Machine Learning and Feature Attribution for Hydroclimate Modeling

By

Shiheng Duan
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

ATMOSPHERIC SCIENCE

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Paul Ullrich

Shuhua Chen

Erwan Monier

Committee in Charge

2022

Contents

Abstract

Artificial intelligence (AI) has been sparked by significant advancements in Graphic Processing Units (GPUs). Machine learning (ML) and deep learning (DL) models have been widely employed for Earth system modeling due to their ability to fit any contiguous functions. Especially for hydro-climate systems, DL models have been adopted to simulate the processes based on our current understandings. Although the accuracy and performance are comparable or even better than process-based models, DL models are often referred to as 'black box' models since people can not intuitively understand why and how the model produces the desired results, necessitating the need for explainable AI and trustworthy AI. Additionally, when generalizing the models to datasets out of the training range, the trust of DL models is lacking since unlike process-based models, DL models do not explicitly satisfy physical constraints, and as a result, they are more likely to generate nonphysical results. In this dissertation, ML and DL models are adopted to simulate and analyze the components of Earth's hydro-climate system, including snowpack, streamflow and precipitation. These studies focus on both accuracy and interpretations from the DL models. As for precipitation, it is analyzed with a feature-based analysis together with ML models to understand the precipitation characteristics and meteorological drivers.

Several DL models are benchmarked with 20 basins in California for streamflow simulations. The model sensitivities with respect to input variables and input time window size reflect the unique streamflow dynamics over the Sierra Nevada basins. Although there are no explicit physical constraints in the DL model, an idealized test proves the mass conservation, providing confidence in future projection analyses. The future projections validate the distinct dynamic features over the Sierra Nevada basins once again.

In the snowpack simulation, three DL models are developed and tested with observational stations across the Western US. The DL models can achieve comparable accuracy compared with a process-based dataset. A permutation-based explainable AI method is applied to understand the importance of each input variable, which highlights the critical roles of precipitation and temperature in snowpack modeling. When DL models are extrapolated to generate gridded snow estimates, the extrapolation problem arises. It is alleviated with a simple transformation to the output variable, and this method is proved to be applicable to all of the DL models that have been examined. Finally this generalized DL model is used to generate climate projections and investigate the response of snowpack with respect to climate change.

The precipitation analysis focuses on the mean precipitation and extreme precipitation events in the North American Monsoon area. The monsoon domain is first identified using a ML model from a gridded precipitation dataset, and it is further delineated into subregions to better represent local precipitation characteristics. A linear orthogonal method is used to decompose the mean precipitation time series and projects it onto various modes. The monsoon ridge and moisture surges along the Gulf of California are present in the first modes, representing the seasonal-background of precipitation, whereas the second modes are more associated with shorter-time phenomena, such as upper-level disturbances and mid-troposphere lows. Feature-based analysis is conducted to reveal the meteorological causes for extreme precipitation. Five synoptic features and one mesoscale feature are examined and assigned as potential drivers for extreme precipitation. Finally, the feature-based analysis are linked to the linear modes, and moisture surges are found to be more connected with the first mode whereas tropical cyclones are more correlated with the second modes, particularly for extreme precipitation events.

Acknowledgments

First and foremost, I would like to express my gratitude to Professor Paul Ullrich, who first introduced me to the topic of atmospheric science and inspired me to investigate machine learning and deep learning applications. His patience and advice are priceless to me. I am also grateful to my dissertation committee members who assisted me in refining and directing my research.

I am also grateful to all of my mentors and collaborators, including but not limited to Dr. David John Gagne from National Center for Atmospheric Research, Dr. Mark Risser and Dr. Alan Rhoades from Lawrence Berkeley National Laboratory. Dr. David Hall and Prof. Grey Nearing, who helped me with specific deep learning models and architectures, are greatly appreciated.

Conducting my PhD study at UC Davis has been a wonderful experience. Further thanks to all the professors, students and staffs in the atmospheric science graduate group. In particular I would like to thank MetroIT for building and maintaining the Tempest computing system. I am also grateful for the Walter and Margaret Milton Graduate Atmospheric Science Award, which helped me purchase a GPU to accelerate deep learning models.

