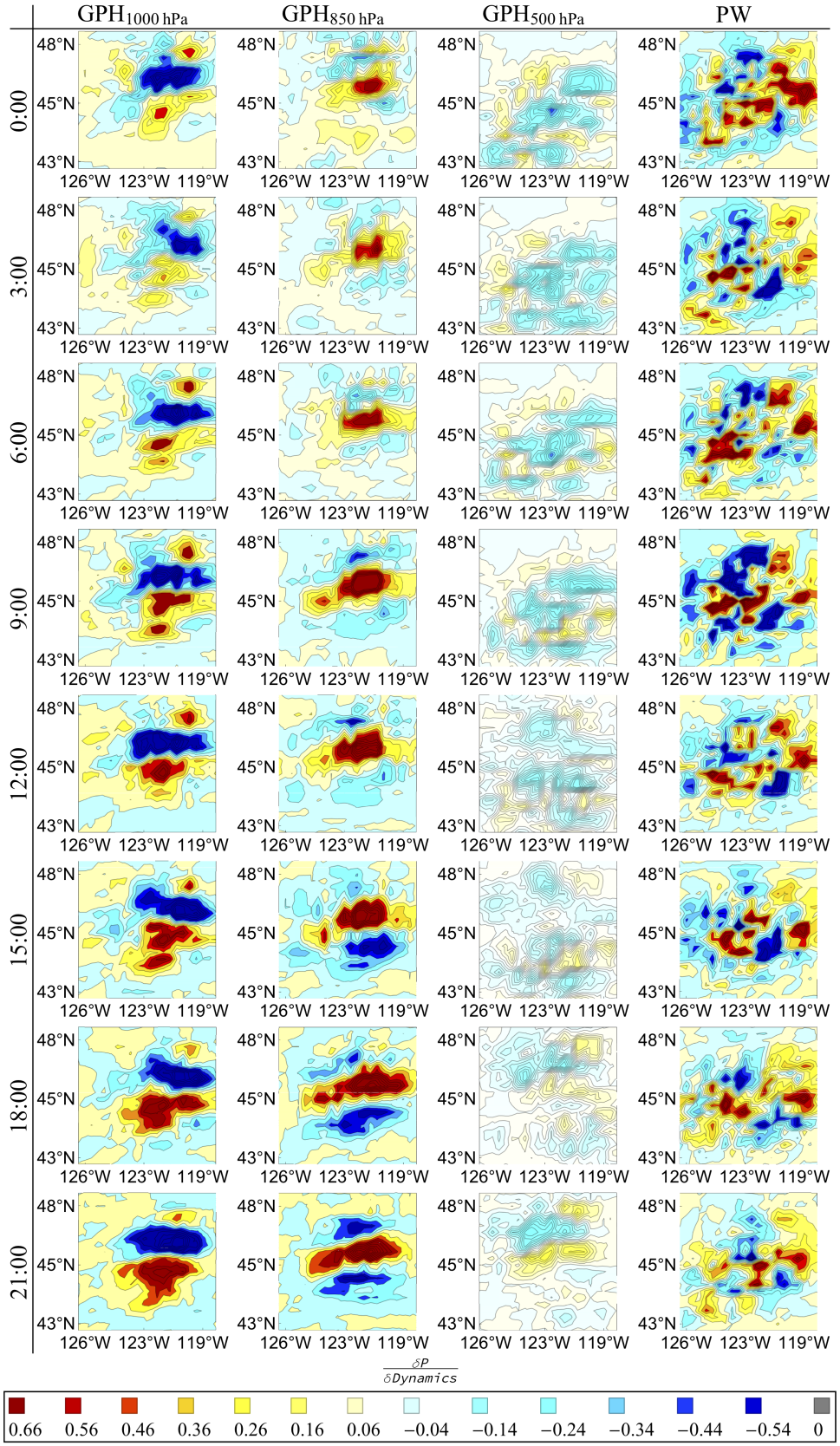


Figure 3.9: Perturbation sensitivity analysis for the November 7$^{\text{th}}$ 2006 storm event.

Baseline Models in Comparison Experiments

Principal Component Analysis

Principal Component Analysis (PCA) seeks a linear orthonormal transformation to "better" re-express high dimensional data [145, 171]. The objective is to filter out the noise and reveal the hidden dynamics based on a linear transformation. Specifically, let $\mathbf{X}$ be an $m \times n$ matrix, where $m \times n$ indicates n measurements for an m-dimensional vector space. Itry to propose an orthonormal transformation using an $m \times m$ matrix $\mathbf{P}$, so that, for the resulting $m \times n$ matrix $\mathbf{PX}$:

1. The correlation between different rows is 0.

2. The first row has the largest variance, each succeeding row in turn has the highest variance under the constraint that it is orthogonal to the preceding rows.

The row vectors are named Principal Components (PCs) of $\mathbf{X}$. The name of "principal" comes from the assumption that the directions with larger variances are considered to be more important in representing the variability of $\mathbf{X}$. Here I apply PCA to extract the leading PCs of the high-dimensional dynamical field (25×25 for each predictor variable) as potential input for precipitation estimation.

Consider the Precipitable Water (PW) field in the form of $25 \times 25 \times n$, where n is the sample size. Ifirst reshape the data into the form of $(25 \times 25) \times n = 625 \times n$ by flattening the spatial grids. The reshaped data matrix is denoted as $\mathbf{X}$. Itry to come up with a 625×625 transformation matrix $\mathbf{P}$, so that, the first row of $\mathbf{PX}$ holds the greatest variance among all 625 rows, the second row holds the second greatest variance, and so on. Also, different rows are linearly uncorrelated.

To estimate $\mathbf{P}$, let $\{\lambda_1, \lambda_2, ..., \lambda_{625}\}$ denote the eigenvalues of the covariance matrix for $\mathbf{X}$, i.e., $\mathbf{XX}^T$, λ is ordered from large value to small value. The corresponding orthonormal eigenvectors are $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{625}\}$. It can be easily proved that, if each row $\mathbf{p}_i$ for $\mathbf{P}$ equals to ${\mathbf{e}_i}^T$, then:

- The empirical variance for the ith row for $\mathbf{PX}$ is $\frac{\lambda_i}{n-1}$.
- The rows for $\mathbf{PX}$ are linearly uncorrelated.

Thus, for $\mathbf{PX}$, the first row holds the greatest variance, followed by the second row, and so on; also, the rows are uncorrelated. The ith row of $\mathbf{PX}$ is the ith PC for $\mathbf{X}$.

To extract the leading PCs for the GPH and PW field, for each variable, Ifirst estimate the transformation matrix of $\mathbf{P}$ based on the training set samples. Then, Imultiply $\mathbf{P}$ on the training/validation/test predictors to estimate PCs. The leading PCs are used as model input in corresponding comparison experiment. Figure 3.10 shows the variance of the leading 256 PCs for the $\text{GPH}_{1000\text{hPa}}$, $\text{GPH}_{850\text{hPa}}$, $\text{GPH}_{500\text{hPa}}$, and PW field.

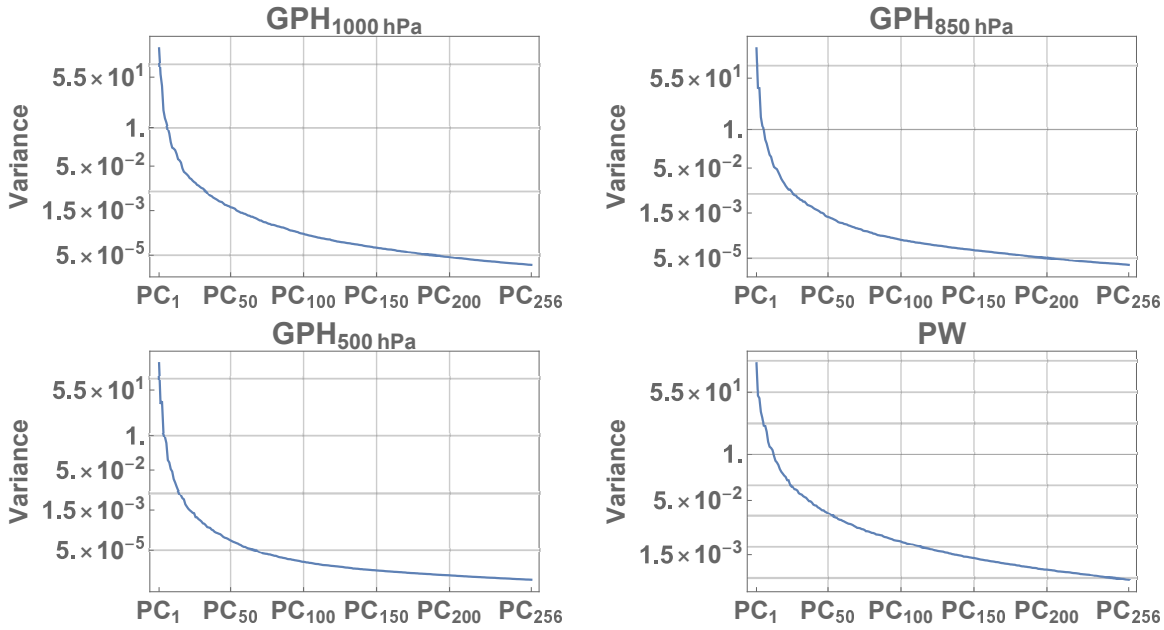


Figure 3.10: The variance (in logarithmic scale) for the leading 256 PCs of the $\text{GPH}_{1000\text{hPa}}$, $\text{GPH}_{850\text{hPa}}$, $\text{GPH}_{500\text{hPa}}$, and PW field.

Linear Regression

Consider estimating $\mathbf{y}$ based on a linear combination of $\mathbf{X}$:

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \epsilon$$

Here $\mathbf{y}$ and ϵ are n-dimensional vectors, $\mathbf{X}$ is a $n \times m$ matrix, $\mathbf{b}$ is m-dimensional vector; $\mathbf{y}$ is the predictand, $\mathbf{X}$ is the predictor, $\mathbf{b}$ is the parameters that need to be calibrated, ϵ is the estimation error; n denotes that there are n observations, m indicates the dimension of the input variable is m. Iapply the least square error criterion to estimate the optimal $\hat{\mathbf{b}}$. The square error (E) is represented as follows:

$$E = \epsilon^T \cdot \epsilon = (\mathbf{y} - \mathbf{X} \cdot \mathbf{b})^T \cdot (\mathbf{y} - \mathbf{X} \cdot \mathbf{b})$$

To obtain optimal $\hat{\mathbf{b}}$, Irequire $\frac{\partial E}{\partial \hat{\mathbf{b}}} = 0$, thus:

$$\frac{\partial E}{\partial \hat{\mathbf{b}}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}} = 0$$

which gives:

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Nearest Neighbor

Nearest Neighbor is among the simplest and most popular form of machine learning approaches. It predicts the output for an input using the average output value for the k closest training examples in the feature space. Here Iset the k value as 4. I use Euclidean distance to measure the distance between samples.

Random Forest

Random forests (RFs) are prediction models based on ensembles of decision trees (DTs) grown from a randomized variant of the tree induction algorithm [110]. Compared to DTs, RFs usually make better predictions by alleviating the problem of overfitting. Regarding the aspect of how random perturbations are introduced into the induction procedure for building DTs, there are many RF variants. Here Iadopt the classical Breiman RF [22], which

combines *bagging* and *random variable selection* for growing trees. To explain Breiman RF, Ifirst introduce the basic building-blocks of DTs, bagging, random variable selection, then Iillustrate how these blocks are assembled to consist the RF algorithm. For comprehensive description and explanation of RF, see [22] and [110].

Decision Tree: Decision trees apply directed acyclic graph to recursively partition the input space into sub-spaces and assign same prediction value to each terminal subspace. To fit a DT for a specific problem, Ishould clarify:

1. The computation strategy for splitting the tree nodes, starting from the root node of the input data.

2. The criteria of when to stop splitting.

For the first aspect, we usually adopt a greedy searching strategy: we seek an optimal partition that maximizes the "information gain" for current split step. For the precipitation regression task here, I use the squared error loss as the "information gain" measurement.

For the second aspect, I pre-define a set of hyper-parameters to determine the stopping criteria. The hyper-parameters include: 1) the minimum sample size for each node, 2) the maximum depth for the DT, 3) the minimum decrease of squared error loss required for adding a partition. To apply these stopping criteria, Ifirst fully develop all nodes by recurrently split the predictor space, then, I apply these criterion to prune the over-complicated trees.

Bagging: The *bagging* technique, which is also known as *bootstrap aggregating*, refers to the strategy of selecting bootstrap samples from the training examples. The samples are drawn at random with replacement. We can build DTs based on each bootstrap sample. The resulting DTs are averaged to reach the ensemble prediction.

Random Variable Selection

For problems with high dimensional predictors, to find the optimal split for each node of a DT, we might consider making the partition based on many random subsets of the

predictor dimensions. The resulting trees are thus structurally different, since they make predictions based on different aspects of the predictor. This strategy is known as *random variable selection*.

The Breiman Random Forest Model: The Breiman RF model assembles the building-blocks introduced above. The model combines many different DT into an ensemble, and introduces random variation with *bagging* and *random variable selection*. The resulting ensemble of trees is averaged to produce an overall prediction, which reduces overfitting while allowing for complex individual learners. The specific steps for building the Breiman RF model is listed as follows:

1. Pick random sample of size n with replacement from the data. Each of the n samples construct a bootstrap sample space.

2. Build regression tree based on each subset of bootstrap samples. When picking the best split for a node, a random subset of input dimensions is selected to be searched over, rather than finding the best split across all input dimensions.

3. Make prediction using the mean of n trees' predictions.

I use the *Wolfram Mathematica* machine learning toolbox to implement the PCA, linear regression, nearest neighbor, and random forest algorithm [215].

Chapter 4

Benchmarking Quantitative Precipitation Forecast using a Composite of Numerical Modeling and Deep Neural Networks

4.1 Background

Empirical experiments in Chapter 3 suggest that, by learning from sufficient examples, a deep convolutional neural network can effectively extract useful features from the resolved dynamical field estimates for more accurate precipitation predictions [141]. This promising result at daily, grid scale motivates a more ambitious exploration of deep neural network's potential for quantitative precipitation forecast (QPF), which informs the expected amount of precipitation accumulated over a specified time period over a specified area [26], usually at a much higher spatiotemporal resolution.

QPF is crucial for both practical applications as well as the modeling and understanding of the climate system. From an application perspective, accurate QPF benefits various

practices ranging from flash flood forecast to long-term ecological, agricultural, and water resources management [94]. From a modeling viewpoint, QPF is widely considered as one of the most stringent challenges for numerical weather/climate modeling, making it an ultimate criterion for model evaluation and diagnosis [184]. Besides, accurate QPF consists a primary requisite for modeling the land processes and land-atmosphere interactions in coupled general circulation models (GCMs).

GCMs, together with numerical weather prediction models (NWPMs), are arguably the only reliable tools for QPF at short to extended ranges. Boosted by 1) consistent growth in computation capacity, 2) accumulation in understandings of atmospheric dynamics and physics, and 3) development of modern observation networks and data assimilation techniques, numerical models have demonstrated impressive skills in capturing the spatiotemporal variation of precipitation at relatively coarse resolution [9, 143]. However, current models usually fail in revealing many critical details of the precipitation processes, such as location, timing, intensity, or total accumulation [179, 184].

As our focus shifts from coarse to fine scale, precipitation tends to be more patchy in space, and more spiky in time. These irregular characteristics stem from the manifestation of individual formation and growth of precipitating clouds, which exhibit complex coupling with their embedding atmospheric fluid dynamics [78]. In current GCMs and NWPMs, the cloud and precipitation processes can not be explicitly resolved, and are mostly parameterized based on a mix of empirics, phenomenological laws, and closing assumptions. Deficiencies in parameterizing the relevant unresolved processes contribute to a major source of uncertainty in numerical modeling, and have attracted significant research interests from the modeling community [178].

Efforts in improving precipitation related parameterizations are supposed to benefit from a clarification of predictability limit under specific model configurations and at specific measurement scopes. Predictability quantifies the extent to which it is possible to predict a system within the constraints of uncertainties from input and model formulation [140]. To

clarify what is possibly predictable offers insights into the sources of prediction limits, therefore directs further progresses to mitigate these deficiencies. Here, I focus on a particular aspect toward understanding precipitation predictability: Given the atmospheric dynamics that are realistically resolved at certain resolution, what is the predictability limit of the precipitation process?

It is difficult to tackle the proposed problem based on a process-based approach. This is because, the inevitable errors from 1) initial field estimations, 2) dynamical forcings, and 3) parameterizations exhibit complex interactions through model's iterative dynamical simulations, with all sources of errors quickly revealing themselves in models' precipitation outputs [184]. Besides, at microphysical scale, precipitation is typically simulated as a diagnostic variable that results from prognostic simulations of various hydrometeor development processes, while we do not have direct observations to verify the estimates of the latent state variables. Overall, to track, attribute and rectify the precipitation prediction errors has been widely considered as an extremely daunting task.

Here, I bypass explicit considerations of complicated hydrometeor developments at cloud microphysics scale, and focus only on the specific aspect of predicting the precipitation process. I investigate precipitation predictability by seeking statistical connections between the high resolution precipitation observations and their associated atmospheric dynamical and moisture analysis. It should be noted that the objective is not to propose a "black-box" statistical model to replace existing cloud microphysics and cloud cumulus parameterization schemes. Instead, I try to clarify to what extend of accuracy QPF can be as viewed from a data-driven perspective, thus offering directions for improving precipitation related parameterizations. The strategy here basically follows what is proposed in Chapter 3, but brings particular opportunities and challenges. Specifically, provided with consistently observed high resolution precipitation and atmospheric analysis data, the amount of data for model training and validation would be significantly increased. On the other hand, high resolution precipitation data are more irregular in their spatiotemporal distribution, and the

observations are not consistently available, which requires careful treatment in modeling.

The rest of this chapter proceeds as follows: Section 4.2 builds a unique long-range hourly precipitation dataset by collecting, processing, and cleaning data from various sources. Based on this dataset, Section 4.3 tests the applicability of a suite of deep neural network architectures for estimating precipitation process at hourly, point scale. Section 4.4 uses dynamical forecast experiments to further testify the model's robustness and applicability in real world QPF tasks. Conclusions are drawn in Section 4.6.

4.2 Study Area and Data

4.2.1 Study Area

The West Coast of the United States is selected as the study area. The spatial coverage and elevation map are shown in Figure 4.1. As have been explained in Chapter 2, this region receives a majority of its precipitation during the cool season. The precipitation system tend to be strongly synoptically forced by extratropical cyclones. Strong precipitation forms as the cyclone cold front sweeps up water vapor in the warm sector of the cyclone, causing a narrow band of high water vapor content to form ahead of the cold front at the base of the warm conveyor belt airflow [33]. Such a phenomenon is manifested as filaments of enhanced water vapor in satellite imageries, and is termed "atmospheric river" (AR). ARs have been recognized as the major storm sources [147] and drought busters [36] for the West Coast, contributing to approximately 30%–50% to its annual precipitation accumulation [37]. The spatiotemporal scale of this precipitation mechanism is explicitly considered in preparing the data and designing the proposed model. Dynamical simulation of two typical AR events are also carried out for detailed verification and analysis, which will be described in detail in later sections.

4.2.2 Data Sources and Dataset Construction

Hourly Precipitation Observations

High spatiotemporal resolution precipitation records are valuable for a wide range of applications, including hydraulic infrastructure design [138], climate monitoring and variability assessment [146]. However, these data are usually not consistently available due to their high cost. Here I compare hourly precipitation observations from various sources and concatenate the quality controlled data to compose a peculiar long term (1980/1/1- 2018/12/31) hourly precipitation observation dataset. The data are then applied for training and validating the proposed models.

Ground-based gauges, satellite remote sensing, and radar are three of the major sources for high resolution precipitation monitoring. Notable discrepancies may arise among their estimates for specific observation areas and periods [175, 134], making quality control a necessity. Here I use reanalysis precipitation products as reference to conduct the quality control. The motivation is justified as follows: I admit the fact that precipitation estimations from atmospheric reanalysis may have severe deficiencies due to parameterizations and their associated uncertainties; in fact, to improve model's precipitation products constitute a major target for this research work. On the other hand, atmospheric reanalysis applies an unchanging data assimilation scheme to systematically ingest available observations into dynamical simulations, which hold fixed configuration as well. The disparities between reanalysis precipitation and the "true" precipitation can thus be considered to fluctuate within an unchanging envelope of variability. I infer this disparity based on simple skill metrics (i.e., correlation coefficient (r) and root mean square error (RMSE)) between precipitation reanalysis and a solid hourly precipitation observation dataset, namely the NOAA (National Oceanic and Atmospheric Administration) CPC (Climate Prediction Center) Hourly US Precipitation dataset [66] (referred as P_{CPC} thereafter). Similar measurements between reanalysis precipitation and other sources of precipitation observations are calculated for

periods beyond the coverage of P_{CPC}. The results are applied for determining the qualified data. The specific data sources and quality control results are illustrated in the following part.