Lastly, I would like to express my deepest gratitude to my beloved parents. They have always been supportive. Without their support and encouragement, I would not be able to complete my studies.

Chapter 1 Introduction

1.1 Introduction of Hydroclimate Modelling

Water cycle is an important component of Earth's climate system. It has influential impacts on Earth's climate and climate change can alter the water cycle. Hydroclimate modelling focuses on the interaction between climate and water cycle processes. It often involves predict and project hydrological variables given certain forcing inputs. This study focuses specifically on the prediction and projection of streamflow and snowpack. Several process-based models have been developed to simulate the hydroclimate system based on the governing equations. For example, Variable Infiltration Capacity model (VIC) is a large-scale semi-distributed model that can simulate various hydrological components including streamflow, snowpack and soil moisture (Liang et al., 1994). The Weather Research and Forecasting Model Hydrological modeling system (WRF-Hydro) is a land-atmosphere coupling framework that can be used for a variety of studies from flash flood prediction to water resource seasonal forecast (Gochis et al., 2020). These process-based models typically use meteorological variables as the forcing data, such as precipitation, temperature, vapor pressure, and radiation. They also require certain static features (less time-variant) like albedo, leaf area index, and vegetation canopy cover fraction.

Despite these process-based models, efforts are being made to develop data-driven models since biases in the model simulation arise either from the governing equations or parameterizations, which are constrained by our understanding of the hydroclimate system. Additionally, it takes a lot of computational power to solve the governing equations, and this cost grows dramatically with both spatial and temporal resolutions. The data-driven models don't have explicit physical constraints. They are trained directly with data from observations or experiments. The framework of data-driven models are similar as the process-based models: meteorological forcing variables and necessary static features are used as inputs. Hydrological variables are the corresponding targets.

This framework is applied in this study to predict streamflow and snowpack. With forcing data obtained from climate model simulations, projections are also simple to generate within the same framework.

In addition to prediction and projection, cluster analysis is an important tool to understand the hydroclimate system, particularly for precipitation given its spatial-temporal heterogeneity. It is natural to use data-driven models for clustering tasks due to their abilities to aggregate high-dimensional data. In this study, we focus on the North American Monsoon area and its precipitation characteristics. A ML model is used to identify and delineate the North American Monsoon domain, which enables subsequent feature-based analysis.

1.2 Introduction of Machine Learning Models

Artificial intelligence (AI) is a broad area of computer science. It refers to a system with intelligence as opposed to natural objects like animals. Machine learning (ML) represents algorithms and methods that learn how to perform tasks from data. There are three categories in machine learning models: supervised learning, unsupervised learning and reinforcement learning. Supervised learning is a mapping from input variables to output targets (e.g. linear regression). When the target is a real-valued vector, it is called regression, and when it is categorical label, it is called classification. Unlike the paired input-output mapping, unsupervised learning is to find the patterns in the data without any labels (e.g. self-organizing maps). In reinforcement learning, it will optimize a reward function given the environment and actions of an agent (e.g. AlphaGo Zero).

Deep learning particularly represents the models using the artificial neural network (ANN) architecture. There has been different model architectures for various tasks. For example, recurrent neural networks (RNN) are designed for time-series tasks, and 2-dimensional convolutional neural networks (CNN) are backbones for multiple image-related models. In this research, we mainly tackled 1-dimensional time series tasks. Here is a brief introduction of the related ML and DL model architectures.

1.2.1 Self-Organizing Maps

Self organizing map (SOM) is an unsupervised ML model that takes high-dimensional data as input and creates spatially organized internal representations of input vectors. A typical SOM has two layers: an input layer, which takes input vectors, and a competitive layer, where a specific number of nodes are aligned in a topological structure. This topology represents the similarity among different clusters. Similar clusters are located in nearby nodes and further distance represents more distinct clusters. The SOM is optimized by competitive training. When the training sample is fed into the network, the similarity between the sample and the weight of each node is calculated and the best matching unit (BMU) is the node with the maximum similarity. Then the weights of the BMU and the neighbor nodes, which are determined by a neighbor function, are updated. After several iterations, each training sample has been assigned to a corresponding BMU, which is the resulting cluster label.

1.2.2 Artificial Neural Network

Artificial neural networks (ANNs), sometimes referred to as feed-forward neural networks or multi-layer perceptrons (MLPs), are fully-connected neural networks that all the nodes in one layer are connected with nodes in the next layer. Each layer takes output from previous layer as inputs, and a linear transformation is conducted. The nonlinearity is injected by activation functions, such as rectified linear functions (ReLU) or sigmoid functions. Remarkably, MLP can be viewed as a universal approximator. With specific hyperparameters, it can fit any smooth function to any accuracy level (Hornik et al., 1989). However, its ability is limited when temporal and/or spatial features are important.

1.2.3 Long-Short Term Meomry

RNN models are popular for time series tasks. For a given input at time step t, the model will generate a corresponding output, and also a hidden state at time step t, which is used as the input for time step $t+1$. This recursive structure can 'memorize' information from previous steps, which is favorable for time series tasks. However, the recursive computation poses challenges when calculating the gradient. During back-propagation, the gradients of loss function with respect to model weights are calculated for each time step. The gradients can be extremely small or large, especially with a long input time window size. This problem is referred to as 'gradient vanishing' or 'gradient vanishing'. To alleviate this situation with long input sequences, several gates are designed in the RNN models to control the information flow. One of the most popular gated RNN is long-short term memory (LSTM) model. In the LSTM model, there are three gates: input gate, forget gate and output gate. The following equations describe a general LSTM cell:

$$f_t = \sigma(U_f h_{t-1} + W_f x_t + b_f) \tag{1.1}$$

$$i_t = \sigma(U_i h_{t-1} + W_i x_t + b_i) \tag{1.2}$$

$$c'_t = \tanh(U_c h_{t-1} + W_c x_t) \tag{1.3}$$

$$c_t = f_t \circ C_{t-1} + i_t \circ c'_t \tag{1.4}$$

$$o_t = \sigma(U_o h_{t-1} + W_o x_t + b_o) \tag{1.5}$$

$$h_t = o_t \circ \tanh(c_t) \tag{1.6}$$

where t represents the time step, c the cell memory state and h the cell activation or cell output. W and U are weights and b stands for bias. f, i and o are the forget, input and output gates, respectively. The forget gate controls the fraction of stored information to be dropped at current time step, input gate controls how much information from input is stored and output gate determines the fraction of past and current information in the cell output. $\circ$ is the Hadamard product operation or element-wise product. σ (sigmoid function) and tanh (hyperbolic tangent function) are activation functions in LSTM cells. The gated design of LSTM enables it to keep and drop information from the previous time steps, which is naturally a desired architecture for time series tasks. A detailed figure representing the gates and outputs is depicted in Figure 1.1. Besides the LSTM model, there is another type of gated RNN called Gated Recurrent Unit (GRU). Unlike the LSTM model with three gated, there are only two gates in the GRU model.