As is clarified, the NOAA CPC Hourly US Precipitation dataset (P_{CPC})[66] is employed as referential precipitation observations for quality control, model training and verification. The data are obtained by gridding quality-controlled hourly-scale station observations into $2^\circ \times 2.5^\circ$ boxes. The data cover 140°W – 60°W , 20°N – 60°N, from 1948 to 2002.

For period from 2002 to 2018, I consider three candidate hourly precipitation datasets. The first is gauge-based dataset (referred as P_{Gauge} thereafter), which consists hourly precipitation observations from approximately 3000 automated rain gauges across Contiguous United States and parts of other regions in North America. The data are collected and organized by the National Centers for Environmental Prediction (NCEP) in cooperation with the Office of Hydrology (OH). I obtain data that cover 1995 to 2018 from the following website:

https://rda.ucar.edu/datasets/ds507.5/

For most gauges, there are unavoidable occasional discontinuities and anomalies, which have been systematically labeled out. The gauge distribution and observation frequency are displayed in Figure 4.1.

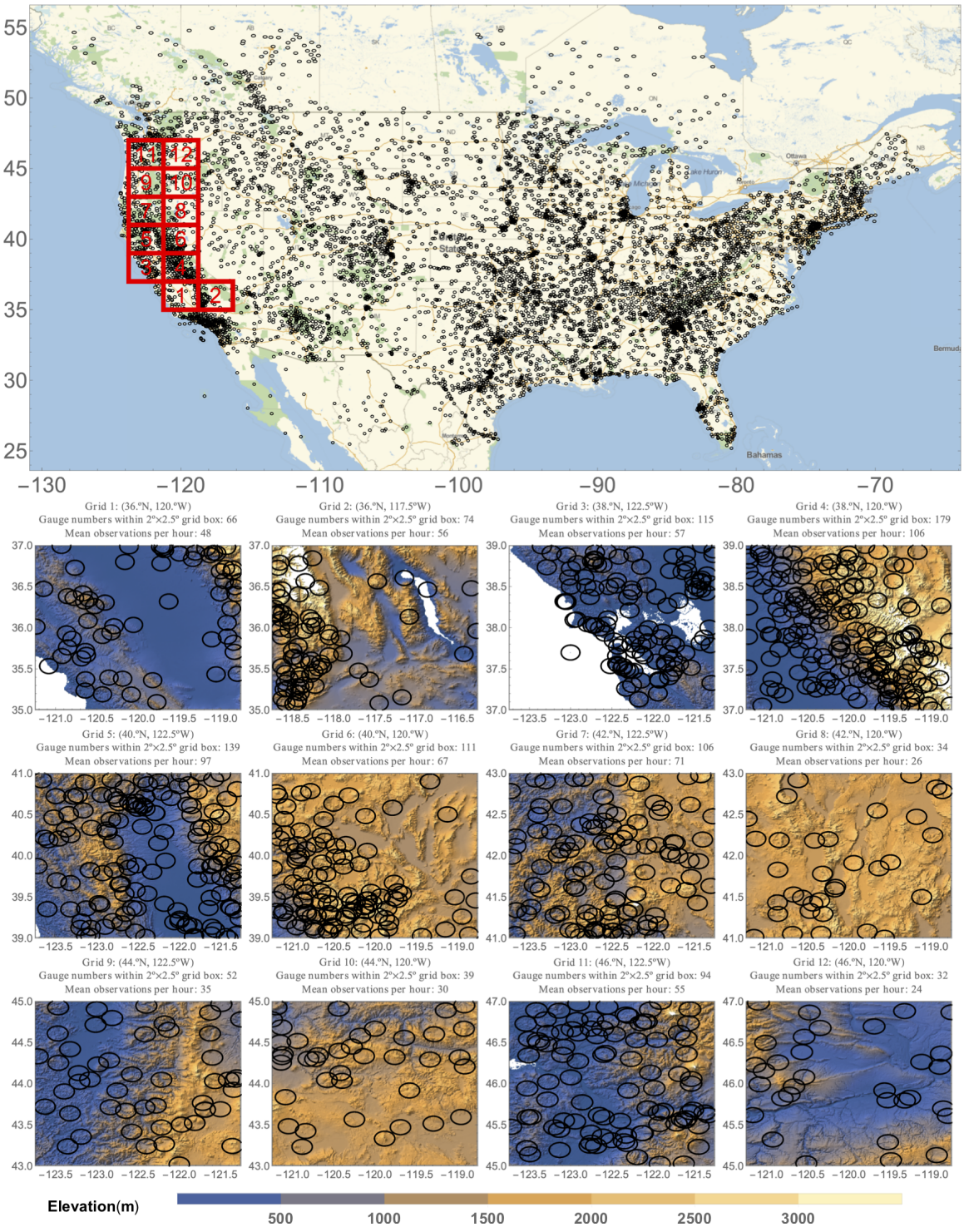


Figure 4.1: Top: Distribution of precipitation gauges across North America. Gauges are labeled with 10 km range rings. The red grids indicate the study regions. Bottom: detailed gauge distribution for the 12 grids labeled with 10 km range rings. The total gauge number and mean number of available observations per hour within each grid box are denoted. Background color indicates elevation, for which the data are obtained from United States Geological Survey [53].

The second dataset is the Stage IV gauge-adjusted precipitation product from National Centers for Environmental Prediction (NCEP) Environmental Modeling Center (EMC) [104] (referred as $P_{StageIV}$ thereafter). The data are based on the high-resolution Doppler Next Generation Weather Radar (NEXRAD) network [47] and the National Weather Service (NWS) River Forecast Center (RFC) precipitation processing system [165]. For the study area, the RFC precipitation processing system adopts the Mountain Mapper approach [76, 134], which adjusts radar estimation for climatological variations due to topography and wind directions. The data are obtained from the following website:

https://data.eol.ucar.edu/dataset/21.093

The hourly, 4 km resolution data from 2002/10/1 to 2018/12/31 are used here. The grid data contain missing observations due to topographical blockings or operational issues. The mean coverage of available mosaicked observations is shown in Figure 4.2.

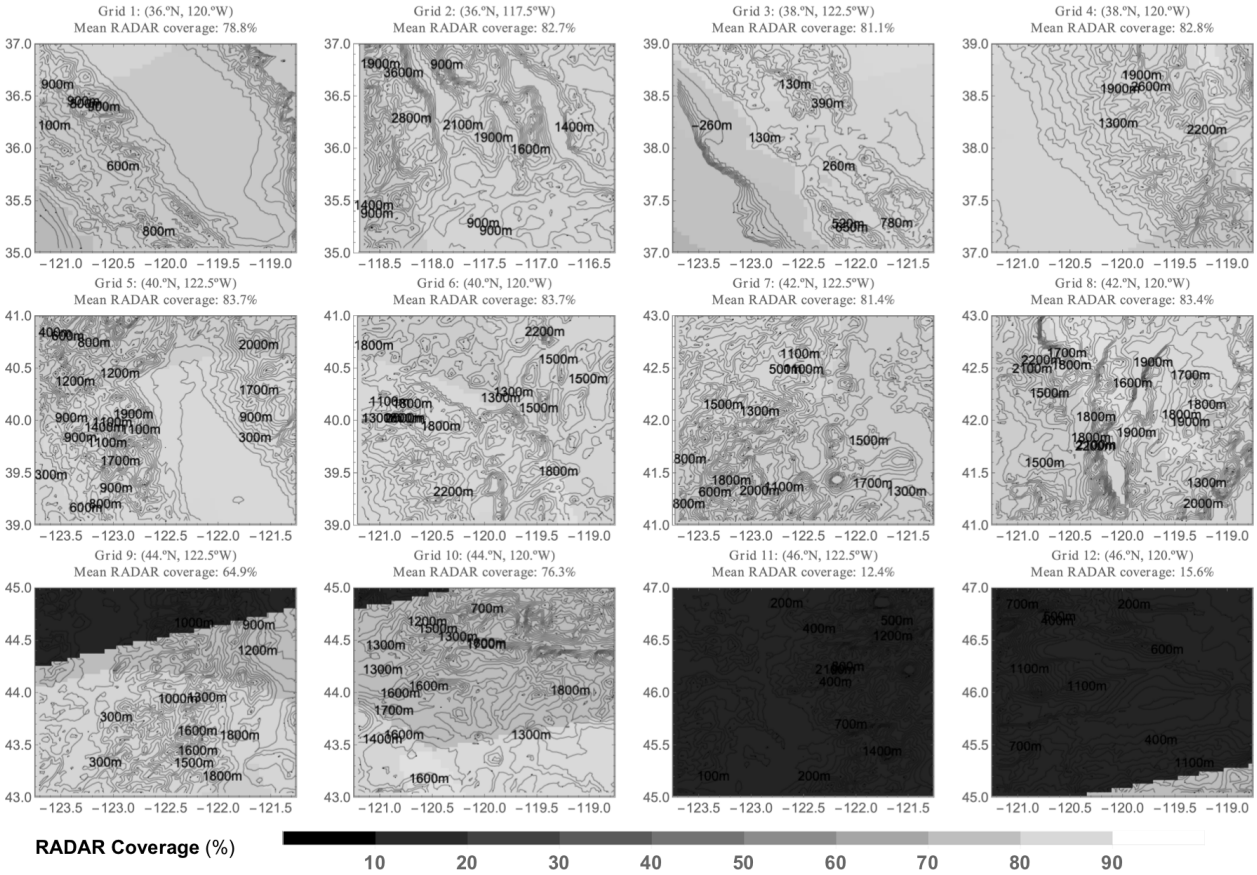


Figure 4.2: Mean coverage of available mosaicked observations from Stage IV precipitation product for the study area. Contours show the elevation.

The third and last dataset is satellite remote sensing product from PERSIANN-CCS (Precipitation Estimation from Remotely Sensed Imagery using an Artificial Neural Networks-Cloud Classification System [72], referred as $P_{Satellite}$ thereafter). PERSIANN-CCS extracts local and regional cloud features from infrared (10.7 μm) geostationary satellite imagery in estimating fine scale ($0.04^\circ \times ^\circ 0.04$, every 30 min) rainfall distribution. The data are obtained from the data portal built by Center for Hydrometeorology and Remote Sensing, University of California, Irvine [135, 136]: http://chrsdata.eng.uci.edu.

I apply two precipitation reanalysis products from NASA (National Aeronautics and Space Administration) MERRA-2 (Modern-Era Retrospective analysis for Research and Applications, Version 2) project as reference to benchmark the candidate precipitation products. MERRA-2 is the latest atmospheric reanalysis produced by NASA GMAO (Global Modeling

and Assimilation Office). Its main components are the GEOS (Goddard Earth Observing System) atmospheric dynamical model [125] and the 3-D variational GSI (Gridpoint Statistical Interpolation) assimilation system [92]. The GSI assimilation system combines disparate historical meteorological observations with underlying atmospheric dynamical simulations in a physically consistent manner, enabling production of realistic, gridded dataset for past climate [51]. The dataset covers the whole globe from 1980 to present, with spatial resolution of $0.5° \times 0.625°$. There are two sets of hourly precipitation analysis products from MERRA-2, namely the raw model's precipitation estimation (referred as P_{MERRA2}) and the bias-corrected estimation (referred as P_{MERRA2_C}). P_{MERRA2} is generated by the GEOS atmospheric dynamical models within the cycling MERRA-2 system, while P_{MERRA2_C} is generated by merging and disaggregating quality-controlled daily-scale observations to P_{MERRA2}. The bias correction details can be found in [150].

To perform the quality control, I first upscale all the precipitation products to $2° \times 2.5°$ to match the spatial resolution of P_{CPC}. Then, P_{CPC}, P_{Gauge}, $P_{StageIV}$, and $P_{Satellite}$ are compared against P_{MERRA2} and P_{MERRA2_C}. The comparison is based on skill score of r and RMSE. Results are summarized in Table 4.1.

	Comparing with P_{MERRA2}								Comparing with P_{MERRA2_C}							
	r				RMSE				r				RMSE			
	P_{CPC}	P_{Gauge}	$P_{StageIV}$	$P_{Satellite}$	P_{CPC}	P_{Gauge}	$P_{StageIV}$	$P_{Satellite}$	P_{CPC}	P_{Gauge}	$P_{StageIV}$	$P_{Satellite}$	P_{CPC}	P_{Gauge}	$P_{StageIV}$	$P_{Satellite}$
Grid 1	0.54	0.49	0.26	0.03	0.22	0.25	0.26	0.52	0.59	0.78	0.37	0.04	0.19	0.14	0.18	0.48
Grid 2	0.42	0.28	0.05	0.00	0.17	0.23	0.25	0.49	0.47	0.56	0.07	0.01	0.15	0.17	0.22	0.47
Grid 3	0.62	0.58	0.34	0.08	0.34	0.36	0.39	0.55	0.63	0.73	0.44	0.1	0.3	0.26	0.31	0.50
Grid 4	0.66	0.7	0.13	0.11	0.51	0.5	0.73	0.74	0.64	0.76	0.14	0.11	0.31	0.24	0.5	0.53
Grid 5	0.5	0.53	0.19	0.21	0.33	0.35	0.37	0.43	0.57	0.79	0.28	0.28	0.35	0.26	0.41	0.45
Grid 6	0.27	0.25	0.07	0.08	0.26	0.36	0.3	0.47	0.54	0.64	0.17	0.16	0.22	0.28	0.29	0.46
Grid 7	0.53	0.61	0.06	0.31	0.29	0.29	0.64	0.4	0.62	0.78	0.08	0.34	0.22	0.19	0.6	0.35
Grid 8	0.38	0.41	0.10	0.25	0.21	0.2	0.22	0.37	0.43	0.54	0.14	0.27	0.17	0.12	0.14	0.35
Grid 9	0.67	0.76	0.17	0.16	0.42	0.39	0.59	0.61	0.71	0.83	0.19	0.19	0.3	0.23	0.42	0.47
Grid 10	0.39	0.38	0.07	0.14	0.25	0.24	0.25	0.38	0.42	0.47	0.08	0.17	0.16	0.16	0.15	0.34
Grid 11	0.62	0.7	0.19	0.08	0.52	0.46	0.67	0.7	0.66	0.77	0.2	0.08	0.42	0.31	0.51	0.55
Grid 12	0.32	0.32	0.09	0.08	0.23	0.21	0.2	0.34	0.43	0.48	0.14	0.10	0.18	0.14	0.12	0.31

Table 4.1: Comparing MERRA2 precipitation products (raw/bias corrected, denoted by P_{MERRA2}/P_{MERRA2_C}) with precipitation observations from (1)NOAA's CPC Hourly US Precipitation dataset (P_{CPC}), (2) gauge precipitation product from NCEP and OH (P_{Gauge}), (3) NWS/NCEP stage IV precipitation product ($P_{StageIV}$), and (4) Remote sensing precipitation from the PERSIANN-CCS ($P_{Satellite}$). All data are of hourly scale. I select P_{MERRA2} and P_{MERRA2_C} that cover time period 1980-2018 with spatial resolution of $0.5° \times 0.625°$., P_{CPC} covers period of 1980-2002 with spatial resolution of $2° \times 2.5°$. $P_{StageIV}$ and P_{Gauge} cover period of 2002-2018. Resolution of $P_{StageIV}$ and $P_{Satellite}$ are 4 km. P_{Gauge} contains point-wise observations. All data except P_{CPC} are spatial averaged to $2° \times 2.5°$ for comparison.

Among P_{Gauge}, $P_{StageIV}$, and $P_{Satellite}$, P_{Gauge} shows largest and most consistent agreement with P_{CPC} in comparison with P_{MERRA2}/P_{MERRA2_C}. Also, P_{CPC} and P_{Gauge} have better r and RMSE score, compared to remote sensing estimations from $P_{StageIV}$ and $P_{Satellite}$. Previous works have highlighted that radar network scarcity and mountain blockage introduce significant inaccuracy for radar precipitation estimation in the West Coast [113], which is confirmed from the results here. $P_{Satellite}$ infers precipitation from indirect and limited data sources (i.e., cloud top temperature), and generally shows little skill at our measurement scope. However, it is noteworthy that the skill for $P_{Satellite}$ is promising for certain regions, such as for Grid 7 and Grid 8. To sum up, results here demonstrate that P_{Gauge} is the most consistent data with P_{CPC}. Thus, I concatenate P_{CPC} (from 1980–2002) and P_{Gauge} (from 2002–2018) to compose the long term hourly precipitation dataset.

4.2.3 Precipitation Events Segmentation

Precipitation appears as impulse event in nature. The precipitating process typically lasts for several hours up to multiple days once atmospheric water condenses and starts to precipitate. In between of two precipitation events is a no-precipitation period. Only the data from precipitation periods are considered to provide informative samples for learning QPF, while data from no-precipitation periods provide little informative information. The inclusion of non-informative data is believed to hinder efficient and effective model training. This motivates me to construct a precipitation event dataset by extracting precipitation events from consecutive historical climate series. A precipitation event segmentation algorithm is carried out by filtering the long series of hourly precipitation records based on the following two criteria:

1. The start and end of a precipitation event are featured by 24 consecutive hours with precipitation rate less than 0.01 mm/hour.

2. The maximum precipitation rate for a precipitation event should be larger than 0.5

mm/hour.

4.2.4 Atmospheric Dynamics

I use the realistically estimated atmospheric dynamical and moisture field data as predictors for precipitation estimation. The data are obtained from NASA MERRA-2 project as well. Details of the dataset has been introduced in Section 4.2.2.

I select the following variables from the MERRA-2 dataset to represent the dynamical and moisture forcings: the geopotential height (GPH) at 1000 hPa, 850 hPa, and 500 hPa, the total column water vapor (TQV), liquid water (TQL), and ice water (TQI). For each $2^\circ \times 2.5^\circ$ grid box, I cut out the corresponding 3-D dynamical/moisture field that centers around the target region and covers 25×25 $0.5^\circ \times 0.625^\circ$ grids. The resulting $n \times 6 \times 25 \times 25$ tensor is applied to estimate the precipitation process for a n-hour precipitation event.

4.3 Deep Neural Network for Precipitation Estimation

4.3.1 Problem Formulation

The precipitation prediction problem is framed as a statistical regression task, i.e., estimating precipitation at hourly, gauge-point scale based on the surrounding atmospheric dynamics and moisture field estimates. Considering the fact that most precipitation events in the study area are associated with synoptic scale extratropical cyclones [60], gauge observations within each $2^\circ \times 2.5^\circ$ geogrid box (delineated with red boxes in Figure 4.1) are estimated based on a same regression model that relates synoptic circulation patterns to the spatial distribution of precipitation. Considering the coherent life cycle of extratropical cyclongenesis processes, the regression model takes input of a whole time sequence of the input field from a precipitation event, and generates output of precipitation time series for corresponding period. The general

form of the regression model is written as follows:

$$E(\mathbf{P}_{t_1:t_n}|\mathbf{X}_{t_1:t_n}) = f(\mathbf{X}_{t_1:t_n};\boldsymbol{\theta}) \tag{4.1}$$

Here E denotes probabilistic expectation, $\mathbf{P}_{t_1:t_n}$ denotes the precipitation time series for a precipitation event from Hour t_1 to t_n. $\mathbf{P}_{t_i}$ is the precipitation observation vector at t_i. Ideally, $\mathbf{P}_{t_i}$ is preferred to include valid observations from all gauges within the considered geogrid box at t_i. f is the functional form of the model. $\mathbf{X}_{t_1:t_n}$ is the atmospheric dynamical field estimates from t_1 to t_n. Specifically, $\mathbf{X}_{t_i}, i \in [1, n]$ is a $6\times25\times25$ tensor, which represents 6 channels of dynamical/moisture feature distributions over 25×25 $0.5° \times 0.625°$ geogrids. $\boldsymbol{\theta}$ are the parameters of f, which should be calibrated using data. The data are composed of atmospheric analysis and precipitation record data pairs: $\{\mathbf{X}^k_{t_1:t_{n^k}} \rightarrow \mathbf{P}^k_{t_1:t_{n^k}}|k = 1, 2, ...\}$, k indexes precipitation event. The learning process applies the data to optimize f and $\boldsymbol{\theta}$ in order to make the model perform well for predicting future precipitation events.