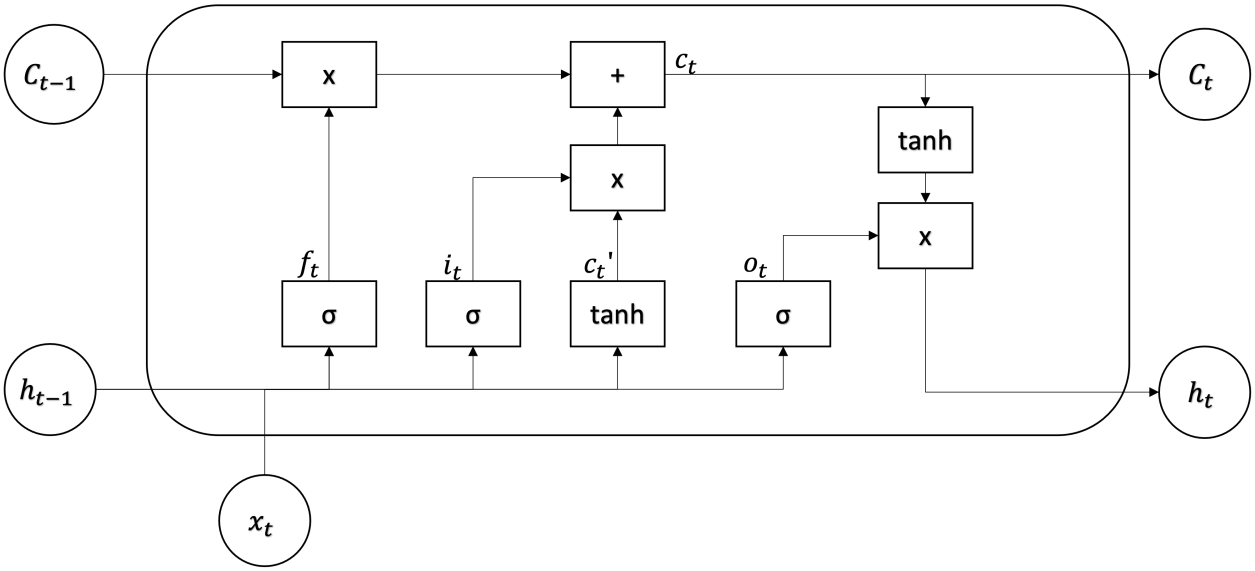

Figure 1.1: LSTM cell architecture. σ represents the sigmoid function. x with subscript denotes the input to the cell, h is the cell activation and c is the cell memory state. i, f, o represent the output after input, forget and output gates respectively. The letter 'x' denotes the element-wise product.

1.2.4 Temporal Convolutional Neural Network

The inherent recursive and gated architectures are desirable to handle the time dependency, while the computational cost is high. The calculations of cell states and outputs at time step t have to wait until the completion of time step $t - 1$, since the cell states from previous time step are required. There has been research indicating the convolution-based architectures can achieve similar performance as recurrent networks (Bai et al., 2018). This type of temporal convolutional neural network (TCNN) utilizes dilated causal convolution operations and residual connection. The dilated convolution expands the receptive field by skipping consecutive inputs. The causal convolution constrains each time frame to only connect with previous time steps, following the causation relationship. Since the convolution kernel only gets limited input frames even with dilation, to capture a long-term dependency, there will be multiple stacked TCNN layers to cover the whole input time window. With such deep architectures, the gradient will probably vanish or explode, which is a similar situation as RNN models. To help the training of deep CNN models, there is a residual connection after two convolutional layers, which provides a shortcut for the

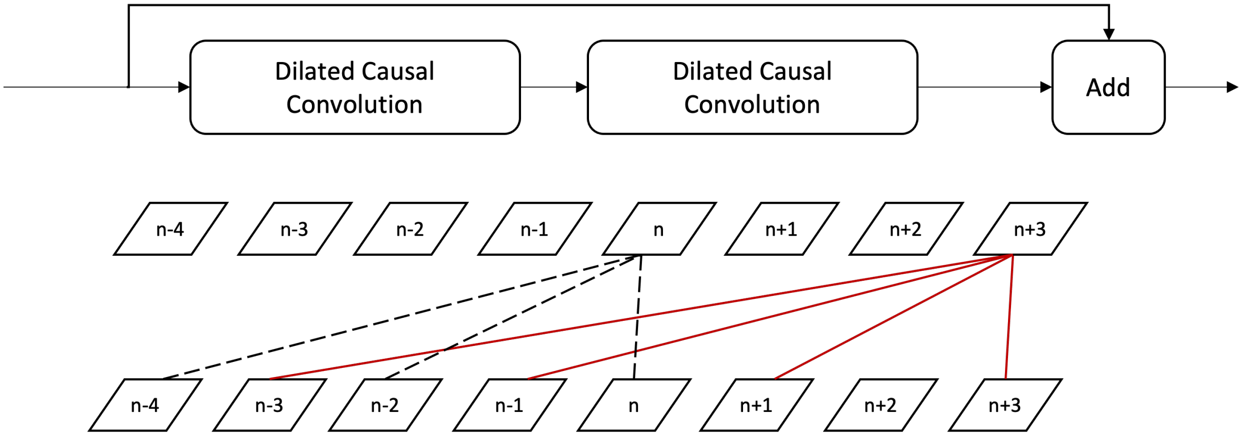

Figure 1.2: TCNN model architecture. The upper panel shows the residual connection and lower panel depicts the dilated causal convolution. Discontinuous input time frames represent the dilated convolutional kernels. Each time step is not allowed to be connected to future steps, assuring the causal relationship.

gradient and has been proven effective in various applications (He et al., 2016). The advantage of TCNN model is that the convolution kernels in the same layer are independent. Without the recursive feature of RNN models, the computation can be paralleled for TCNN models.