4.3.2 Loss Function

A loss function, denoted by L, quantifies the difference between model predictions and "ground-truth" observations of the predictand. L measures model performance and directs the learning process. In our problem setting, observations are not consistently available. For period from 1 January 1980 to 30 September 2002, we only have grid spatial average precipitation record from NOAA CPC, for period from 1 October 2002 to 31 December 2018, there are frequent missing records for most gauges. To make full use of the available data, I customize our loss function based on a "masked" mean square error metric. Specifically, I assign each precipitation observation vector $\mathbf{P}_{t_i}$ a mask vector $\mathbf{M}_{t_i} = \{m^1_{t_i}, m^2_{t_i}, ..., m^G_{t_i}\}^T$. Here G denotes the number of gauges within the considered $2° \times 2.5°$ grid box; $m^j_{t_i} = 1$ if observation is available for the jth gauge at time t_i, otherwise $m^j_{t_i} = 0$. With the mask

vector clarified, I define L that balances spatial distribution estimation error and spatial mean estimation error:

$$L = \frac{1}{K}\sum_{k=1}^{K}\left\{\frac{1}{n_k}\sum_{i=1}^{n_k}\left[(\mathbf{P}_{t_i}\cdot\mathbf{M}_{t_i} - \hat{\mathbf{P}}_{t_i}\cdot\mathbf{M}_{t_i})^2 + \lambda(||\mathbf{P}_{t_i}\cdot\mathbf{M}_{t_i}||_1 - ||\hat{\mathbf{P}}_{t_i}\cdot\mathbf{M}_{t_i}||_1)^2\right]\right\} \quad (4.2)$$

K is the number of precipitation events, n_k is the duration hour of the kth precipitation event. $\mathbf{P}_{t_i}/\hat{\mathbf{P}}_{t_i}$ is n dimensional model's precipitation estimation/observation vector at Hour t_i, n is number of gauges. $\mathbf{M}_{t_i}$ is the mask vector at t_i. $||\cdot||_1$ is L1 norm operator that outputs sum of the absolute values of the input vector. Thus, $||\mathbf{P}_{t_i}\cdot\mathbf{M}_{t_i}||_1/||\hat{\mathbf{P}}_{t_i}\cdot\mathbf{M}_{t_i}||$ represents the spatial average estimation/observation at t_i. The first term on the right side represents spatial distribution estimation error, the second term represents the spatial average estimation error. λ is a weighting factor that balances these two error terms.

4.3.3 Deep Neural Network Models

Artificial neural networks (ANNs) cover a continually-evolving family of machine learning methods, which approximate complicated functions through composing simple functions in hierarchical computing graphs. I use a simple example of multilayer perceptron (MLP) to illustrate the basic concepts of ANN and introduce the proposed models. A MLP is arguably the most popular class of ANN. Mathematically, a n-layer MLP is a chain of matrix multiplications and element-wise non-linearities:

$$\begin{aligned}&\mathrm{MLP_n}(\mathbf{X}) = \mathbf{f}^{\{n\}}(\cdots(\mathbf{f}^{\{2\}}(\mathbf{f}^{\{1\}}(\mathbf{X})))\cdots)\\&\mathbf{f}^{\{i\}}(*) = g^{\{i\}}[\mathbf{W}^{\{i\}}\cdot * + \mathbf{b}^{\{i\}}]\end{aligned} \quad (4.3)$$

$\mathbf{X}$ denotes the input vector; $\mathbf{f}^{\{i\}}$ is the ith layer operator that transforms the representation of the data using a linear transformation (through matrix multiplication $\mathbf{W}^{\{i\}}$), shifting (through bias vector $\mathbf{b}^{\{i\}}$), followed by an element-wise non-linear function (through element-

wise nonlinear activation function $g^{\{i\}}$). Since the model is differentiable, we can apply gradient descent method to minimize a pre-defined loss function in order to make the model exhibit desired behavior. This process is widely known as backpropagation training [155].

Deep neural networks (DNNs) are ANNs with multiple hidden layers. By composing many layers that each transforms data representation at one level into a representation at a higher, slightly more abstract level [96], DNNs can automatically learn customized feature representations for specific tasks. Besides being deeper, modern DNNs have developed efficient and effective architecture variations that scale well for high dimensional structured data. For instance, deep convolutional networks have demonstrated state-of-the-art performance in processing imagery data [93], deep networks with recurrent [69], attention [31], and memory [57] modules have brought about breakthroughs in sequential learning problems, such as natural language processing, speech and audio modeling [223]. A blending of the spatial/temporal modules have shown particular advantage in video and motion prediction [202], which have been recognized to share striking similarities to many dynamic geoscience problems [151]. Here I develop two sets of DNN architectures that input the dynamical/moisture field time sequences and output hourly-scale spatial distribution as well as spatial mean precipitation estimates. Each of them adopts and composes particular spatial/temporal modules for the treatment of the predictor's spatiotemporal structures.

1. Convolutional Neural Network

This section describes the convolutional neural network that searches, extracts, and synthesizes spatial features from the dynamical/moisture field for precipitation estimation. The model here makes explicit use of the data spatial structure but does not consider temporal connections within a precipitation event. Details about model architecture are introduced as follows.

The dynamical/moisture field time series is denoted with a $n \times c \times h \times w$ tensor. Here n is the duration hour of the considered precipitation event, $c = 6$ represents 6 variables,

$h, w = 25$ is the latitude/longitude span of the predictor field. The CNN model operates on the 3-D dynamical/moisture field at each time step. As illustrated in Figure 4.3, the CNN applies a set of convolution kernels to go through the $c \times h \times w$ input snapshot. The kernels are $c' \times c \times a \times b$ tensors composed of trainable parameters. Each convolution operation is carried out by computing the dot product between the kernel and a particular input patch, followed by shifting and element-wise non-linearity:

$$\mathbf{Y}_{p,q}^{c'} = \mathbf{f}(\mathbf{W}^{c' \times c \times a \times b} \cdot \mathbf{X}_{p,q}^{c \times a \times b} + \mathbf{b}^{c'}) \tag{4.4}$$

Here the upper index labels the variable's dimension, the lower index labels the geoposition. $\mathbf{X}_{p,q}^{c \times a \times b}$ represents a $c \times a \times b$ dimensional input patch around location (p, q) in the $h \times w$ input field. (p, q) shifts as we scan $\mathbf{X}^{c,a,b}$ with $\mathbf{W}^{c' \times c \times a \times b}$ by a pre-defined stride. The scanning is performed using element dot product between $\mathbf{X}_{p,q}^{c \times a \times b}$ and the convolution kernel $\mathbf{W}^{c' \times c \times a \times b}$. $a \times b$ is named the *receptive field* of the kernel. The result is further transformed by adding a bias vector $\mathbf{b}'$ followed by a non-linear transformation f. The final convolution result is a c' dimensional vector ($\mathbf{Y}_{p,q}^{c'}$) at location (p, q). Equation 4.4 can be interpreted as applying c' learnable filters to seek salient features from the input field while maintaining the spatial structure of the input. Preferably, the network will learn filters that activate when they see critical local dynamical patterns on the first layer. Fostered by spatial down-sampling, which is achieved by the pooling operation, the network may eventually learn a synoptic atmospheric pattern that promotes precipitation on higher layers of the network.

The pooling layers act to coarsely grain local semantically similar features into one [97]. Through down-sampling, the higher layer convolutions work on extracted local features, which enables learning higher level abstractions on the expanded receptive field [96]. I adopt maximum pooling that computes the maximum of a local patch of units in one feature map.

I extract spatial salient features by stacking multiple stages of convolution, non-linearity, and pooling layers. The extracted features are further processed by a MLP that outputs

a G dimensional vector, which represents precipitation estimates for G gauges within the considered geogrid. By performing dot-product with the pre-defined mask vector $\mathbf{M}$, I obtain precipitation estimates at gauges with available observations. I apply off-hand tools from the deep learning software library to analytically derive the derivative of the loss function with respect to the parameters of the computing graph. The parameters are then optimized based on backpropagation. Details about model implementation are introduced in the following sections.

2. Recurrent Convolutional Neural Network

I explicitly consider the temporal connection within a precipitation event by including a hidden state variable after spatial features extraction from CNN. This is achieved by stacking a recurrent network module upon the CNN. Basics about recurrent neural network (RNN) are briefly introduced below. The specific model architecture is explained thereafter.

RNNs are neural networks with hidden state variables:

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{X}_t, \boldsymbol{\theta}) \tag{4.5}$$

$\mathbf{h}_t$ denotes the hidden state variable at time t, f is the transition function that updates the hidden state $\mathbf{h}_t$ based on previous state $\mathbf{h}_{t-1}$ and t-step input $\mathbf{X}_t$. $\boldsymbol{\theta}$ denotes parameter vector. For instance, in a vanilla RNN, the state transition function takes the following parameteric form:

$$\mathbf{h}_t = \tanh\Big(\mathbf{W}_{\mathbf{h}} \cdot \mathbf{h}_{t-1} + \mathbf{W}_{\mathbf{X}} \cdot \mathbf{X}_t + \mathbf{b}\Big) \tag{4.6}$$

We can unfold Equation 4.5 along time by iteratively applying Equation 4.5 on itself:

$$\begin{aligned}\mathbf{h}_t &= f(\mathbf{h}_{t-1}, \mathbf{X}_t, \boldsymbol{\theta}) \\ &= f\Big(f(\mathbf{h}_{t-2}, \mathbf{X}_{t-1}, \boldsymbol{\theta}), \mathbf{X}_t, \boldsymbol{\theta}\Big) \\ &= f\Big(f\big(f(\mathbf{h}_{t-3}, \mathbf{X}_{t-2}, \boldsymbol{\theta}\big), \mathbf{X}_{t-1}, \boldsymbol{\theta}), \mathbf{X}_t, \boldsymbol{\theta}\Big) \\ &= f\Big(f(f(f(...))), \mathbf{X}_t, \boldsymbol{\theta}\Big)\end{aligned} \tag{4.7}$$

Some key implications of Equation 4.7 are summarized as follows:

- The transition function f and its parameter $\boldsymbol{\theta}$ are shared at each computing time step along the sequential modeling process.
- The hidden variable $\mathbf{h}_t$ is preferred to serve as a summary of task-relevant aspects of the past input sequence up to time t [56].
- We can run the unfolded computational graph forward through entire sequence to compute the loss, and run it backward through entire sequence to compute gradient. This indicates that RNN can potentially be trained using backpropagation through time.

For the last aspect above, gradient-based training of RNN has been recognized difficult for basic recurrent architectures, such as for Equation 4.6. The difficulty originates from the fact that, the partial derivatives of the loss function L with respect to the parameters tend to vanish or blow up as error signals flow backwards in time [67, 12, 68]. Although it is possible to clip the gradients as they blow up [144], a vanishing gradient prevents effective learning of long-term dependencies. Here I adopt the Long Short-Term Memory (LSTM) recurrent network [69] to mitigate the vanishing gradient problem. A LSTM architecture maintains two hidden states in its computational graph, one is the conventional hidden state vector $\mathbf{h}_t$, the other is a memory vector $\mathbf{c}_t$. LSTM adopts four interacting modules to read from, write

to, or reset the memory vector $\mathbf{c}_t$:

$$\begin{pmatrix} \mathbf{i} \\ \mathbf{f} \\ \mathbf{o} \\ \mathbf{g} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \mathbf{W} \begin{pmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{pmatrix} \tag{4.8}$$

Here σ/tanh is element-wise sigmoid/hyperbolic-tangent function that squashes each element of a vector to $(0, 1)/(-1, 1)$. Assume the input vector $\mathbf{x}$ of dimension p, and hidden state vector $\mathbf{h}$ as well as memory vector $\mathbf{c}$ of dimension q, $\mathbf{W}$ is thus a $(4q, p + q)$-dimensional transition matrix, $\mathbf{i}$, $\mathbf{f}$, $\mathbf{o}$, and $\mathbf{g}$ are q-dimensional vectors. With these modules defined, the memory vector $\mathbf{c}_t$ and $\mathbf{h}_t$ are updated based on the following equations:

$$\mathbf{c}_t = \mathbf{f} \odot \mathbf{c}_{t-1} + \mathbf{i} \odot \mathbf{g} \tag{4.9}$$

$$\mathbf{h}_t = \mathbf{o} \odot \tanh(\mathbf{c}_t) \tag{4.10}$$

Equation 4.9 and 4.10 tells how $\mathbf{i}/\mathbf{g}$, $\mathbf{f}$, and $\mathbf{o}$ work as binary gates that control whether memory cell $\mathbf{c}_t$ is updated, whether it is reset to zero, and whether its local state is revealed in the hidden vector, respectively [89]. The memory cell $\mathbf{c}_t$ is updated based on two mechanisms that cooperates in an additive manner. The first mechanism tells how $\mathbf{c}_t$ maintains memory from past state, which is described by $\mathbf{f} \odot \mathbf{c}_{t-1}$; the second mechanism tells how to update $\mathbf{c}_t$ based on input $\mathbf{x}_t$ and past hidden state $\mathbf{h}_{t-1}$, which is described by $\mathbf{i} \odot \mathbf{g}$. The additive interaction of the two mechanisms allows error gradient on $\mathbf{c}_t$ to be distributed through time without suffering from vanishing/blowing up gradient, thus enabling learning long-term dependencies [89].

I use LSTM to relate the extracted spatial feature sequence to the corresponding time's

precipitation distribution. Specifically, I start from mapping a CNN to the time series of predictors. The extracted spatial feature time series are then used as input for the LSTM RNN for precipitation estimation. The specific model architecture is illusrated in Figure 4.3.

3. Implementation Details

Data Preparation: The data are normalized before being applied for modeling. Each class of the predictor is normalized by subtracting its mean value (μ) and dividing by its standard deviation (σ). Here μ and σ are scalars that are calculated based on the flattened predictor field for all precipitation event cases. After normalization, I divide the data into non-overlapping training/validation/test sets. Data for 2014–2018 are used as test set for model assessment. The rest of the precipitation events data are shuffled, 80%/20% used for model training/validation. The training set is applied to directly optimize the model, while the validation set is frequently evaluated along the training process for hyper-parameter tuning and preventing from overfitting.

Model Architecture: I rely on empirical experiments to determine network architecture and hyperparameters here. For implementation convenience, I only consider equal convolution kernels (same channel size and receptive field) for CNN. I try different network architecture, learning rate, batch-size and training iterations to decide the optimal setting. I admit that the result could be significantly different by adopting alternative architectures. The specific network architecture and hyper-parameters options are listed in Table 4.2.

Training I use stochastic gradient descent (SGD) for model training [20]. SGD uses a stochastic approximation of whole batch gradient in backpropogation to alleviate the high computing cost in evaluating the derivatives for the global loss function. For training convolutional recurrent net, I calculate stochastic gradient of loss function based on chunks of the sequence instead of the whole sequence to alleviate computation burden. I adopt the early stopping strategy to regularize the model: The training process is terminated when further training improves performance only for the training set but not for the validation set.

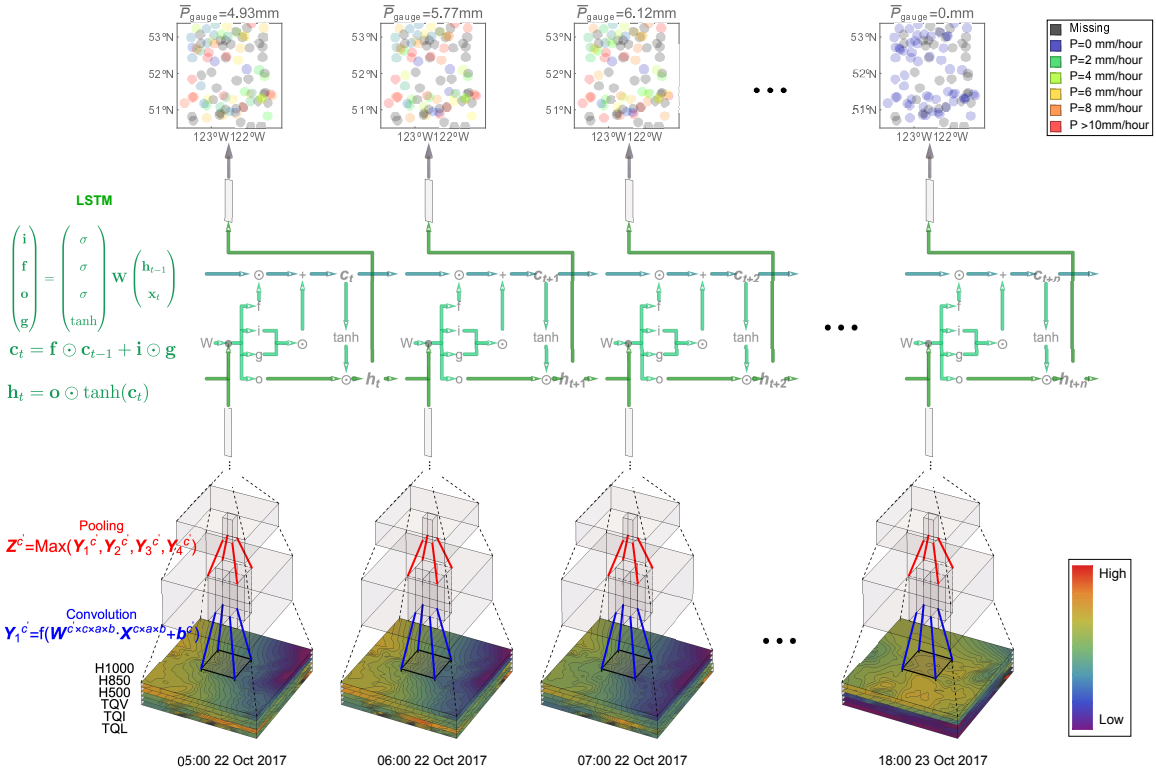


Figure 4.3: Illustration of the convolutional recurrent neural network model. The bottom colored stacked frames show the predictors, which are composed of every-hour geopotential height (GPH) field at 500, 850, and 1,000 hPa, as well as the total column liquid water, ice water and water vapor field. The specific region I consider here is the Grid 11 in Figure 4.1, which covers 40°N-52°N,115°W-130°W. Data are normalized by subtracting mean and divided by standard variance. Orange/blue indicates high/low values, as shown in the bottom-right legend. The blue lines represent a convolution operation applied on the dynamical/moisture field. The red lines represent the pooling operation that down-samples the local features. Several stages of convolution and pooling layers are stacked for extracting salient spatial features. The extracted feature time series are combined with the hidden state variable through a LSTM RNN for precipitation estimation. Information flow though the memory and hidden state cells of LSTM is labeled with green arrows. The observed precipitation distribution for the target geogrid is shown on the precipitation map at the top of the figure.

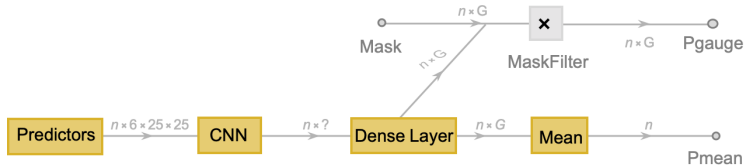
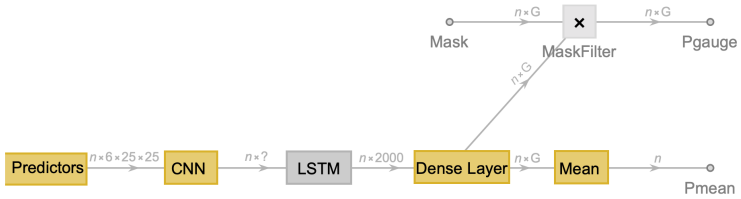
Architecture/Hyperparameter	Options
CNN Architecture	 $\text{Conv}_{24\times3\times3} \times 3 \rightarrow \text{Pool}_{2\times1} \rightarrow \text{Conv}_{24\times3\times3} \times 3 \rightarrow \text{BN} \rightarrow \text{FC}_{2000} \rightarrow \text{Drop} \rightarrow \text{BN} \rightarrow \text{FC}_{2000} \rightarrow \text{BN} \rightarrow \text{FC}_G$ $\text{Conv}_{48\times3\times3} \times 3 \rightarrow \text{Pool}_{2\times1} \rightarrow \text{Conv}_{48\times3\times3} \times 3 \rightarrow \text{BN} \rightarrow \text{FC}_{2000} \rightarrow \text{Drop} \rightarrow \text{BN} \rightarrow \text{FC}_{2000} \rightarrow \text{BN} \rightarrow \text{FC}_G$ $\text{Conv}_{64\times3\times3} \times 3 \rightarrow \text{Pool}_{2\times1} \rightarrow \text{Conv}_{64\times3\times3} \times 3 \rightarrow \text{BN} \rightarrow \text{FC}_{2000} \rightarrow \text{Drop} \rightarrow \text{BN} \rightarrow \text{FC}_{2000} \rightarrow \text{BN} \rightarrow \text{FC}_G$
RCNN Architecture	 $\text{Conv}_{24\times3\times3} \times 3 \rightarrow \text{Pool}_{2\times1} \rightarrow \text{Conv}_{24\times3\times3} \times 3 \rightarrow \text{BN} \rightarrow \text{LSTM}_{2000} \rightarrow \text{Drop} \rightarrow \text{BN} \rightarrow \text{FC}_{1000} \rightarrow \text{BN} \rightarrow \text{FC}_G$ $\text{Conv}_{48\times3\times3} \times 3 \rightarrow \text{Pool}_{2\times1} \rightarrow \text{Conv}_{48\times3\times3} \times 3 \rightarrow \text{BN} \rightarrow \text{LSTM}_{2000} \rightarrow \text{Drop} \rightarrow \text{BN} \rightarrow \text{FC}_{1000} \rightarrow \text{BN} \rightarrow \text{FC}_G$ $\text{Conv}_{64\times3\times3} \times 3 \rightarrow \text{Pool}_{2\times1} \rightarrow \text{Conv}_{64\times3\times3} \times 3 \rightarrow \text{BN} \rightarrow \text{LSTM}_{2000} \rightarrow \text{Drop} \rightarrow \text{BN} \rightarrow \text{FC}_{1000} \rightarrow \text{BN} \rightarrow \text{FC}_G$
Learning Rate	0.01/0.001/0.0001/0.00001
Mini Batch Size	32 for CNN and 4 for RCNN
Dropout Probability	0.2/0.5

Table 4.2: Model architectures and hyperparameters considered in the experiment. For the model architecture, $\text{Conv}_{c\times a\times b}$ represents convolutional layer with channel size of c and receptive field of $a \times b$, followed by batchnormalization and ReLU activation function :$\text{ReLU}(x) = max(0, x)$. $\text{Pool}_{2\times1}$ is maximum pooling layer with receptive field of $2times2$ and stride of 1. BN is batchnormalization, FC_n is fully connected layer with neuron size of n, followed by ReLU. Drop is dropout layer.