1.2.5 Self-Attention Model (Transformer)

Vaswani et al. (2017) introduced Transformer, a self-attention based encoder-decoder model for Natural Language Processing (NLP), and since then, lots of self-attention based models have been designed and investigated for time series problems. In the encoder portion of Transformer, the input vectors are embedded by a dense layer (also called embedding layer). The self-attention layer takes the embedded inputs and extracts the temporal features, which are then used as input for the decoder part. In this study, only the encoder part is used. Figure 1.3 depicts the principle for self-attention. It can be viewed as a fully connected layer but with dynamical weights representing the pairwise relationships of the input time steps (Lin et al., 2021). The detailed formula is shown below:

$$q^i = W_q a^i \tag{1.7}$$

$$k^i = W_k a^i \tag{1.8}$$

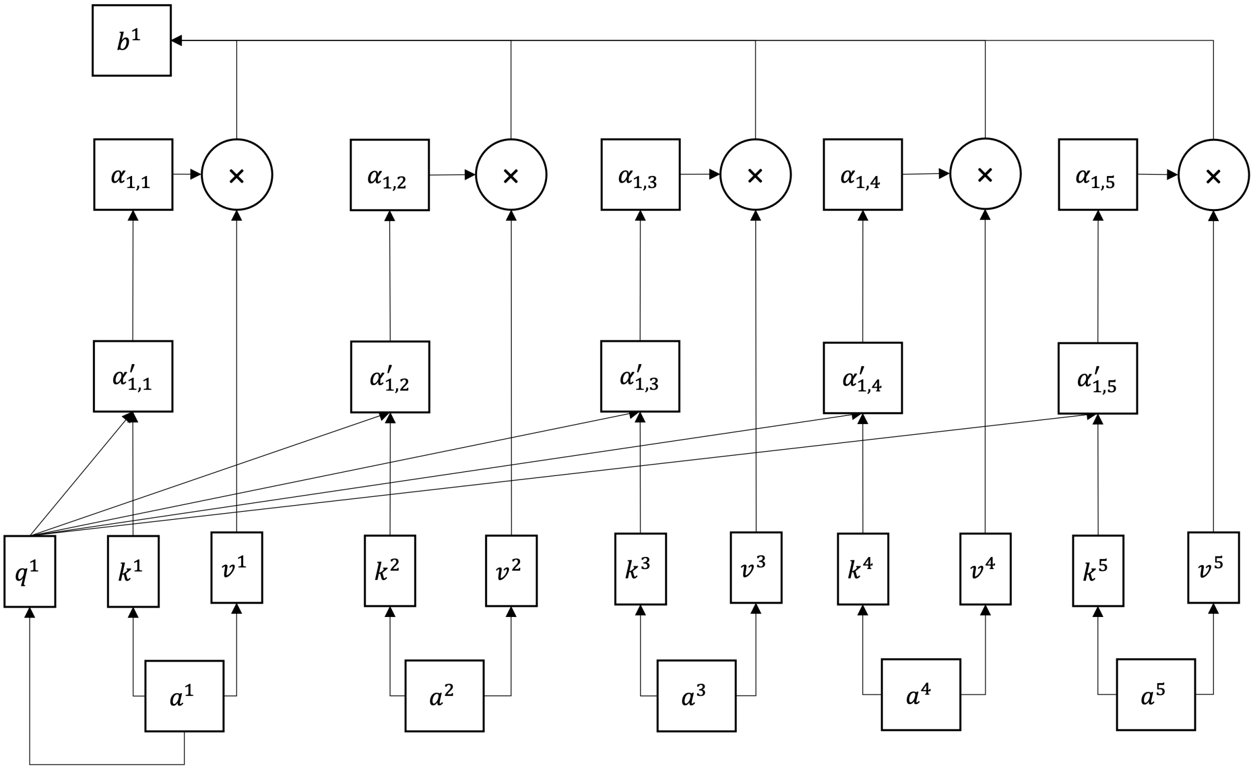

Figure 1.3: Self-attention diagram. a represents the input vector and superscript stands for time steps. q, k, and v are query, key and value respectively. b is the output of self-attention layer.