4.3.4 Evaluation Metrics

Two deterministic skill metrics, namely correlation coefficient score (r), root mean square error (RMSE) are used to measure model's performance at various measurement scopes. The formula of r and RMSE are written as follows:

$$r = \frac{E[(\hat{P} - \overline{\hat{P}})(P - \overline{P})]}{\sigma_{\hat{P}}\sigma_P} \tag{4.11}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{P}_i - P_i)^2} \tag{4.12}$$

$P/\hat{P}$ is precipitation estimation/observation at considered measurement scope, E denotes expectation, σ denotes standard variance.

4.4 Results

This section shows the results for hourly QPF using the proposed deep neural networks. Models' performances for the test set (2014–2018) are evaluated against gauge observations, and compared with the skill of MERRA2 precipitation products (P_{MERRA2} and $\text{P}_{\text{MERRA2}_C}$). The evaluation is carried out at different spatial scales in correspondence to the resolutions of different reference data. I consider $2^\circ \times 2.5^\circ$ grid scale (following P_{CPC}) and gauge-point scale for the evaluation here. The specific results are presented as follows.

4.4.1 Evaluation at $2^\circ \times 2.5^\circ$ Spatial Scale

Column 1 and Column 2 of Figure 4.4 and 4.5 show examples of precipitation process simulations at $2^\circ \times 2.5^\circ$ spatial scale. The reference observation data are obtained by averaging

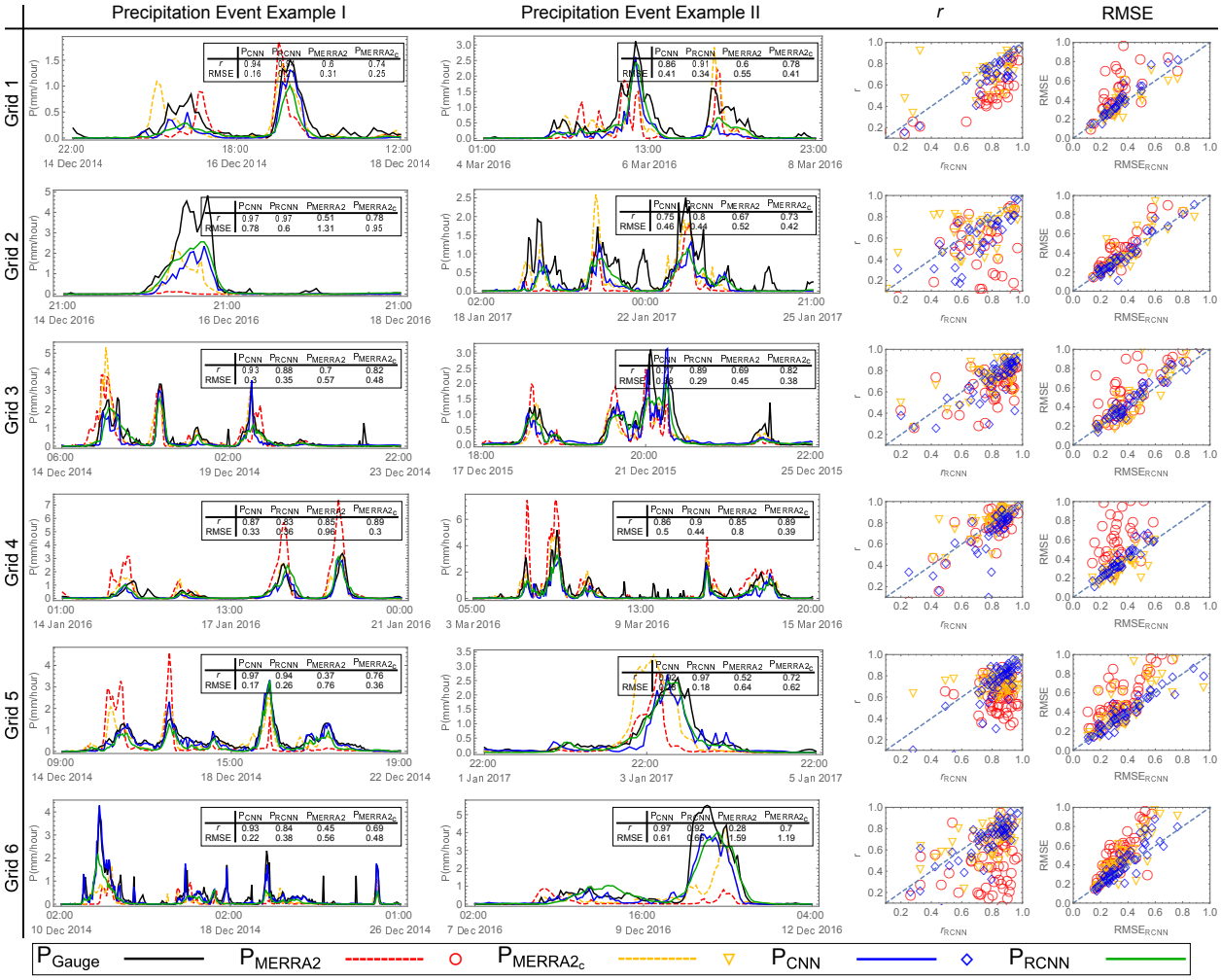


Figure 4.4: Column 1 and Column 2 show examples of precipitation process simulations for Grid 1 to Grid 6. Column 3 and Column 4 compares the r and RMSE score of P_{MERRA2}/P_{MERRA2_C}/ P_{CNN} against P_{RCNN}.

gauge observations within the considered grid-box. The configuration of the CNN and RCNN models are determined based on their performances for the validation set. The MERRA2 precipitation products are used as baseline. A summary of r and RMSE score evaluated for all the test set precipitation events are shown in Column 3 and Column 4, respectively. Since RCNN generally shows optimal performance among the considered models, I use P_{RCNN} to benchmark the rest models: For scatter plots in Column 3, points below $1:1$ line indicates lower r score compared to P_{RCNN}; similarly, for scatter plots in Column 4, points above $1:1$ line indicates higher RMSE compared to P_{RCNN}.

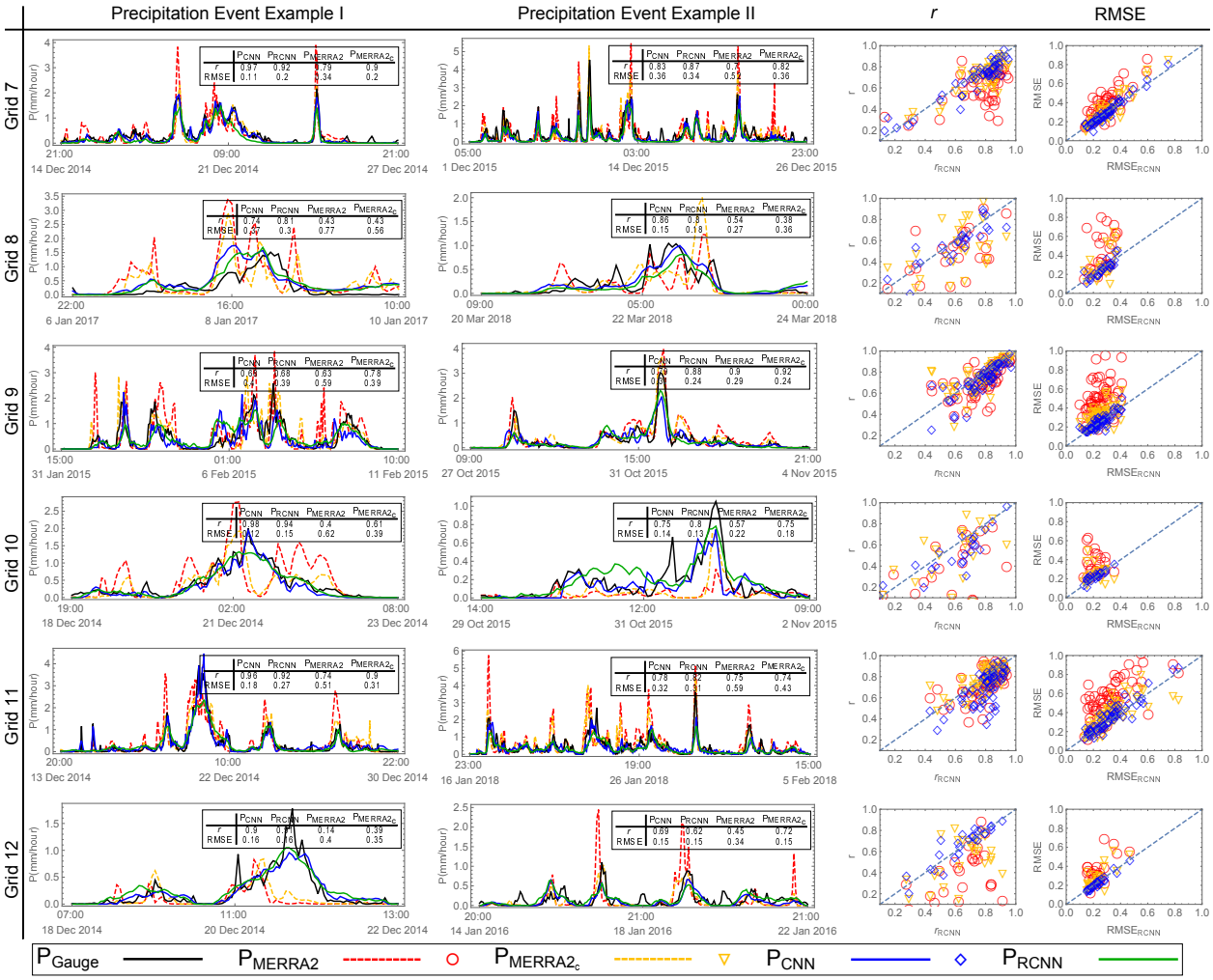


Figure 4.5: Similar as Figure 4.4 but for Grid 7 to Grid 12.

Deficiency Type	Description	Examples
Overestimation	Precipitation>>Observation	Grid 3, 10:00 Dec 14, 2014 Grid 4, 24:00 Jan 17, 2016 Grid 11, 01:00 Jan 17, 2018
Underestimation	Precipitation<<Observation	Grid 1, 12:00 Mar 6, 2016 Grid 2, 21:00 Dec 16, 2016 Grid 6, 02:00 Dec 10, 2014
Precocious warning	Predicting too early arrival of precipitation peaks.	Grid 1, 24:00 Dec 16, 2014 Grid 5, 23:00 Jan 3, 2017 Grid 10, 02:00 Dec 21, 2014
Delayed warning	Predicting too late arrival of precipitation peaks.	Grid 6, 24:00 Dec 10, 2014 Grid 6, 20:00 Dec 10, 2016 Grid 10, 08:00 Nov 1, 2015

Table 4.3: Classification of deficiencies of MERRA2 precipitation products.

A detailed inspection of the precipitation process simulation examples shows that the DNN models can fetch up various sources of deficiencies of P_{MERRA2}, such as over/under estimation, peaking shift. I summarize the analyses in Table 4.3.

Table 4.4 gives a quantitative recap of skills for the considered models. While Figure 4.4 and Figure 4.5 show evaluation results at event-wise scale, Table 4.4 calculate same skill scores based on all the simulations from different events. Some of the key messages from Table 4.4 are listed as follows:

- **P_{CNN}/P_{RCNN} V.S. P_{MERRA2}:** For all the 12 $2^\circ \times 2.5^\circ$ geogrids, P_{CNN} and P_{RCNN} show better performance compared to P_{MERRA2}. The average r and RMSE skills are significantly improved.

- **P_{CNN}/P_{RCNN} V.S. P_{MERRA2_C}:** P_{CNN} shows better performance compared to P_{MERRA2_C} for 8 out of the 12 geogrids as measured by r and RMSE; P_{RCNN} shows better performance compared to P_{MERRA2_C} for 11 out of the 12 geogrids.

- **P_{CNN} V.S. P_{RCNN}:** For all the 12 grid boxes, P_{RCNN} shows better performance than P_{CNN}, with an average r/RMSE skill improvement of 0.036/0.025.

Overall, the results suggest a better performance of P_{CNN} and P_{RCNN} compared to

	r				RMSE			
	P_{MERRA2}	P_{MERRA2_C}	P_{CNN}	P_{RCNN}	P_{MERRA2}	P_{MERRA2_C}	P_{CNN}	P_{RCNN}
Grid 1	0.495	0.736	0.722	**0.766**	0.609	0.404	0.419	**0.395**
Grid 2	0.318	0.605	0.648	**0.725**	0.572	0.446	0.446	**0.426**
Grid 3	0.658	0.764	0.763	**0.824**	0.779	0.562	0.579	**0.518**
Grid 4	0.780	**0.834**	0.781	0.816	0.834	**0.401**	0.443	0.405
Grid 5	0.521	0.737	0.819	**0.848**	0.712	0.572	0.476	**0.459**
Grid 6	0.341	0.695	0.822	**0.866**	0.722	0.546	0.459	**0.398**
Grid 7	0.617	0.745	0.791	**0.833**	0.464	0.364	0.347	**0.332**
Grid 8	0.425	0.501	0.504	**0.513**	0.507	0.333	0.280	**0.277**
Grid 9	0.754	0.802	0.809	**0.818**	0.575	0.369	0.285	**0.278**
Grid 10	0.447	0.539	0.548	**0.594**	0.378	0.269	0.244	**0.227**
Grid 11	0.770	0.820	0.796	**0.828**	0.635	0.447	0.397	**0.370**
Grid 12	0.373	0.510	0.672	**0.679**	0.379	0.304	0.249	**0.237**
Average	0.542	0.691	0.723	**0.759**	0.597	0.418	0.385	**0.360**

Table 4.4: Comparing precipitation estimation performance based on r and RMSE score. The precipitation estimates at hourly, $2^\circ \times 2.5^\circ$ from original MERRA2 precipitation product (P_{MERRA2}), MERRA2 bias corrected precipitation product (P_{MERRA2_C}), CNN estimation (P_{CNN}), and RCNN estimation (P_{RCNN}) are compared against gauge average observations for period from 2015 to 2018. The average skill score are shown in the bottom row. The best performance for each comparison group are labeled with bold typeface and underline.

P_{MERRA2} and P_{MERRA2_C} at a $2^\circ \times 2.5^\circ$ spatial scale, with r improved from 0.55 to 0.75 on average. Various types of parameterization-related precipitation estimation errors are highlighted by comparing the MERRA2 and DNN precipitation process estimations.

4.4.2 Evaluation at Gauge-point Scale

Figure 4.6 to Figure 4.8 shows the gauge-point, hourly QPF evaluation results using r and RMSE score. The evaluation are based on all the solid observations from all the test set's precipitation events. Some of the key findings are listed as follows:

- The average r skill score measured at gauge-point, hourly scale are of the magnitude order of 0.3–0.4. For the best-simulated gauges, the r skill scores can reach the magnitude order of 0.6–0.7.

- Although RCNN outperforms CNN in estimating grid average precipitation, many grids show better performance of CNN in estimating precipitation at gauge-point scale. This

indicates that the recurrent module in RCNN might make over consideration of the temporal interconnections of a precipitation process.

- Gauges that are close to each other may show distinct skill scores.

4.5 Dynamical Forecast Experiment

In order to test the model's robustness and its applicability in real world precipitation forecast, retrospective dynamical forecast experiments are carried out for two typical atmospheric river land-falling events (00:00 UTC 13 October 2016 – 23:00 UTC 17 October 2016 and 00:00 UTC 19 October 2017 – 23:00 UTC 23 October 2017). Both events are selected from the test set in order to make objective evaluations.

The experiment employs a dynamical downscaling model (i.e., WRF-ARW Version 4) that is forced by 1) historical operational forecasts from Global Forecast System (GFS), and 2) corresponding periods' atmosphere reanalysis from GFS. Then, the downscaled atmospheric dynamics and moisture fields are used as input for the neural network model to make precipitation estimations. The models and data for this experiment are described as follows.

4.5.1 Numerical Models

Global Forecast System

The Global Forecast System (GFS) is a global weather forecast model that serves as the cornerstone for operational forecast in the National Centers for Environmental Prediction (NCEP). The system includes four separate models that simulate the earth atmosphere, ocean, land/soil, and sea ice, respectively. Model's initial state estimation is produced by the Global Data Assimilation System (GDAS) that merges satellite and conventional meteorological observations from various sources. GFS is run routinely four times per day at

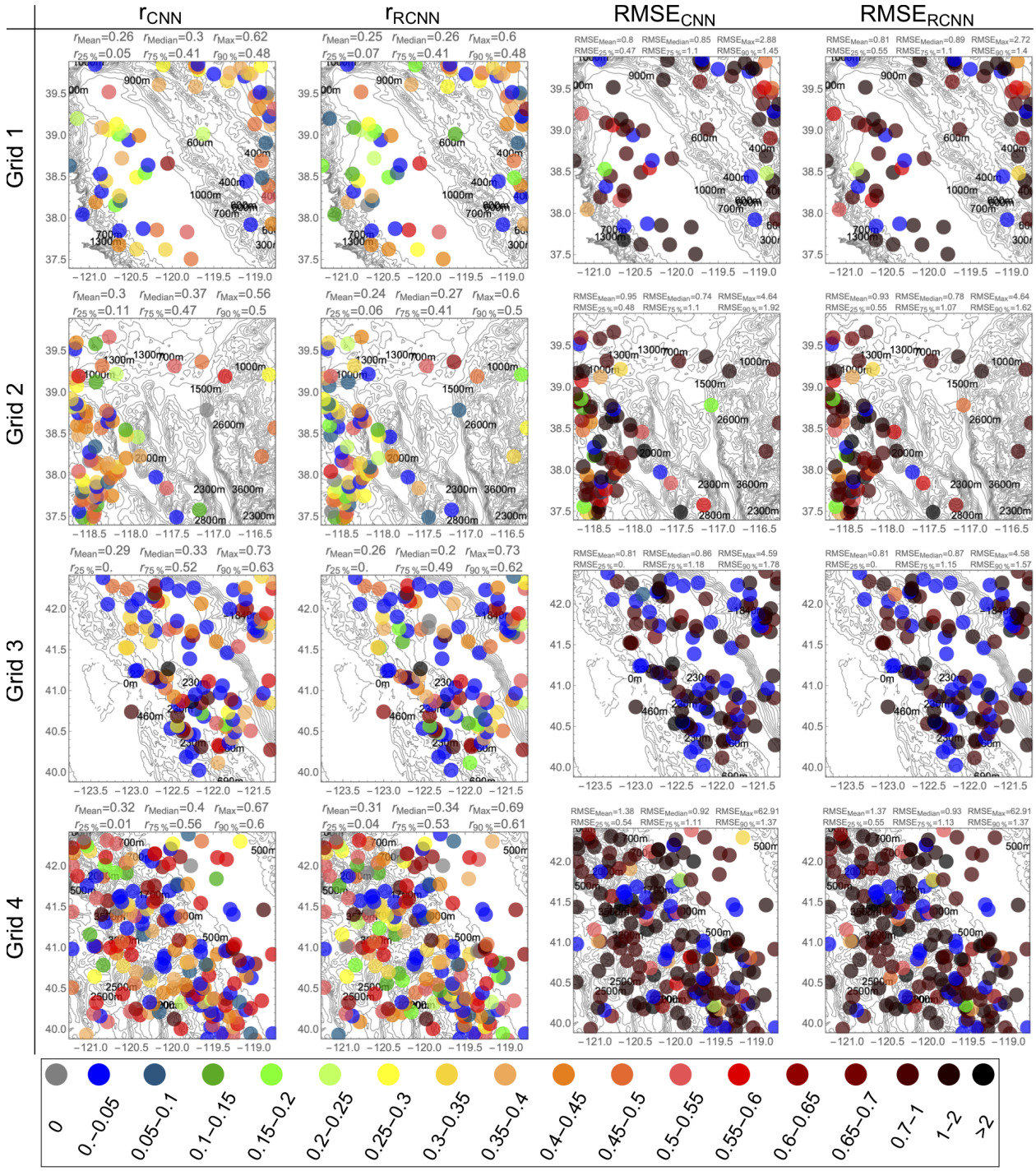


Figure 4.6: r and RMSE skill score evaluated at gauge-point and hourly scale for Grid 1-4. The contour lines show the elevation data. The skill scores are labeled with colored disks.

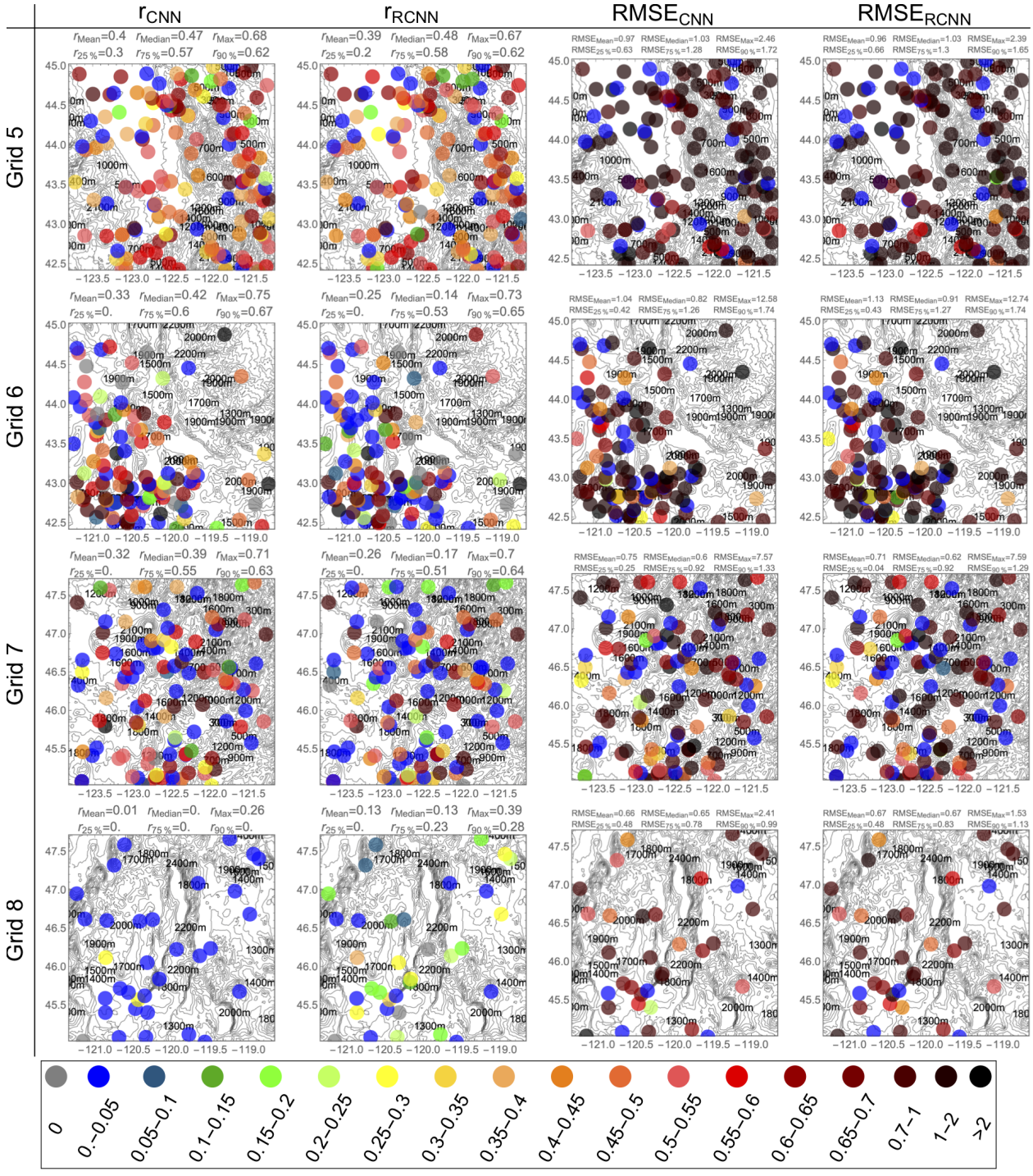


Figure 4.7: Similar as Figure 4.4 but for Grid 5 to Grid 8.

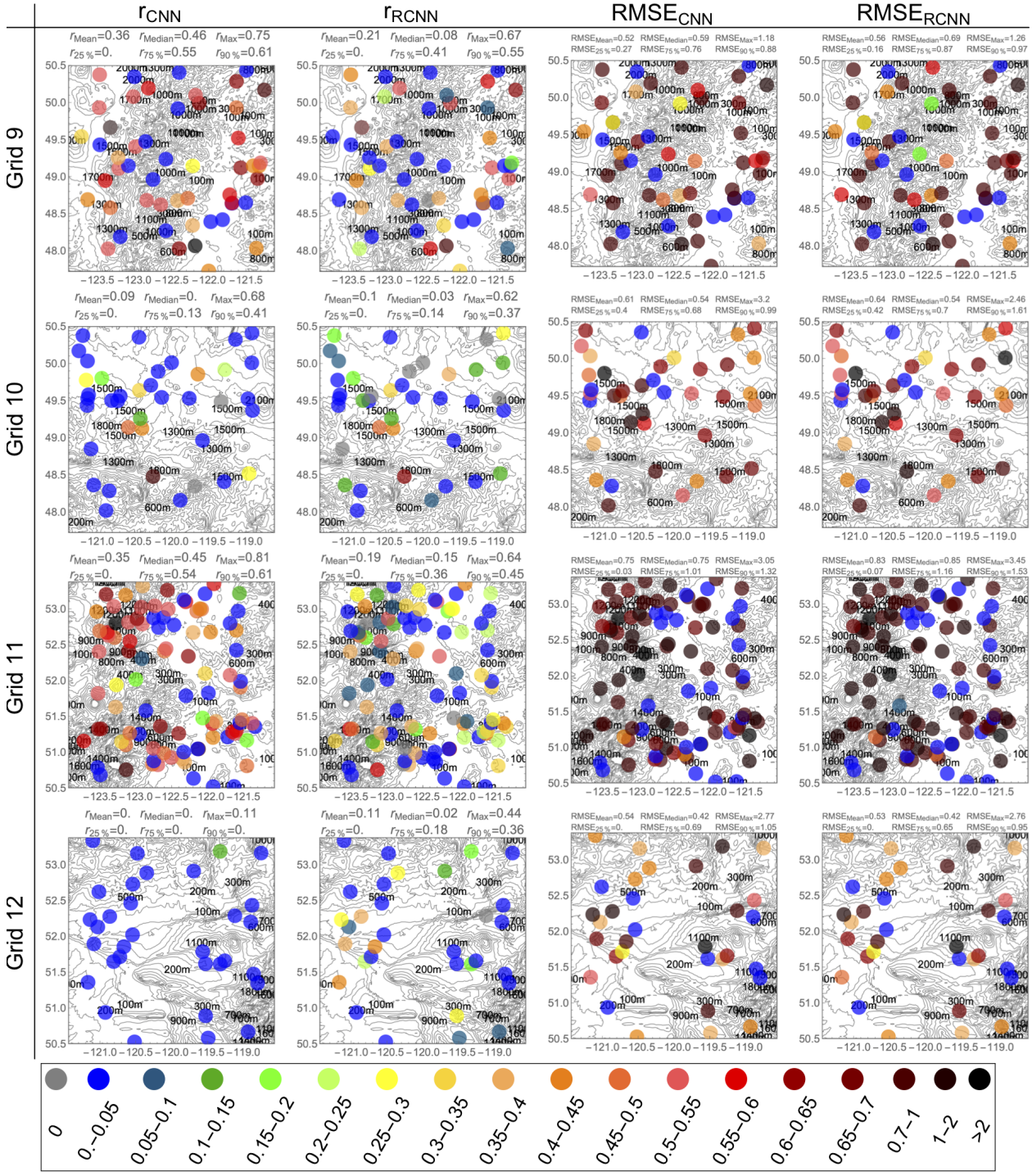


Figure 4.8: Similar as Figure 4.6 and Figure 4.7 but for Grid 9 to Grid 12.

00:00, 06:00, 12:00 and 18:00 UTC out to 192 hours. The base horizontal resolution is 28 km. Details about model dynamics, physical parameterization, analysis, and configuration can be found in [156, 157]. Here the forecast and reanalysis data from GFS for the two selected events are downloaded from the following website:

https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs

The data are of $1^\circ \times 1^\circ$ horizontal resolution and 3 hour temporal resolution.

Weather Research and Forecasting Model

In order to 1) obtain the input variables at required resolution for the neural network model, and 2) fill in the spatiotemporal missing values in a physically sound manner, dynamical downscaling is carried out with initial and boundary conditions constrained by the GFS forecast/analysis. Here, I use the Advanced Research core of the Weather Research and Forecasting Model Version 4 (WRF-ARW V4) for dynamical downscaling.

WRF is a nonhydrostatic, mesoscale numerical weather prediction model. The ARW dynamical solver performs numerical spatiotemporal integration of the atmospheric primitive equations on terrain-following coordinates [174]. The unresolved processes are empirically estimated as functions of the resolved variables based on multiple choices of parameterization schemes. WRF-ARW has been intensively investigated for quantifying forecast errors and exploring predictability limit for the West Coast [27, 117]. The existing works offer rich legacy for guiding model configuration and parameterization selection. Following previous settings [23, 27, 117], the WRF model is set as follows: two one-way nested domains are applied (domains delineated in Figure 4.9), the outer domain comprises 100×100 30-km grid points centered at (45°N,120°W), the inner domain comprises 180×180 10-km grid points centered at (42°N,121°W). The vertical grids contain 40 σ levels with the top pressure level at 50 hPa. The integral time steps for the two domains are 108s and 36s. The subgrid-scale parameterization options for each six of the major physical processes are listed in Table 4.5.

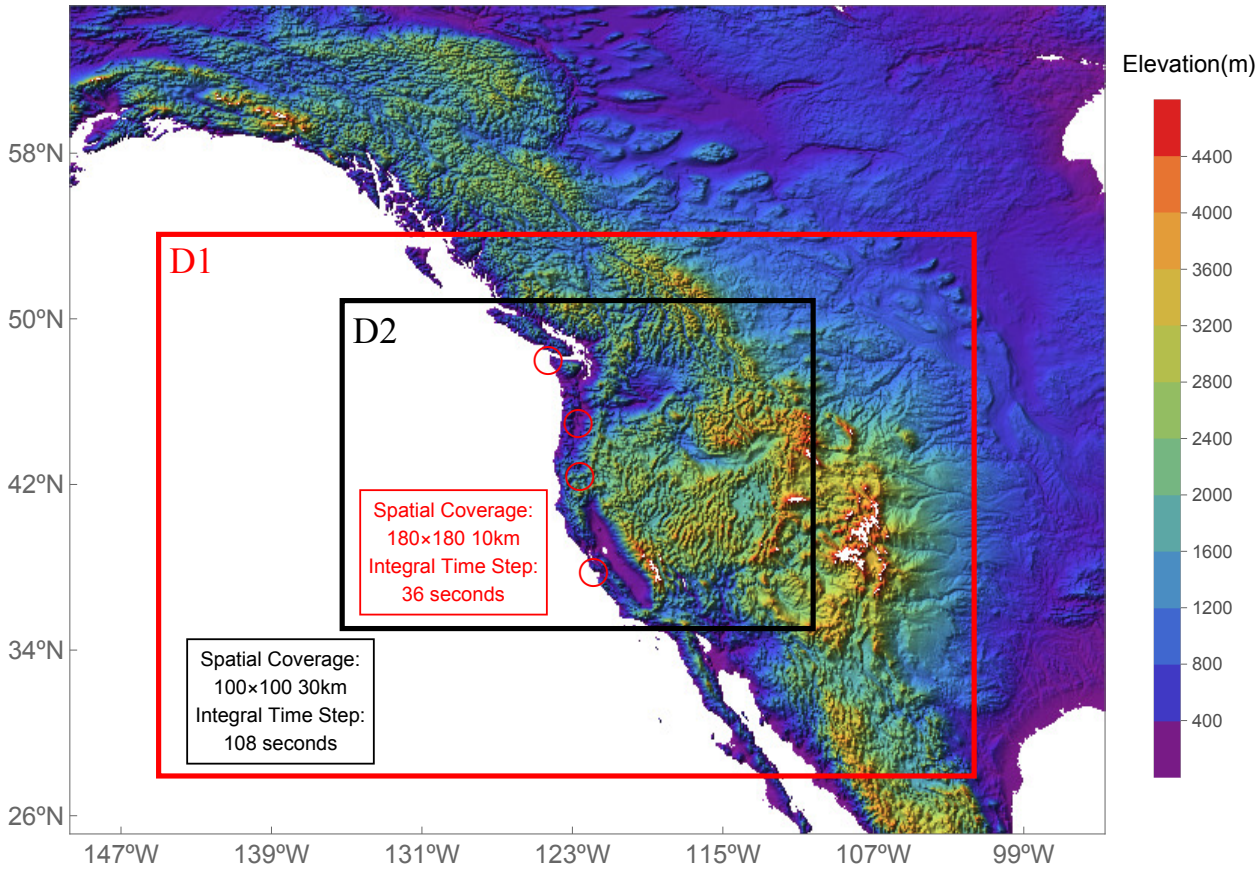


Figure 4.9: Two nested domains in the WRF configuration. The spatial coverage and integral time step for each domain is labeled. The red circles along the coast denote the positions of atmospheric sounding observations.

Physical Process	Parameterization Option
Cumulus Cloud	Kain-Fritsch Scheme [86] (only for domain 1)
Microphysics	Purdue Lin Scheme [30]
Short-wave Radiation	Dudhia Shortwave Scheme [41]
Long-wave Radiation	RRTM Longwave Scheme [123]
Planetary Boundary Layer	Yonsei University Scheme (YSU) [71]
Surface	Unified Noah Land Surface Model [185]

Table 4.5: Parameterization options for WRF-ARW dynamical downscaling

I use the Purdue Lin Scheme [30] to represent the cloud and precipitation processes in the inner domain. The scheme parameterize the following forms of hydrometeors at a grid point: vapor, cloud water, cloud ice, rain, snow, and hail. The total column ice water, liquid water and water vapor are computed by integrating hydrometeors of different phases over the pressure coordinate for the whole column.

Since the coordinate for WRF-ARW is terrain following, GPH at pressure levels above terrain surface pressure are excluded. Besides, GPH at required pressure levels below terrain surface pressure are not necessarily available in model's output. I apply the barometric formula to extrapolate/interpolate the GPH to obtain the neural network required pressure heights:

$$P = P_b \cdot [\frac{T_b}{T_b + L_b(h - h_b)}]^{\frac{g_0 \cdot M}{R^* \cdot L_b}} \tag{4.13}$$

Here P_b is static pressure measured in Pa, T_b is standard temperature measured in K, $L_b \approx -6.5 \times 10^{-3} K/m$ is standard temperature lapse rate for troposphere, h is height above sea level (m), h_b is base height, $g_0 \approx 9.81 m/s^2$ is gravitational acceleration, $M \approx 2.9 \times 10^{-2} kg/mol$ is the molar mass of the Earth's air. $R^* \approx 8.31 J/mol/K$ is universal gas constant. Equation 4.13 can be readily derived by combining ideal gas law, atmospheric hydrostatic equilibrium equation, and constant atmospheric lapse rate.

4.5.2 Atmospheric River Land-falling Events

Dynamical Field Verification

The integrated water vapor (IWV) is applied as a primary indicator for the verification of the dynamical forecast. IWV represents the total atmospheric column's moisture content, which has been widely used for describing AR characteristics [133, 37]. IWV is calculated

by integrating the mixing ratio along all pressure levels:

$$\text{IWV} = \int_{P_b}^{P_t} \frac{q}{g} dP \tag{4.14}$$

P_b/P_t is the base/top pressure of the considered column, q is the mixing ratio of water vapor, measured in kg/kg. g is gravitational constant $9.8m/s^2$. IWV is thus measured in kg/m^2.

Figure 4.10 shows the IWV and wind field for the two dynamical forecast experiments. Compared to reanalysis-forced simulation, forecasts in both cases show relatively satisfactory performance for lead time from Day-0 up to Day-3 (Column 1 to Column 4). In Case 1 (00:00 UTC 13 October 2016 – 23:00 UTC 17 October 2016), the initial state estimate shows over-estimation of water vapor concentration off the coastal region. The forecast model develops too quick a cyclone center off British Columbia compared to reanalysis simulation. Case 2 (00:00 UTC 19 October 2017 – 23:00 UTC 23 October 2017) has more realistic initial estimates, showing an informative forecast up to Day 4.

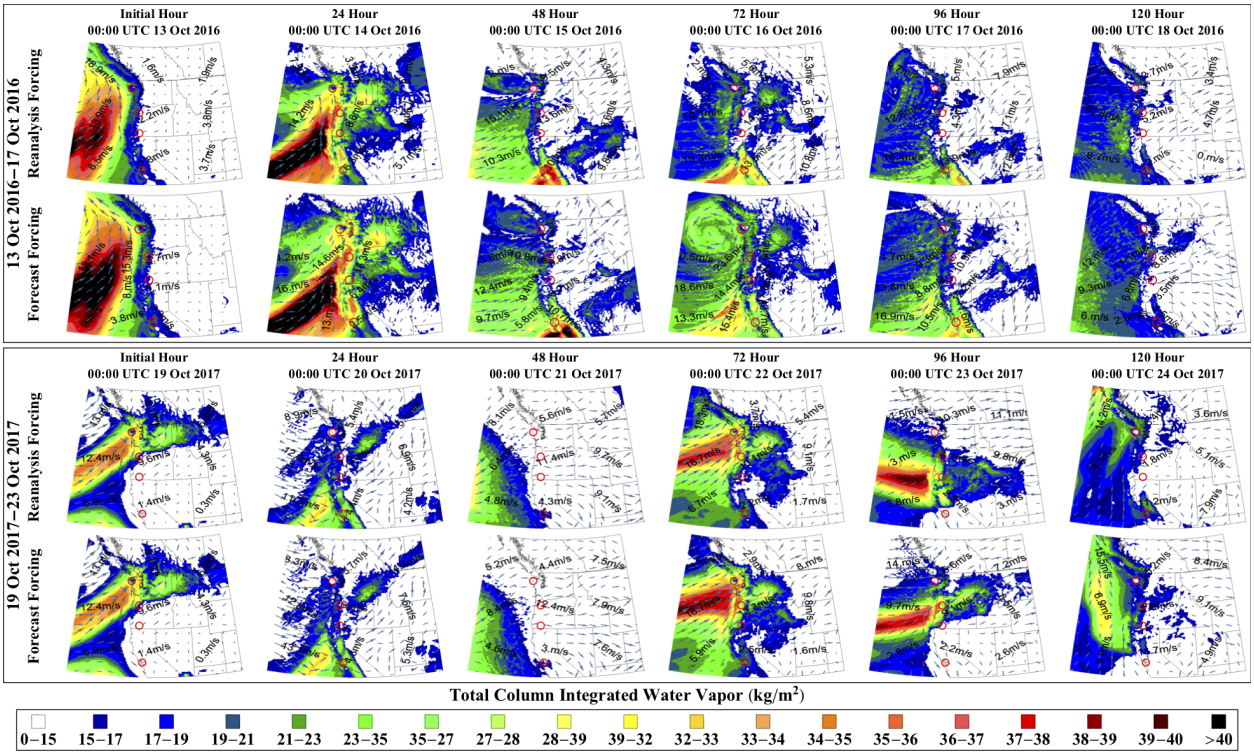


Figure 4.10: Integrated water vapor (IWV) and wind field forecast for Domain 2. Row 1–2 shows the case of 00:00 UTC 13 October 2016 – 23:00 UTC 17 October 2016; Row 3–4 shows the case of 00:00 UTC 19 October 2017 – 23:00 UTC 23 October 2017. Row 1 and Row 3 show results forced by GFS reanalysis; Row 2 and Row 4 show results forced by operational GFS forecasts that start at the beginning of the event. Column 1–6 show the dynamical analysis/forecast at forecast lead Day 0 to Day 5. The red circle along the coast denotes the positions of the soundings that measure the vertical profile of the atmosphere. The sounding data are applied for quantitative dynamical forecast verification.