$$v^i = W_v a^i \tag{1.9}$$

$$\alpha'_{i,j} = \langle k^j a^i \rangle \tag{1.10}$$

$$\alpha_{i,j} = \frac{e^{\alpha'_{i,j}}}{\sum_j e^{\alpha'_{i,j}}} \tag{1.11}$$

$$b^i = \sum(\alpha_{i,j} v^j) \tag{1.12}$$

In these equations, ij denotes the time steps, a is in input to the self-attention. W_q, W_k and W_v are weights to calculate query q, key k and value v. α' is the dot product of key and query and after normalized by the softmax function, α represents the attention score. The sum of value v, weighted by attention scores, is the output b for the attention layer.

Similar with the multiple kernels in the CNN model, the Transformer encode layer also uses several independent self-attention, which is referred to as multi-head, to capture different features

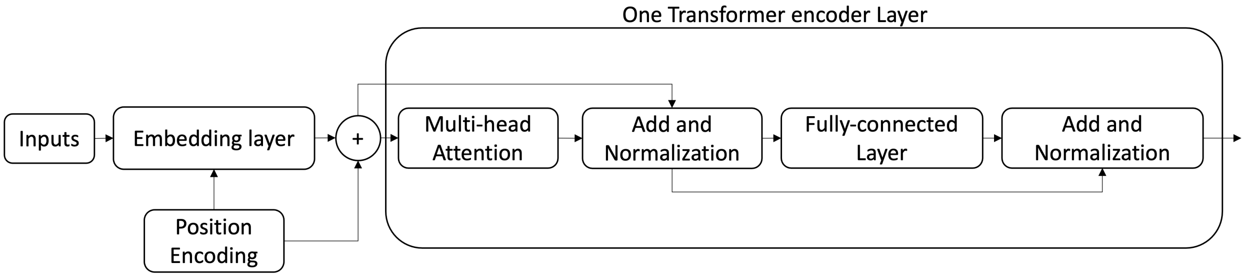

Figure 1.4: Transformer encoder layer architecture.

from the input sequence. There will be a fully-connected layer (feedforward layer) after the self-attention layer to form a complete encoder layer.

To be noted, when calculating q, k, v in the self-attention layers, there is no information about the vector locations in the input time series unlike the inherent input order in LSTM and causal convolution in TCNN. So there is a positional encoder to inject location information into the model. There are various methods to encode positions and we adopted the formula from Vaswani et al. (2017):

$$\text{Encoded}_{x,2i} = \sin \frac{x}{5000^{2i/d}} \tag{1.13}$$

$$\text{Encoded}_{x,2i+1} = \cos \frac{x}{5000^{2i/d}} \tag{1.14}$$

x represents the time steps in the input series, i stands for the ith dimension of the input vector and d for the total number of input dimensions. This location encoding is added to the embedded input time series, as shown in Figure 1.4.

Chapter 2 Streamflow Prediction and Projection in California

2.1 Background

Streamflow is an undeniably important hydrologic quantity for agriculture, society and ecosystems. While historical records of streamflow have been indispensable in informing us of the probability associated with particular flow conditions, it is unclear to what degree these predictions are valid under future meteorological conditions in light of climate change. Failure to correctly predict reservoir inputs has the potential to lead to reservoir failure, such as was witnessed recently with the Oroville resrvoir spillway collapse (White et al., 2019). Long-term projections of streamflow that capture the climatology of streamflow within each watershed are further useful for informing water management strategy. Models for streamflow prediction and projection can be generally divided into two categories: physically-based models and data-driven models (Shen, 2018). Since physically-based hydrological models typically require significant computational expense and extensive calibration of land surface characteristics, machine learning (ML) models are being increasingly employed for streamflow prediction, especially Artificial Neural Networks (ANNs) (Atieh et al., 2017; Gao et al., 2010; Noori and Kalin, 2016; Peng et al., 2017), Support Vector Machines (SVMs) (Huang et al., 2014; Kisi and Cimen, 2011) and recurrent netwokrs like Long-Short Term Memory (LSTM) (Feng et al., 2019; Kratzert et al., 2019a; Le et al., 2019; Yan et al., 2019).