Figure 4.11 and Figure 4.12 shows the geopotential height and moisture predictions at the selected sounding locations. Estimations from WRF simulations forced by GFS forecast ($\text{WRF}_{\text{Forecast}}$), WRF simulations forced by GFS reanalysis ($\text{WRF}_{\text{Analysis}}$), and MERRA2 are compared with sounding observations. Results here suggest satisfying matching between the three products and observations. However, it is noteworthy that there are disagreements between WRF simulation and MERRA2 for upper level (i.e, 500 hPa) GPH estimates.

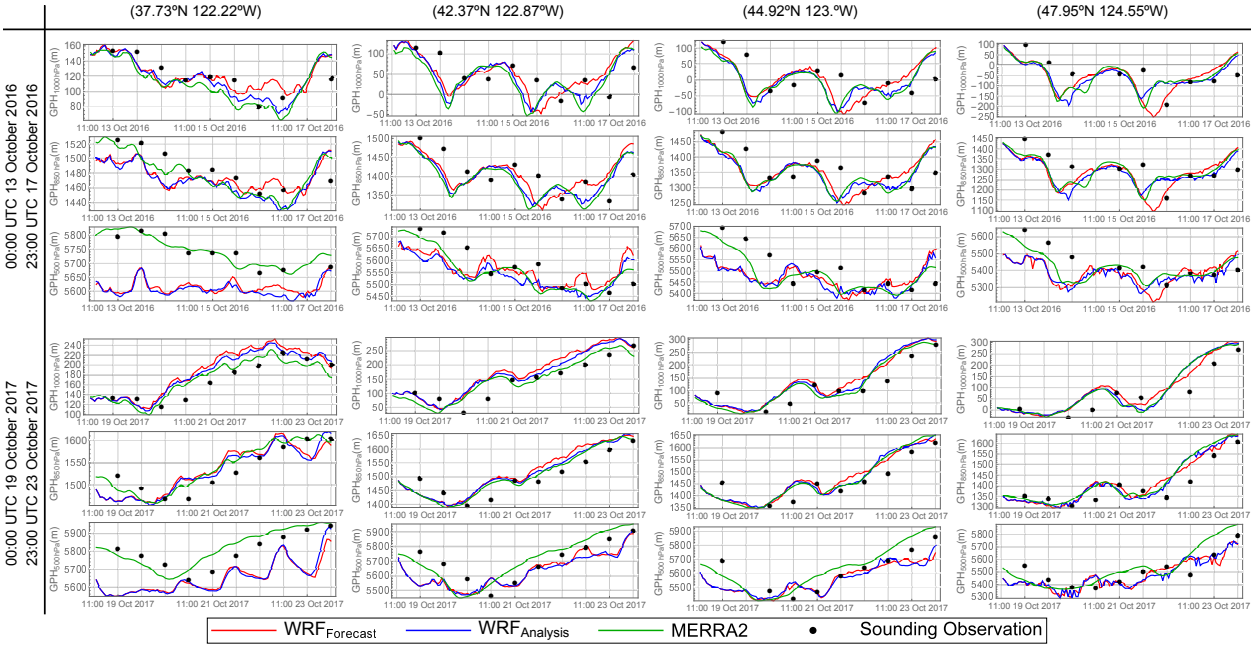


Figure 4.11: Comparing predictions of $GPH_{1000hPa}$, GPH_{850hPa}, and GPH_{500hPa} at sounding locations. The red line shows estimations from WRF simulations forced by GFS forecast ($WRF_{Forecast}$), blue line shows estimations from WRF simulations forced by GFS reanalysis ($WRF_{Analysis}$), green line shows MERRA2 estimation. Sounding observations at 00:00, 12:00 for each day are labeled with black points.

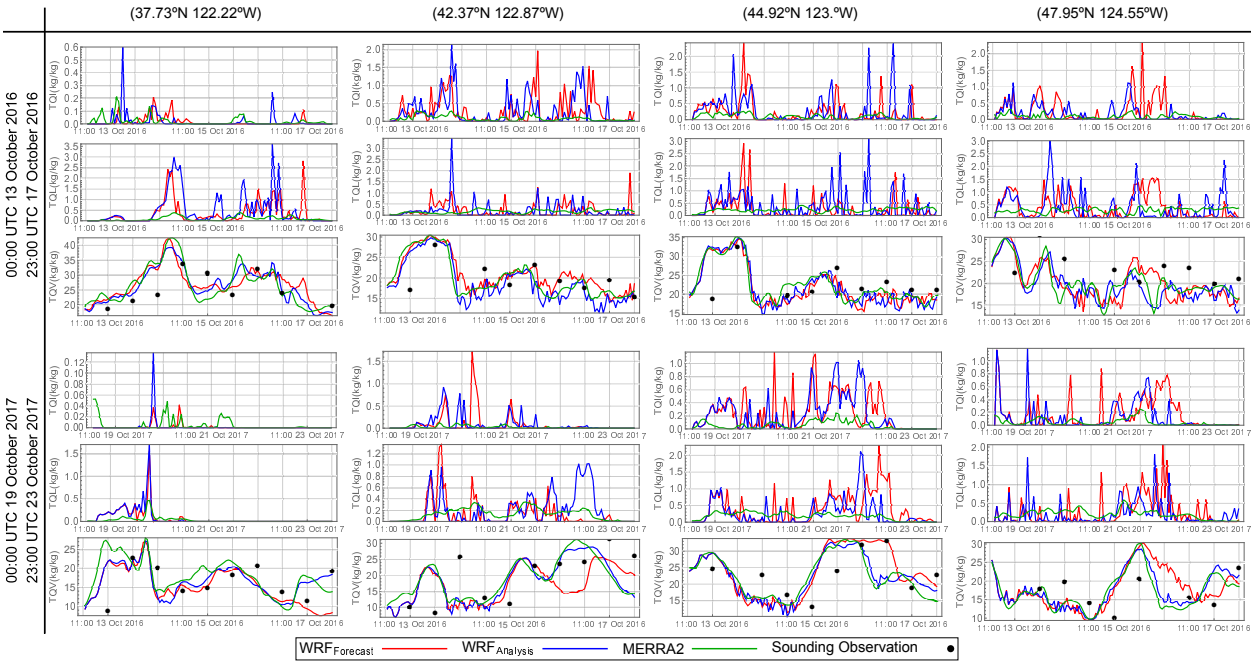
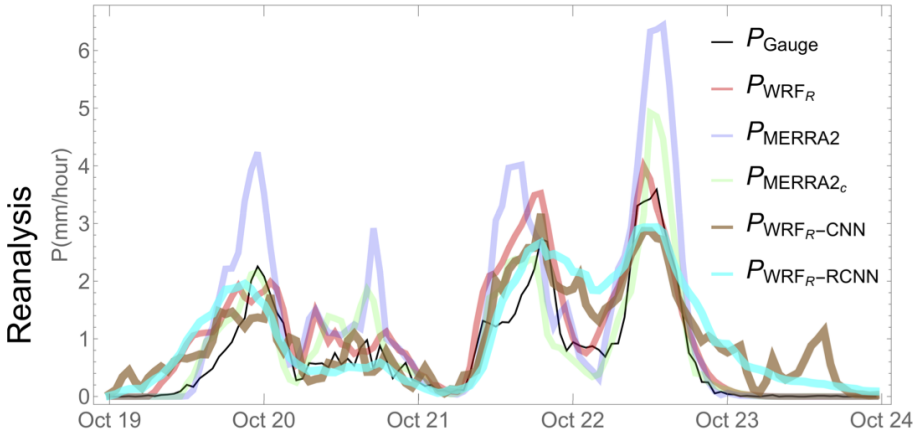


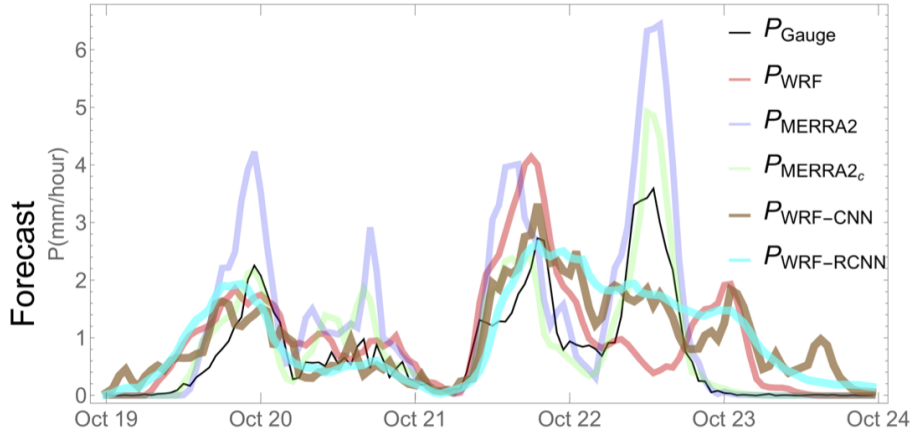
Figure 4.12: Comparing predictions of TQI, TQL, and TQV at sounding locations. The red line shows estimations from WRF simulations forced by GFS forecast ($\text{WRF}_{\text{Forecast}}$), blue line shows estimations from WRF simulations forced by GFS reanalysis ($\text{WRF}_{\text{Analysis}}$), green line shows MERRA2 estimation.

Precipitation Verification

Figure 4.13 shows the precipitation time series for 00:00 UTC 19 October 2017 – 23:00 UTC 23 October 2017 at Grid 8 in the selected domain. The case is selected since the dynamical verification in the previous section shows that the WRF model better captures the dynamical evolution for this AR event.



	P_{WRF_R}	P_{MERRA2}	P_{MERRA2_c}	P_{WRF_R-CNN}	P_{WRF_R-RCNN}
r	0.93	0.9	0.9	0.86	0.83
RMSE	0.47	0.98	0.46	0.52	0.58



	P_{WRF}	P_{MERRA2}	P_{MERRA2_c}	$P_{WRF-CNN}$	$P_{WRF-RCNN}$
r	0.49	0.9	0.9	0.67	0.66
RMSE	0.94	0.98	0.46	0.71	0.74

Figure 4.13: Precipitation time series for 00:00 UTC 19 October 2017 – 23:00 UTC 23 October 2017 at Grid 8 in the selected domain. The top figure shows precipitation observations, precipitation estimates from MERRA2, precipitation estimates from WRF that are forced by GFS reanalysis, as well as precipitation estimates from neural network models that are forced by WRF products. Similar denotations for the bottom figure, but the WRF simulations are forced by GFS forecast.

The key findings in Figure 4.13 are summarized as follows:

- Considering precipitation estimates forced by reanalysis dynamical field:

  1. Compared to MERRA2 precipitation products (P_{MERRA2}, P_{MERRA2_C}), dynamical downscaling using WRF (P_{WRF_R}) significantly improves precipitation estimate accuracy for the selected case.

2. Precipitation estimates using deep neural network models ($P_{\text{WRF-CNN}}$, $P_{\text{WRF-RCNN}}$) capture the general evolution. However, the accuracy is not as good as P_{WRF_R}.

- Considering precipitation estimates forced dynamical forecasts:

1. Dynamical forecast error is considerably revealed in WRF's precipitation product, as we compare P_{WRF_R} in the top figure with P_{WRF} in the bottom figure.
2. Precipitation estimates from the deep neural network models ($P_{\text{WRF-CNN}}$, $P_{\text{WRF-RCNN}}$) are robust to their input errors from dynamical estimations for accurate precipitation predictions in the selected case.

4.6 Conclusions

The idea of "end-to-end" learning for inferring unresolved precipitation process based on resolved atmospheric field is further explored in this Chapter for hourly scale quantitative precipitation forecast. Hourly precipitation observations from various sources are collected, quality controlled, and concatenated to compose a peculiar long term (1980/1/1- 2018/12/31) high temporal resolution precipitation observation dataset. A general framework for statistically modeling of spatiotemporal data and making use of inconsistently available observations is developed.

Results show that hourly precipitation predictions using the deep neural network model give $r \approx 0.8$ at $2^\circ \times 2.5^\circ$ spatial scale, while the baseline numerical model achieved $r \approx 0.5$. The best performance at hourly, gauge-point scale reaches the order of $r \approx 0.6$ for some gauges. However, there is high variance for skills in estimating precipitation at such a stringent scale. To further test the proposed model in practical forecasts, dynamical retrospective forecast experiments for two atmospheric river land-falling events are carried out using the Weather Research and Forecasting (WRF) model. The WRF dynamical simulations are used to force the trained neural network model for alternative precipitation process estimations.

Simulations verified the consistency and robustness of the proposed approach. It should be noted that the methods here are not intended to replace precipitation-related parameterization schemes using a "black box" model, rather, the objective is to give a benchmark for precipitation prediction from a data-driven perspective, and offer directions for improving precipitation related parameterizations.

Chapter 5

Conclusions

5.1 Key Findings

This dissertation conducted an assessment of prediction skills and an exploration of predictability for the precipitation process. The West Coast of United States is selected as the main study area. Data from various sources, including observations, numerical retrospective forecasts, numerical analysis and predictions are employed. Simple statistical analysis and complicated deep neural network modeling are explored to seek evidences for improving precipitation predicion using a composite of models and data. The key findings are listed as follows.

5.1.1 Assessment of Precipitation Prediction Skills

A seamless assessment of precipitation prediction skill for daily range up to subseasonal scale range is conducted. The evaluation is based on the Subseasonal-to-Seasonal Prediction Project retrospective forecast database (S2S). The evaluated models are frequently restarted through the past 20 more years, offering a unique opportunity for comprehensive and less-biased evaluation. The prediction skill–lead time relationship is assessed using both deterministic and probabilistic skill scores. The key findings in this assessment work are

listed as follows:

1. For Week 1 forecast, the evaluated models show advantageous precipitation prediction skills. The r/NSE/ROC score is approximately of the order of 0.8/0.7/0.8 for this period. By spatial averaging, the skill score can be further improved.

2. For Week 2, models show large variations regarding their performances. The Week-2 mean precipitation forecast from the best-performing model (i.e., ECMWF) is of considerable value, with $r > 0.6$, NSE> 0.35, and ROC score > 0.7.

3. Beyond Week 2, predictions generally provide little deterministic skill. For this range period, probabilistic evaluation of ensemble forecasts using the $\overline{\text{CRPS}}$ shows significant advantage of ensemble forecast over deterministic forecast.

4. Considering the performance difference of the S2S models, the informative predictable range may differ by up to $6-7$ days across different models. For the short range, models with higher resolution tend to have better performances (JMA, KMA, ECCC, and ECMWF). For medium to extended range, ensemble mean predictions show significant better performance compared to deterministic predictions. The best performing models for this range period are the ECCC, ECMWF, and JMA. For Week 3–4 forecast, although there is essentially no useful deterministic forecast skill, the ECMWF model still shows advantage over the rest models. Results here can benefit model selections for practical forecasts and multi-model ensemble predictions.

5.1.2 Opportunity of Predictability for Extended to Subseasonal Range

The opportunity of predictability for extended to subseasonal range precipitation prediction is explored based on statistical analysis and numerical modeling. The key findings are listed as follows:

1. Through investigating the impact of ENSO on the West Coast precipitation distribution and models' prediction skill, I found a spatial see-saw effect for ENSO to modulate precipitation distribution and prediction skill:

    - During El Niño years, Southern California receives more precipitation in late winter on average, and most models show better extended-range prediction skills.
    - During La Niña years, Oregon receives more precipitation in winter season, with most models showing better extended-range prediction skills.

    For Northern California or Washington, ENSO influences the precipitation distribution, but specific models may either have higher or lower prediction skills depending on the ENSO phases. We assume the excessive precipitation and improved extended-range prediction skills accompany the meridional shift of baroclinic systems as modulated by ENSO. This predictability difference related to ENSO phase will be useful for extended-range prediction applications.

2. The impact of MJO on the West Coast precipitation distribution and models' prediction skill is explored.

    - To assess the impact of MJO on precipitation distribution, we examined the average precipitation anomalies conditioned on the MJO phases and days after MJO phases. Our results show that MJO systematically modulates the region's precipitation distribution. The time lag (here, up to three weeks) for MJO to manifest its effects provides valuable potential for skillful predictions at the extended range.
    - Regarding the impact of the MJO on GCM's extended-range precipitation prediction skills, we verified that for certain MJO phases (especially, Phases 5-6 and 7-8), some S2S models can well capture the MJO-associated teleconnections, improving Week 3–4 prediction skills. However, for hindcasts initialized during active MJO in Phase 3–4, most models show lower extended-range prediction skills as

compared to MJO-quiescent cases, suggesting that the forecast opportunity may also be a curse if models have deficiencies in capturing the MJO or the related teleconnections.

5.1.3 Improving Precipitation Estimation with Convolutional Neural Network

The Convolutional Neural Network (CNN) model is introduced to the geo community to improve precipitation estimation accuracy at daily, grid scale. The CNN model stacks several convolution and pooling operators to extract the intricate spatial circulation features for precipitation estimation. Instead of applying pre-engineered feature extractors, the model applies "end-to-end" learning. Specifically, the kernels that are used to search and extract the salient features from the resolved dynamical field are optimized by backpropagating the precipitation estimation error through the convolutional layers. Thus, the learned features are determined based on the relation between the predictors and the predictand for the exact learning target. Also, through hierarchical convolution, we can well disintegrate dominant circulation features of different geometric properties and from different locations.

The key findings from the experiments and analyses for this work are summarized as follows

1. Through the case study of precipitation estimation, I demonstrate that the CNN is a promising approach for climate downscaling. In this sense, the work here is closely related to [191] and [28].

2. The CNN precipitation estimation problem is illustrated and formulated with realistic precipitation event cases. I justify the motivation by relating the model to the classical phenomenological understandings of precipitation mechanism (i.e., the Bergen school cyclone model).

3. I restrict the predictors to the variables that are directly resolved by discretizing the

atmospheric dynamics equations. The physically solid and comprehensive information from atmospheric dynamical modeling has not been touched in previous CNN downscaling applications.

4. The model can seamlessly be incorporated in numerical precipitation prediction. Compared to the raw precipitation product from numerical models, the model here shows enhanced precipitation estimation when trained with abundant data.

5. The performance improvement provides important implications for improving precipitation-related parameterization schemes using a data-driven approach. In this sense, the work here is closely related to [52] and [149].

6. I examine the impact of the network architectures on model performance. The results demonstrate the advantage of deep CNN to conventional fully connected Artificial Neural Networks for precipitation downscaling.

7. I provide simple visualization and analyzing approaches to interpret the models and their results.

8. Through comparing the performance between CNN and fully connected neural network, linear regression, nearest neighbor, and random forest, Iempirically verify the effectiveness of CNN for precipitation estimation.

5.1.4 Benchmarking Quantitative Precipitation Forecast using a Composite of Numerical Modeling and Deep Neural Networks

The idea of "end-to-end" learning for inferring unresolved precipitation process based on resolved atmospheric field is further explored for hourly scale quantitative precipitation forecast. Hourly precipitation observations from various sources are collected, quality controlled,

and concatenated to compose a unique long term (1980/1/1- 2018/12/31) high temporal resolution precipitation observation dataset. A general framework for statistically modeling of spatiotemporal data and making use of inconsistently available observations is developed. Hourly precipitation predictions using the deep neural network model give $r \approx 0.8$ at $2^\circ \times 2.5^\circ$ spatial scale, while the baseline numerical model achieved $r \approx 0.5$. The best performance at hourly, gauge-point scale reaches the order of $r \approx 0.6$ for some gauges. However, there is high variance for skills in estimating precipitation at such a stringent scale. To further test the proposed model in practical forecasts, dynamical retrospective forecast experiments for two atmospheric river land-falling events are carried out using the Weather Research and Forecasting (WRF) model. The WRF dynamical simulations are used to force the trained neural network model for alternative precipitation process estimations. Simulations verified the consistency and robustness of the proposed approach. It should be noted that the methods here are not intended to replace precipitation-related parameterization schemes using a "black box" model, rather, the objective is to give a benchmark for precipitation prediction from a data-driven perspective, and offer directions for improving precipitation related parameterizations.