Instead of directly simulating physical processes, ML models mimic the physical rules from historical datasets to develop a functional relationship between inputs and outputs. The learning process largely consists of repeated matrix algebra to adjust the weights in the models, which makes it amenable to acceleration by graphics processing units (GPUs). Further, because ML is broadly applicable across a variety of industries and fields, significant investments have been made in the software supporting its use. Compared with physically-based models, ML models are generally faster to train and can operate with essentially any predictors (Kratzert et al., 2019a). However, the structure of the model and predictor selection are important since they determine the model performance. The general principle governing these models is to build a simple, easy to train model with all the necessary predictors – while avoiding redundant predictors – and ensuring the relationships being clear and direct.

Significant research on ML data-driven models for streamflow has been directed towards data preprocessing, with the purpose being to reduce the number of degrees of freedom in the input dataset and so make any underlying patterns or relationships easier to be identified by ML algorithms. Streamflow at a single gauge station is a fairly traditional 1D time-series dataset, but one that is composed of different components at a variety of frequencies. Consequently Kisi and Cimen (2011) used the discrete wavelet transform (DWT) with SVM for monthly streamflow prediction. The DWT was used to decompose streamflow into high-frequency and low-frequency components, referred to as the "details" and "approximation" in their study, respectively. The approximation, which is the low-frequency component, acts as the baseflow while other high-frequency details represent the variation with shorter period. Their results demonstrated preprocessing with DWT increased the prediction accuracy compared with a model leveraging the raw series. Analogously, Peng et al. (2017) employed the empirical wavelet transform (EWT). Unlike DWT, the EWT decomposition consisted of only three modes, which were used for an ANN model and a residual component. Huang et al. (2014) introduced the empirical mode decomposition (EMD) method for streamflow preprocessing. They decomposed the original data into five intrinsic mode functions and a residual. Instead of removing the residual, they retained it and excluded the high-frequency intrinsic mode function, producing better performance compared with the model using only the original data series. Although these preprocessing steps can simplify the streamflow series and increase performance, they also introduce additional hyperparameters and uncertainty into the model which may impact model robustness.

ML model research has also focused on limiting the choice of predictors – including both input variables and time window size – so as to reduce the number of inputs (Rasouli et al., 2012). Since traditional ML models do not generally incorporate comprehensive physical relationships, ML model developers can focus on only the predictors that explain the most output variability. For streamflow, the most common predictors are precipitation (P) and streamflow (Q) over some historical time period. However, other predictors have been explored as well, informed by our understanding of the system's physical drivers; for instance, Rasouli et al. (2012) investigated several climate indices as predictors, and demonstrated that these can be beneficial for prediction with long lead times up to 7 days. If one only uses precipitation and historical streamflow as predictors, the 1-day lag streamflow prediction problem can be expressed as: Identify a function f so that the predicted daily time series

$$\hat{Q}_t = f(P_{t-N}, P_{t-N+1}, \ldots, P_t; Q_{t-N}, Q_{t-N+1}, \ldots, Q_{t-1}) \tag{2.1}$$

satisfies $\hat{Q}_t \approx Q_t$ (measured under some prescribed metric). Here the subscript represents the time index and N represents the number of historical time points used for prediction. N must typically be large enough to incorporate all historical information relevant to prediction of streamflow at present, but large values of N can lead to increased model complexity which can in turn reduce performance. The value of N is thus usually decided by calculating the autocorrelation or partial correlation; Yaseen et al. (2016) used this approach for monthly streamflow prediction, eventually deciding on a time lag of five months.