Overall, this work conducted a systematical evaluations of precipitation prediction skills across a spectrum of critical scales and ranges. Sources of predictability at subseasonal scale are explored based on a composite of statistical analysis and numerical prediction. The potential of deep learning for seeking evidences in improving precipitation prediction is explored by combining high quality observation data with numerical dynamical predictions.

5.2 Deficiencies and Future works

Results in this thesis suggest that, by combining numerical modeling with data from various sources, it is possible to improve precipitation prediction by a large margin. Advances in deep learning techniques offer powerful tools for seeking evidences and realizing these potential

improvements. It should be admitted that the work here only touches limited part toward this target. Some of the key deficiencies of this dissertation and promising directions for future works are highlighted as follows:

- **Making sufficient use of prior knowledge.** Data-driven models are applied to combine data with prior knowledge in order to generate new knowledge. Through the past decades, the meteorology community has gained important theoretical understandings of atmospheric dynamics and physics. Many of the findings are derived from mathematical analysis of the primitive equations, which offer insights into critical aspects toward understanding the atmospheric processes. On the other hand, machine learning models are very often criticized for ignoring these achievements and offering little accumulative progresses toward reliable predictions. Deep neural networks have flexible structures that can be tailored to encode prior knowledge into the model settings. We hope to leverage better modeling and predictions from the following directions:

  - **Predictor selection and design:**. The geopotential height and total column moisture contents are used as the predictors in the considered models here. Theoretical analysis of the primitive equations shows that many critical aspects that dominate the dynamical environment of precipitation can be extracted through analytical simplification of the governing equations. For instance, the omega equation relates vertical velocity with instantaneous geopotential field, and has been widely employed to assess the development of vertical motions from synoptic charts. To adopt the findings from such theoretical analysis for predictor selection and design may alleviate the data amount requirement, enable the generalization of model for different regions, and improve model accuracy.

  - **Network architecture:** The inclusion of convolution structure in deep neural network modeling significantly improves precipitation estimation accuracy in the considered experiments. We can potentially make more accurate predictions by

further optimizing the network architecture through explicitly considering the critical scales and spatiotemporal coherence of atmospheric dynamics.

  - **Loss function and model regularization:** A well-designed loss function makes it easier for model training, and usually yields more satisfying results. Also, it is important to regularize the models by adding physical constraints to the statistical models.

- **The distinct difficulty in predicting convective precipitation.** While most of the focus here is on stratiform precipitation, convective precipitation holds distinct mechanisms, spatiotemporal scales and characteristics (more intense, shorter duration). Also, it poses a more difficult challenge for most numerical weather prediction models. It is important to make a dedicated verification and modification of the proposed models for convective precipitation predictions.

- **Key problems in precipitation prediction are not sufficiently touched.** As has been highlighted, the ever-accumulating advances in predicting the precipitation process has been achieved due to (1) improvement in resolving atmospheric dynamics, (2) improvement in inferring unresolved cloud and precipitation processes, and (3) improvement in inferring the initial status of the hydrometeor distributions. Any further advances are assumed to come from the above-mentioned aspects. It has been pointed out that data-driven models offer insights about predictability limits that are achievable with better modeling and statistical analysis techniques. However, it is still not clear how we could realize these potentials. Below I briefly discuss some of the potential directions in improving precipitation predictions.

  - **Assimilation of Cloud and Precipitation Data:** The assimilation of cloud and precipitation characteristics has been widely recognized to be crucial for improving precipitation forecasts [43]. However, fundamental challenges remain before we could successfully infer the initial status of hydrometeor distribution from

indirect and insufficient observations. The key challenges include the non-linearity of the forward models, variable non-normality, etc. While some research works have explored the potential of data-driven models in statistical data assimilation tasks [152, 170], it still remains unclear how we can leverage the power of deep learning for more efficient and effective inference of the atmospheric states.

  - **Better modeling of ocean-atmosphere interactions to better reveal teleconnections:** Results in Chapter 2 suggest the existence of opportunity of predictability provided by climate variance from tropical latent heating signals. On the other hand, it is also found that many models have severe deficiencies in utilizing these opportunities for enhanced predictions at extended to seasonal ranges [38, 143]. It is reasonable to assume that provided with better simulation of the ocean-atmosphere interactions at key tropical regions, models can make more informative extended range predictions.

- **Many powerful learning paradigms remain to be explored to bring further advances.**

  - **Reinforcement learning:** The machine learning algorithms explored here belong mostly to the category of supervised learning methods, which learn from *examples* to perform classification or regression tasks. Another powerful learning paradigm, named reinforcement learning, learns from *experiences* to make the agent take optimized actions in an environment, so as to maximize some notion of cumulative reward [180]. The formulation of reinforcement learning problems shares striking similarity with the statistical data assimilation task. With the help of deep neural networks, reinforcement learning has achieved significant progresses in many applications, such as playing games and automatic driving. We wish deep reinforcement learning can as well enhance better merging of models and data for more reliable predictions of the earth system.

– **Transfer learning:** Transfer learning refers to machine learning method where a model developed for a task is reused as the starting point for a model on another task [25]. While the models tested in this thesis work are built separately for each study region, it is reasonable to assume that we could transfer what is learned from one study area to another area. Improvements in this direction would significantly alleviate the data requirement, leverage our understandings about what is learned from the statistical model, and bring potentials for predictions in ungauged regions.

– **Few-shot learning:** There are many classical machine learning datasets for benchmarking learning algorithms. Most of them are artificially balanced [203], where objects of different classes of predictands are approximately evenly distributed. In the real world, the phenomena we focus on are mostly long-tail distributed: we have many trivial examples, but much less samples of extreme events. Existing studies have verified that a less-balanced training dataset results in a corresponding poor modeling performance for rare events. Learning to model the rare events consists a major objective of *few-shot learning*. In future works, it is important to pay special attention to model's performance in modeling the rare and extreme events. Also, it is crucial to employ techniques that stress few-shot learning for the modeling of geophysical processes.

– **Bayesian neural networks:** The prediction of subgrid scale processes in numerical modeling of the atmosphere is of inherent limitation of predictability. It is imperative to know the uncertainty of predictions. While most of the works here focus on deterministic learning, many existing studies have also developed practical uncertainty estimates in deep learning modeling [49]. We wish future works to apply these Bayesian neural networks to help determine model uncertainty properties.

Bibliography

[1] C. Abbe. The physical basis of long-range weather forecasts. *Monthly Weather Review*, 29(12):551–561, 1901.

[2] R. B. Alley, K. A. Emanuel, and F. Zhang. Advances in weather prediction. *Science*, 363(6425):342–344, 2019.

[3] O. Alves, G. Wang, A. Zhong, N. Smith, F. Tseitkin, G. Warren, A. Schiller, S. Godfrey, and G. Meyers. POAMA: Bureau of Meteorology operational coupled model seasonal forecast system. In *Proceedings of National Drought Forum*, pages 49–56, Brisbane, Queensland, Australia, 2003.

[4] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

[5] P. V. Ayar, M. Vrac, S. Bastin, J. Carreau, M. Déqué, and C. Gallardo. Intercomparison of statistical and dynamical downscaling models under the EURO-and MED-CORDEX initiative framework: present climate evaluations. *Climate Dynamics*, 46(3-4):1301–1329, 2016.

[6] M. Bader and W. Roach. Orographic rainfall in warm sectors of depressions. *Quarterly Journal of the Royal Meteorological Society*, 103(436):269–280, 1977.

[7] C. F. Baggett, E. A. Barnes, E. D. Maloney, and B. D. Mundhenk. Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. *Geophys. Res. Lett.*, 44(14):7528–7536, 2017.

[8] J. Bao, S. Michelson, P. Neiman, F. Ralph, and J. Wilczak. Interpretation of enhanced integrated water vapor bands associated with extratropical cyclones: Their formation and connection to tropical moisture. *Monthly weather review*, 134(4):1063–1080, 2006.

[9] P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

[10] E. J. Becker, E. H. Berbery, and R. W. Higgins. Understanding the characteristics of daily precipitation over the United States using the North American Regional Reanalysis. *Journal of Climate*, 22(23):6268–6286, 2009.

[11] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[12] Y. Bengio, P. Simard, P. Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[13] J. Berner, U. Achatz, L. Batté, L. Bengtsson, A. d. l. Cámara, H. M. Christensen, M. Colangeli, D. R. Coleman, D. Crommelin, S. I. Dolaptchiev, et al. Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, 98(3):565–588, 2017.

[14] M. Best, M. Pryor, D. Clark, G. Rooney, R. Essery, C. Ménard, J. Edwards, M. Hendry, A. Porson, N. Gedney, et al. The Joint UK Land Environment Simulator (JULES), model description–Part 1: energy and water fluxes. *Geoscientific Model Development*, 4(3):677–699, 2011.

[15] A. K. Betts, P. Viterbo, and E. Wood. Surface energy and water balance for the Arkansas–Red River basin from the ECMWF reanalysis. *Journal of Climate*, 11(11):2881–2897, 1998.

[16] J. Bjerknes. Extratropical cyclones. In *Compendium of Meteorology*, pages 577–598. Springer, 1951.

[17] J. Bjerknes. Atmospheric teleconnections from the equatorial Pacific. *Mon. Wea. Rev.*, 97(3):163–172, 1969.

[18] V. F. K. Bjerknes. Fields of force. 1906.

[19] N. A. Bond and G. A. Vecchi. The influence of the Madden–Julian oscillation on precipitation in Oregon and Washington. *Wea. Forecasting*, 18(4):600–613, 2003.

[20] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[21] J. Bouvrie. Notes on convolutional neural networks. 2006.

[22] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[23] T. J. Brown, F. Fujioka, and C. Fontana. The california and nevada smoke and air committee (cansac)—an interagency partnership to meet decision-making needs. In *5th Symposium on Fire and Forest Meteorology*, 2003.

[24] M. S. Bukovsky and D. J. Karoly. A brief evaluation of precipitation from the North American Regional Reanalysis. *Journal of Hydrometeorology*, 8(4):837–846, 2007.

[25] L. A. Burke and H. M. Hutchins. Training transfer: An integrative literature review. *Human resource development review*, 6(3):263–296, 2007.

[26] J. S. Bushong. Quantitative precipitation forecast: Its generation and verification at the southeast river forecast center. Georgia Institute of Technology, 1999.

[27] P. Caldwell, H.-N. S. Chin, D. C. Bader, and G. Bala. Evaluation of a wrf dynamical downscaling simulation over california. *Climatic change*, 95(3-4):499–521, 2009.

[28] Y.-C. Chang, R. Acierto, T. Itaya, K. Akiyuki, and C.-P. Tung. A Deep Learning Approach to Downscaling Precipitation and Temperature over Myanmar. In *EGU General Assembly Conference Abstracts*, volume 20, page 4120, 2018.

[29] S.-C. Chen, J. O. Roads, and J. C. Alpert. Variability and predictability in an empirically forced global model. *J. Atmos. Sci.*, 50(3):443–463, 1993.

[30] S.-H. Chen and W.-Y. Sun. A one-dimensional time dependent cloud model. *Journal of the Meteorological Society of Japan. Ser. II*, 80(1):99–118, 2002.

[31] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[32] P. Courtier and J.-F. Geleyn. A global numerical weather prediction model with variable resolution: Application to the shallow-water equations. *Quart. J. Roy. Meteor. Soc.*, 114(483):1321–1346, 1988.

[33] H. F. Dacre, P. A. Clark, O. Martinez-Alvarado, M. A. Stringer, and D. A. Lavers. How do atmospheric rivers form? *Bull. Amer. Meteor. Soc.*, 96(8):1243–1255, 2015.

[34] E. de Bezenac, A. Pajot, and P. Gallinari. Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge. *arXiv preprint arXiv:1711.07970*, 2017.

[35] M. Dettinger. Climate change, atmospheric rivers, and floods in California – a multimodel analysis of storm frequency and magnitude changes. *Journal of the American Water Resources Association*, 47(3):514–523, 2011.

[36] M. D. Dettinger. Atmospheric rivers as drought busters on the us west coast. *J. Hydrometeor.*, 14(6):1721–1732, 2013.

[37] M. D. Dettinger, F. M. Ralph, T. Das, P. J. Neiman, and D. R. Cayan. Atmospheric rivers, floods and the water resources of california. *Water*, 3(2):445–478, 2011.

[38] J. Dias and G. N. Kiladis. The influence of tropical forecast errors on higher latitude predictions. *Geophysical Research Letters*, 46(8):4450–4459, 2019.

[39] P. M. Domingos. A few useful things to know about machine learning. *Commun. acm*, 55(10):78–87, 2012.

[40] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 45–97, 1995.

[41] J. Dudhia. Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *Journal of the atmospheric sciences*, 46(20):3077–3107, 1989.

[42] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[43] R. M. Errico, P. Bauer, and J.-F. Mahfouf. Issues regarding the assimilation of cloud and precipitation data. *Journal of the Atmospheric Sciences*, 64(11):3785–3798, 2007.

[44] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31(1):1–38, 2004.

[45] R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[46] H. J. Fowler, S. Blenkinsop, and C. Tebaldi. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International journal of climatology*, 27(12):1547–1578, 2007.

[47] R. A. Fulton, J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon. The wsr-88d rainfall algorithm. *Weather and Forecasting*, 13(2):377–395, 1998.

[48] N. Gagnon, X. Deng, P. Houtekamer, M. Charron, A. Erfani, S. Beauregard, B. Archambault, F. Petrucci, and A. Giguère. Improvements to the Global Ensemble Prediction System (GEPS) from version 2.0. 3 to version 3.0. 0. *Technical Note*, 2013.

[49] Y. Gal. *Uncertainty in deep learning*. PhD thesis, PhD thesis, University of Cambridge, 2016.

[50] S. Gangopadhyay, M. Clark, and B. Rajagopalan. Statistical downscaling using k-nearest neighbors. *Water Resources Research*, 41(2), 2005.

[51] R. Gelaro, W. McCarty, M. J. Suárez, R. Todling, A. Molod, L. Takacs, C. A. Randles, A. Darmenov, M. G. Bosilovich, R. Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of Climate*, 30(14):5419–5454, 2017.

[52] P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis. Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, 2018.

[53] D. Gesch, M. Oimoen, S. Greenlee, C. Nelson, M. Steuck, and D. Tyler. The National Elevation Dataset. *Photogrammetric Engineering and Remote Sensing*, 68(1):5–11, 2002.

[54] M. Ghil. Hilbert problems for the geosciences in the 21st century. *Nonlinear processes in geophysics*, 8(4/5):211–211, 2001.

[55] H. R. Glahn and D. A. Lowry. The use of model output statistics (MOS) in objective weather forecasting. *Journal of applied meteorology*, 11(8):1203–1211, 1972.

[56] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

[57] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[58] E. D. Gutmann, R. M. Rasmussen, C. Liu, K. Ikeda, D. J. Gochis, M. P. Clark, J. Dudhia, and G. Thompson. A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *Journal of Climate*, 25(1):262–281, 2012.

[59] A. Hannachi, I. Jolliffe, and D. Stephenson. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27(9):1119–1152, 2007.

[60] M. Hawcroft, L. Shaffrey, K. Hodges, and H. Dacre. How much northern hemisphere precipitation is associated with extratropical cyclones? *Geophysical Research Letters*, 39(24), 2012.

[61] M. R. Haylock, G. C. Cawley, C. Harpham, R. L. Wilby, and C. M. Goodess. Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. *International Journal of Climatology*, 26(10):1397–1415, 2006.

[62] H. H. Hendon, B. Liebmann, M. Newman, J. D. Glick, and J. Schemm. Medium-range forecast errors associated with active episodes of themadden–julian oscillation. *Mon. Wea. Rev.*, 128(1):69–86, 2000.

[63] H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, 15(5):559–570, 2000.

[64] R. Higgins, K. Mo, and S. Schubert. The moisture budget of the central United States in spring as evaluated in the NCEP/NCAR and the NASA/DAO reanalyses. *Monthly Weather Review*, 124(5):939–963, 1996.

[65] R. Higgins, J. E. Schemm, W. Shi, and A. Leetmaa. Extreme precipitation events in the western United States related to tropical forcing. *J. Climate*, 13(4):793–820, 2000.

[66] R. W. Higgins, J. E. Janowiak, and Y.-P. Yao. *A gridded hourly precipitation data base for the United States (1963-1993)*. US Department of Commerce, National Oceanic and Atmospheric Administration . . . , 1996.

[67] S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, 91(1), 1991.

[68] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[69] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[70] J. R. Holton and G. J. Hakim. *An introduction to dynamic meteorology*, volume 88. Academic press, 2012.

[71] S.-Y. Hong, Y. Noh, and J. Dudhia. A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly weather review*, 134(9):2318–2341, 2006.

[72] Y. Hong, K.-L. Hsu, S. Sorooshian, and X. Gao. Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *Journal of Applied Meteorology*, 43(12):1834–1853, 2004.

[73] P. K. Hope. Projected future changes in synoptic systems influencing southwest Western Australia. *Climate Dynamics*, 26(7-8):765–780, 2006.

[74] B. Hoskins. The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quart. J. Roy. Meteor. Soc.*, 139(672):573–584, 2013.

[75] B. J. Hoskins and D. J. Karoly. The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.*, 38(6):1179–1196, 1981.

[76] D. Hou, M. Charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D.-J. Seo, M. Pena, et al. Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage iv toward cpc gauge-based analysis. *Journal of Hydrometeorology*, 15(6):2542–2557, 2014.

[77] F. Hourdin, T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, et al. The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98(3):589–602, 2017.

[78] R. A. Houze Jr. *Cloud dynamics*, volume 104. Academic press, 2014.

[79] C. Hutengs and M. Vohland. Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, 178:127–141, 2016.

[80] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[81] M. Jakob Themeßl, A. Gobiet, and A. Leuprecht. Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology*, 31(10):1530–1544, 2011.

[82] K. Japan. *Outline of the operational numerical weather prediction at the Japan Meteorological Agency*. 2013.

[83] D. Jeong, A. St-Hilaire, T. Ouarda, and P. Gachon. Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator. *Climatic Change*, 114(3-4):567–591, 2012.

[84] C. Jones, A. Hazra, and L. M. Carvalho. The Madden–Julian oscillation and boreal winter forecast skill: An analysis of NCEP CFSv2 reforecasts. *J. Climate*, 28(15):6297–6307, 2015.

[85] B.-T. Jong, M. Ting, and R. Seager. El niño's impact on california precipitation: seasonality, regionality, and el niño intensity. *Environmental Research Letters*, 11(5):054021, 2016.

[86] J. S. Kain. The kain–fritsch convective parameterization: an update. *Journal of Applied Meteorology*, 43(1):170–181, 2004.

[87] E. Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003.

[88] I.-S. Kang and H.-M. Kim. Assessment of MJO predictability for boreal winter with various statistical and dynamical models. *J. Climate*, 23(9):2368–2378, 2010.

[89] A. Karpathy, J. Johnson, and L. Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[90] A. Khain, M. Ovtchinnikov, M. Pinsky, A. Pokrovsky, and H. Krugliak. Notes on the state-of-the-art numerical modeling of cloud microphysics. *Atmospheric Research*, 55(3-4):159–224, 2000.

[91] M. F. Khairoutdinov and D. A. Randall. A cloud resolving model as a cloud parameterization in the ncar community climate system model: Preliminary results. *Geophysical Research Letters*, 28(18):3617–3620, 2001.

[92] D. T. Kleist, D. F. Parrish, J. C. Derber, R. Treadon, R. M. Errico, and R. Yang. Improving incremental balance in the gsi 3dvar analysis system. *Monthly Weather Review*, 137(3):1046–1060, 2009.

[93] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[94] R. J. Kuligowski and A. P. Barros. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and forecasting*, 13(4):1194–1204, 1998.

[95] A. Kumar, M. Chen, and W. Wang. An analysis of prediction skill of monthly mean climate variability. *Climate Dyn.*, 37(5-6):1119–1131, 2011.

[96] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[97] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[98] S. Li and A. W. Robertson. Evaluation of submonthly precipitation forecast skill from global ensemble prediction systems. *Mon. Wea. Rev.*, 143(7):2871–2889, 2015.

[99] Y. Li and I. Smith. A statistical downscaling model for southern Australia winter rainfall. *Journal of Climate*, 22(5):1142–1158, 2009.

[100] Y. Lim, S.-W. Son, and D. Kim. MJO prediction skill of the subseasonal-to-seasonal prediction models. *J. Climate*, 31(10):4075–4094, 2018.

[101] H. Lin, G. Brunet, and J. Derome. Forecast skill of the Madden–Julian oscillation in two Canadian atmospheric models. *Mon. Wea. Rev.*, 136(11):4130–4149, 2008.

[102] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[103] Y. Lin, K. Mitchell, E. Rogers, M. Baldwin, and G. DiMego. Test assimilations of the real-time, multi-sensor hourly precipitation analysis into the NCEP Eta model. In *Preprints, 8th Conf. on Mesoscale Meteorology, Boulder, CO, Amer. Meteor. Soc*, pages 341–344, 1999.

[104] Y. Lin and K. E. Mitchell. 1.2 the ncep stage ii/iv hourly precipitation analyses: Development and applications. In *19th Conf. Hydrology*. Citeseer, 2005.

[105] J. Ling, A. Kurzawski, and J. Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, 2016.

[106] Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins, et al. Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv preprint arXiv:1605.01156*, 2016.

[107] P. Lopez. Cloud and precipitation parameterizations in modeling and variational data assimilation: A review. *Journal of the Atmospheric Sciences*, 64(11):3766–3784, 2007.

[108] E. N. Lorenz. A study of the predictability of a 28-variable atmospheric model. *Tellus*, 17(3):321–333, 1965.

[109] W. Lotter, G. Kreiman, and D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[110] G. Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

[111] S. Madadgar, A. AghaKouchak, S. Shukla, A. W. Wood, L. Cheng, K.-L. Hsu, and M. Svoboda. A hybrid statistical-dynamical framework for meteorological drought prediction: Application to the southwestern United States. *Water Resources Research*, 52(7):5095–5110, 2016.

[112] R. A. Madden and P. R. Julian. Description of global-scale circulation cells in the tropics with a 40–50 day period. *J. Atmos. Sci.*, 29(6):1109–1123, 1972.

[113] R. A. Maddox, J. Zhang, J. J. Gourley, and K. W. Howard. Weather radar coverage over the contiguous united states. *Weather and Forecasting*, 17(4):927–934, 2002.

[114] P. Malguzzi, A. Buzzi, and O. Drofa. The meteorological global model GLOBO at the ISAC-CNR of italy assessment of 1.5 yr of experimental use for medium-range weather forecasts. *Wea. Forecasting*, 26(6):1045–1055, 2011.

[115] A. Mamalakis, J.-Y. Yu, J. T. Randerson, A. AghaKouchak, and E. Foufoula-Georgiou. A new interhemispheric teleconnection increases predictability of winter precipitation in southwestern US. *Nature communications*, 9(1):2332, 2018.

[116] D. Maraun, F. Wetterhall, A. Ireson, R. Chandler, E. Kendon, M. Widmann, S. Brienen, H. Rust, T. Sauter, M. Themeßl, et al. Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3), 2010.

[117] A. Martin, F. M. Ralph, R. Demirdjian, L. DeHaan, R. Weihs, J. Helly, D. Reynolds, and S. Iacobellis. Evaluation of atmospheric river predictions by the wrf model using aircraft and regional mesonet observations of orographic precipitation and its forcing. *Journal of Hydrometeorology*, 19(7):1097–1113, 2018.

[118] R. Martínez, M. Johnson, and P. O'Connor. US Hydropower Market Report 2017 Update (April). *Oak Ridge National Laboratory*, 2017.

[119] M. Matsueda and H. Endo. Verification of medium-range MJO forecasts with TIGGE. *Geophys. Res. Lett.*, 38(11), 2011.

[120] A. J. Matthews. Atmospheric response to observed intraseasonal tropical sea surface temperature anomalies. *Geophys. Res. Lett.*, 31(14), 2004.

[121] F. Mesinger, G. DiMego, E. Kalnay, K. Mitchell, P. C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, E. H. Berbery, et al. North American regional reanalysis. *Bulletin of the American Meteorological Society*, 87(3):343–360, 2006.

[122] T. Mitchell, B. Buchanan, G. DeJong, T. Dietterich, P. Rosenbloom, and A. Waibel. Machine learning. *Annual review of computer science*, 4(1):417–433, 1990.

[123] E. J. Mlawer, S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough. Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, 102(D14):16663–16682, 1997.

[124] K. C. Mo and R. Higgins. Tropical convection and precipitation regimes in the western United States. *J. Climate*, 11(9):2404–2423, 1998.

[125] A. Molod, L. Takacs, M. Suarez, and J. Bacmeister. Development of the geos-5 atmospheric general circulation model: Evolution from merra to merra2. *Geoscientific Model Development*, 8(5):1339–1356, 2015.

[126] B. D. Mundhenk, E. A. Barnes, E. D. Maloney, and C. F. Baggett. Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden–Julian oscillation and quasi-biennial oscillation. *NPJ Climate and Atmospheric Science*, 1(1):7, 2018.

[127] J. Murphy. An evaluation of statistical and dynamical techniques for downscaling local climate. *Journal of Climate*, 12(8):2256–2284, 1999.

[128] J. Murphy et al. Predictions of climate change over Europe using statistical and dynamical downscaling techniques. *International Journal of Climatology*, 20(5):489–501, 2000.

[129] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970.

[130] E. National Academies of Sciences, Medicine, et al. *Next generation earth system prediction: strategies for subseasonal to seasonal forecasts*. National Academies Press, 2016.

[131] F. Nebeker. *Calculating the weather: Meteorology in the 20th century*, volume 60. Elsevier, 1995.

[132] J. Neena, J. Y. Lee, D. Waliser, B. Wang, and X. Jiang. Predictability of the Madden–Julian oscillation in the intraseasonal variability hindcast experiment (ISVHE). *J. Climate*, 27(12):4531–4543, 2014.

[133] P. J. Neiman, A. B. White, F. M. Ralph, D. J. Gottas, and S. I. Gutman. A water vapour flux tool for precipitation forecasting. In *Proceedings of the Institution of Civil Engineers-Water Management*, volume 162, pages 83–94. Thomas Telford Ltd, 2009.

[134] B. R. Nelson, O. P. Prat, D.-J. Seo, and E. Habib. Assessment and implications of ncep stage iv quantitative precipitation estimates for product intercomparisons. *Weather and Forecasting*, 31(2):371–394, 2016.

[135] P. Nguyen, M. Ombadi, S. Sorooshian, K. Hsu, A. AghaKouchak, D. Braithwaite, H. Ashouri, and A. R. Thorstensen. The persiann family of global satellite precipitation data: a review and evaluation of products. *Hydrology and Earth System Sciences*, 22(11):5801–5816, 2018.

[136] P. Nguyen, E. J. Shearer, H. Tran, M. Ombadi, N. Hayatbini, T. Palacios, P. Huynh, D. Braithwaite, G. Updegraff, K. Hsu, et al. The chrs data portal for distributing persiann family global satellite precipitation data. In *AGU Fall Meeting Abstracts*, 2018.

[137] OECD. Population (indicator). 2018.

[138] M. Ombadi, P. Nguyen, S. Sorooshian, and K.-l. Hsu. Developing intensity-duration-frequency (idf) curves from satellite-based precipitation: Methodology and evaluation. *Water Resources Research*, 54(10):7752–7766, 2018.

[139] S. A. Orszag. Analytical theories of turbulence. *Journal of Fluid Mechanics*, 41(2):363–386, 1970.

[140] T. Palmer and R. Hagedorn. *Predictability of weather and climate*. Cambridge University Press, 2006.

[141] B. Pan, K. Hsu, A. AghaKouchak, and S. Sorooshian. Improving precipitation estimation using convolutional neural network. *Water Resources Research*.

[142] B. Pan, K. Hsu, A. AghaKouchak, and S. Sorooshian. The Use of Convolutional Neural Network in Relating Precipitation to Circulation. In *AGU Fall Meeting Abstracts*, 2017.

[143] B. Pan, K. Hsu, A. AghaKouchak, S. Sorooshian, and W. Higgins. Precipitation prediction skill for the west coast united states: From short to extended range. *Journal of Climate*, 32(1):161–182, 2019.

[144] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.

[145] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[146] A. F. Prein, R. M. Rasmussen, K. Ikeda, C. Liu, M. P. Clark, and G. J. Holland. The future intensification of hourly precipitation extremes. *Nature Climate Change*, 7(1):48, 2017.

[147] F. M. Ralph, P. J. Neiman, G. A. Wick, S. I. Gutman, M. D. Dettinger, D. R. Cayan, and A. B. White. Flooding on california's russian river: Role of atmospheric rivers. *Geophysical Research Letters*, 33(13), 2006.

[148] E. M. Rasmusson and J. M. Wallace. Meteorological aspects of the El Niño-Southern Oscillation. *Science*, 222(4629):1195–1202, 1983.

[149] S. Rasp, M. S. Pritchard, and P. Gentine. Deep learning to represent sub-grid processes in climate models. *arXiv preprint arXiv:1806.04731*, 2018.

[150] R. H. Reichle and Q. Liu. Observation-corrected precipitation estimates in geos-5. 2014.

[151] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195, 2019.

[152] D. Rey. *Chaos, observability and symplectic structure in optimal estimation*. PhD thesis, UC San Diego, 2017.

[153] A. W. Robertson, A. Kumar, M. Peña, and F. Vitart. Improving and promoting subseasonal to seasonal prediction. *Bull. Amer. Meteor. Soc.*, 96(3):ES49–ES53, 2015.

[154] C.-G. Rossby. Relation between variations in the intensity of the zonal circulation of the atmosphere and the displacements of the semi-permanent centers of action. *J. Mar. Res.*, 2:38–55, 1939.

[155] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[156] S. Saha, S. Moorthi, H.-L. Pan, X. Wu, J. Wang, S. Nadiga, P. Tripp, R. Kistler, J. Woollen, D. Behringer, et al. The ncep climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, 91(8):1015–1058, 2010.

[157] S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-y. Chuang, M. Iredell, et al. The NCEP climate forecast system version 2. *J. Climate*, 27(6):2185–2208, 2014.

[158] P. D. Sardeshmukh and B. J. Hoskins. The generation of global rotational flow by steady idealized tropical divergence. *J. Atmos. Sci.*, 45(7):1228–1251, 1988.

[159] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[160] J. Schmidli, C. Goodess, C. Frei, M. Haylock, Y. Hundecha, J. Ribalaygua, and T. Schmith. Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps. *Journal of Geophysical Research: Atmospheres*, 112(D4), 2007.

[161] T. Schneider, S. Lan, A. Stuart, and J. Teixeira. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, 44(24), 2017.

[162] T. Schonher and S. Nicholson. The relationship between California rainfall and ENSO events. *J. Climate*, 2(11):1258–1269, 1989.

[163] J. T. Schoof and S. Pryor. Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks. *International Journal of climatology*, 21(7):773–790, 2001.

[164] R. Seager, M. Hoerling, S. Schubert, H. Wang, B. Lyon, A. Kumar, J. Nakamura, and N. Henderson. Causes of the 2011–14 california drought. *Journal of Climate*, 28(18):6997–7024, 2015.

[165] D.-J. Seo and J. Breidenbach. Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *Journal of Hydrometeorology*, 3(2):93–111, 2002.

[166] T. Shaw, M. Baldwin, E. Barnes, R. Caballero, C. Garfinkel, Y.-T. Hwang, C. Li, P. O'Gorman, G. Rivière, I. Simpson, et al. Storm track processes and the opposing influences of climate change. *Nature Geoscience*, 2016.

[167] C. Shen. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 2018.

[168] C. Shen, E. Laloy, A. Elshorbagy, A. Albert, J. Bales, F.-J. Chang, S. Ganguly, K.-L. Hsu, D. Kifer, Z. Fang, et al. Hess opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, 22(11):5639–5656, 2018.

[169] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in Neural Information Processing Systems*, pages 5617–5627, 2017.

[170] A. Shirman. *Strategic Monte Carlo and Variational Methods in Statistical Data Assimilation for Nonlinear Dynamical Systems*. PhD thesis, UC San Diego, 2018.

[171] J. Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[172] M. S. Sikder and F. Hossain. Sensitivity of initial-condition and cloud microphysics to the forecasting of monsoon rainfall in south asia. *Meteorological Applications*, 25(4):493–509, 2018.

[173] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[174] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers. A description of the advanced research wrf version 2. Technical report, National Center For Atmospheric Research Boulder Co Mesoscale and Microscale ..., 2005.

[175] M. Smalley, T. L'Ecuyer, M. Lebsock, and J. Haynes. A comparison of precipitation occurrence from the ncep stage iv qpe product and the cloudsat cloud profiling radar. *Journal of Hydrometeorology*, 15(1):444–458, 2014.

[176] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011.

[177] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[178] D. J. Stensrud. *Parameterization schemes: keys to understanding numerical weather prediction models*. Cambridge University Press, 2009.

[179] G. L. Stephens, T. L'Ecuyer, R. Forbes, A. Gettelmen, J.-C. Golaz, A. Bodas-Salcedo, K. Suzuki, P. Gabriel, and J. Haynes. Dreary state of precipitation in global models. *Journal of Geophysical Research: Atmospheres*, 115(D24), 2010.

[180] R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

[181] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[182] J. Tang, X. Niu, S. Wang, H. Gao, X. Wang, and J. Wu. Statistical downscaling and dynamical downscaling of regional climate in china: Present climate evaluations and future climate projections. *Journal of Geophysical Research: Atmospheres*, 121(5):2110–2129, 2016.

[183] Y. Tao, X. Gao, K. Hsu, S. Sorooshian, and A. Ihler. A deep neural network modeling framework to reduce bias in satellite precipitation products. *Journal of Hydrometeorology*, 17(3):931–945, 2016.

[184] F. J. Tapiador, R. Roca, A. D. Genio, B. Dewitte, W. Petersen, and F. Zhang. Is precipitation a good metric for model performance? *Bulletin of the American Meteorological Society*, (2018), 2018.

[185] M. Tewari, F. Chen, W. Wang, J. Dudhia, M. LeMone, K. Mitchell, M. Ek, G. Gayno, J. Wegiel, and R. Cuenca. Implementation and verification of the unified noah land surface model in the wrf model. In *20th conference on weather analysis and forecasting/16th conference on numerical weather prediction*, volume 1115. American Meteorological Society Seattle, WA, 2004.

[186] D. Tian, E. F. Wood, and X. Yuan. CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrology and Earth System Sciences*, 21(3):1477, 2017.

[187] J. Tompson, K. Schlachter, P. Sprechmann, and K. Perlin. Accelerating eulerian fluid simulation with convolutional networks. *arXiv preprint arXiv:1607.03597*, 2016.

[188] K. E. Trenberth, G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. Ropelewski. Progress during toga in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res.: Oceans*, 103(C7):14291–14324, 1998.

[189] K. E. Trenberth and D. P. Stepaniak. Indices of el Niño evolution. *J. Climate*, 14(8):1697–1701, 2001.

[190] S. Tripathi, V. Srinivas, and R. S. Nanjundiah. Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of hydrology*, 330(3-4):621–640, 2006.

[191] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly. Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1663–1672. ACM, 2017.

[192] T. Vilsack and J. T. Reilly. Census of Agriculture–Farm and Ranch Irrigation Survey (2013), United States Department of Agriculture (USDA), National Agricultural Statistics Survey (NASS), 2013.

[193] F. Vitart. Monthly forecasting at ECMWF. *Mon. Wea. Rev.*, 132(12):2761–2779, 2004.

[194] F. Vitart. Evolution of ECMWF sub-seasonal forecast skill scores. *Quart. J. Roy. Meteor. Soc.*, 140(683):1889–1899, 2014.

[195] F. Vitart, C. Ardilouze, A. Bonet, A. Brookshaw, M. Chen, C. Codorean, M. Déqué, L. Ferranti, E. Fucile, M. Fuentes, et al. The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bull. Amer. Meteor. Soc.*, 98(1):163–173, 2017.

[196] F. Vitart, R. Buizza, M. Alonso Balmaseda, G. Balsamo, J.-R. Bidlot, A. Bonet, M. Fuentes, A. Hofstadler, F. Molteni, and T. N. Palmer. The new VAREPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, 134(636):1789–1799, 2008.

[197] F. Vitart and F. Molteni. Simulation of the Madden–Julian oscillation and its teleconnections in the ECMWF forecast system. *Quart. J. Roy. Meteor. Soc.*, 136(649):842–855, 2010.

[198] P. Vogel, P. Knippertz, A. H. Fink, A. Schlueter, and T. Gneiting. Skill of global raw and postprocessed ensemble predictions of rainfall over northern tropical Africa. *Wea. Forecasting*, 33(2):369–388, 2018.

[199] A. Voldoire, E. Sanchez-Gomez, D. S. y Mélia, B. Decharme, C. Cassou, S. Sénési, S. Valcke, I. Beau, A. Alias, M. Chevallier, et al. The CNRM-CM5. 1 global climate model: description and basic evaluation. *Climate Dyn.*, 40(9-10):2091–2121, 2013.

[200] B. Wang, J.-Y. Lee, I.-S. Kang, J. Shukla, C.-K. Park, A. Kumar, J. Schemm, S. Cocke, J.-S. Kug, J.-J. Luo, et al. Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dyn.*, 33(1):93–117, 2009.

[201] W. Wang, M.-P. Hung, S. J. Weaver, A. Kumar, and X. Fu. MJO prediction in the NCEP Climate Forecast System version 2. *Climate Dyn.*, 42(9-10):2509–2520, 2014.

[202] X. Wang, L. Gao, J. Song, and H. Shen. Beyond frame-level cnn: saliency-aware 3-d cnn with lstm for video action recognition. *IEEE Signal Processing Letters*, 24(4):510–514, 2017.

[203] Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[204] N. Weber. Subseasonal Prediction over the Western United States. In *Amer. Geophys. Union Fall Meeting Abstracts*, 2015.

[205] N. J. Weber and C. F. Mass. Evaluating cfsv2 subseasonal forecast skill with an emphasis on tropical convection. *Mon. Wea. Rev.*, 145(9):3795–3815, 2017.

[206] P. Welch. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

[207] P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In *System modeling and optimization*, pages 762–770. Springer, 1982.

[208] M. C. Wheeler and H. H. Hendon. An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction. *Mon. Wea. Rev.*, 132(8):1917–1932, 2004.

[209] M. C. Wheeler, H. Zhu, A. H. Sobel, D. Hudson, and F. Vitart. Seamless precipitation prediction skill comparison between two global models. *Quart. J. Roy. Meteor. Soc.*, 143(702):374–383, 2017.

[210] J. S. Whitaker and K. M. Weickmann. Subseasonal variations of tropical convection and week-2 prediction of wintertime western North American rainfall. *J. Climate*, 14(15):3279–3288, 2001.

[211] C. J. White, H. Carlsen, A. W. Robertson, R. J. Klein, J. K. Lazo, A. Kumar, F. Vitart, E. Coughlan de Perez, A. J. Ray, V. Murray, et al. Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, 2017.

[212] S. Wiewel, M. Becher, and N. Thuerey. Latent-space physics: Towards learning the temporal evolution of fluid flow. *arXiv preprint arXiv:1802.10123*, 2018.

[213] R. L. Wilby and T. Wigley. Precipitation predictors for downscaling: observed and general circulation model relationships. *International Journal of Climatology*, 20(6):641–661, 2000.

[214] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences (International Geophysics Series; V. 91)*. Academic Press, 2011.

[215] Wolfram. Mathematica, Version 11.3, 2018. Champaign, IL.

[216] N. Wood, A. Staniforth, A. White, T. Allen, M. Diamantakis, M. Gross, T. Melvin, C. Smith, S. Vosper, M. Zerroukat, et al. An inherently mass-conserving semi-implicit semi-Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Quart. J. Roy. Meteor. Soc.*, 140(682):1505–1520, 2014.

[217] T. Wu, L. Song, W. Li, Z. Wang, H. Zhang, X. Xin, Y. Zhang, L. Zhang, J. Li, F. Wu, et al. An overview of BCC climate system model development and application for climate change studies. *Journal of Meteorological Research*, 28(1):34–56, 2014.

[218] P. Xie, M. Chen, and W. Shi. Cpc unified gauge-based analysis of global daily precipitation. In *Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc*, volume 2, 2010.

[219] P. Xie, M. Chen, S. Yang, A. Yatagai, T. Hayasaka, Y. Fukushima, and C. Liu. A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeor.*, 8(3):607–626, 2007.

[220] Y. Xie, E. Franz, M. Chu, and N. Thuerey. tempogan: A temporally coherent, volumetric gan for super-resolution fluid flow. *arXiv preprint arXiv:1801.09710*, 2018.

[221] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, pages 802–810, 2015.

[222] B. Yarnal and H. F. Diaz. Relationships between extremes of the Southern oscillation and the winter climate of the Anglo-American Pacific Coast. *Int. J. Climatol.*, 6(2):197–219, 1986.

[223] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[224] B. Yu and K. Kumbier. Three principles of data science: predictability, computability, and stability (pcs). *arXiv preprint arXiv:1901.08152*, 2019.

[225] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[226] H. Zhu, M. C. Wheeler, A. H. Sobel, and D. Hudson. Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Mon. Wea. Rev.*, 142(4):1556–1569, 2014.