A common feature in early data-driven streamflow prediction models is that the input variables were independent of time when fed into the ANNs or SVMs. For example, there were no connections within each layer of dense ANNs, and consequently the network could not "remember" past states. Under such architectures, temporal features in the predictors that may be vital for time series prediction might be neglected. To deal with this problem, some recurrent ML models have been adapted to recognize time dependent features (Le et al., 2019; Yan et al., 2019). Among such models, the most commonly used network (at present) is the LSTM. Kratzert et al. (2019a) used LSTM and Catchment Attributes for Large-Sample Studies dataset (CAMELS) to predict streamflow over CONUS. Their results demonstrated that the LSTM model is capable of extracting temporal features and the results from the ML model can then be used to interpret the physical characteristics of different basins. Feng et al. (2019) added the previous flow rate as data integration, which improves the prediction accuracy of LSTM model. They also employed a convolution data integration method, although the resulting model did not outperform feeding observations directly into LSTM model.

Although there are many ML prediction models, not all can be directly employed for long-term projection. Under future climate change scenarios driven by increased greenhouse gas concentrations, the U.S. West is expected to experience more precipitation and higher surface temperature (Huang and Ullrich, 2017; Ullrich et al., 2018). It is similarly expected that the resultant streamflow patterns will also change. In ML models, since the model is developed and trained with a prescribed training dataset, it is generally expected that the target variable is the same in both training and testing sets. In the real world, however, the statistical properties of the target variable may be changing in time (for instance, under climate change). Under such scenarios, the prediction model may be inconsistent with future projection data, a problem referred to as concept drift (Tsymbal, 2004). Although streamflow can be used in a predictive model framework such as equation 2.1 (an initial-boundary value problem), a simple substitution of $\hat{Q}$ for Q to produce a projection model can lead to errors in streamflow that accumulate over time, potentially biasing the projection.Consequently, projection models must be more heavily constrained to external forcing data, which can restrict the selection of ML model. In the context of projection, Koirala et al. (2014) used the Catchment-based Macro-scale Floodplain Model (CaMa-Flood) with runoff from CMIP5 models as input to derive streamflow under different climate scenarios. Gao et al. (2010) used an ANN and ECHAM5/ MPI-OM model output to derive monthly projection for Huaihe River Basin. These studies demonstrated the potential for ML in streamflow projection.

In this work, a ML-based modeling system for estimating future daily streamflow in California under climate change is developed and validated. After intercomparison among various ML models, a general Temporal Convolutional Neural Network (TCNN) is selected as our candidate system. Although CNNs have not been typically employed for streamflow prediction and projection – being more widely known for image processing – recent work has shown that they exhibit comparable performance to recurrent networks for time series problems (Bai et al., 2018). Consequently our study aims to further establish that TCNNs are competitive for streamflow forecasting with only atmospheric forcing data. Model sensitivities to input variables and time window size are investigated to develop optimal configurations for each basin. With the ML-based streamflow model in hand, future streamflow projections are constructed through the end of the 21st century using statistically downscaled LOCA meteorology as input. To the best of the our knowledge, this is the first work to assess TCNNs for streamflow projection with only atmospheric forcing data. The comprehensive study of the model's sensitivity to covariates and time window size are further novelties of this study. Although this work identifies a strategy for production of future streamflow projections, future work is needed to validate the methodology against physical constraints and investigate the impacts these changes may convey.

2.2 Data and Models

2.2.1 CAMELS

The Catchment Attributes for Large-Sample Studies (CAMELS) dataset provides the hydrologic data for this study (Newman et al., 2014). The CAMELS dataset contains gauge streamflow data and forcing data for 671 basins that feature minimal human disturbance and at least 20 years of data over CONUS. The forcing data is provided as a basin average from NLDAS, Daymet and Maurer, and includes precipitation, day length, solar radiation, and temperature. The streamflow time series data is obtained from USGS gauge stations. The dataset covers 40 watersheds in California, which we have downselected to the 20 watersheds without missing values for this study. Figure 2.1 shows the location, HUC8 identifier, and name of these watersheds. Based on the location of these watersheds, we divided them into five categories and there are the corresponding abbreviations: NC for Northern California (Basin 11381500, 11451100, 11475560, 11522500 and 11528700), SN for Sierra Nevada (Basin 10343500, 11264500, 11266500 and 11284400), SC for Southern California (Basin 10258500, 10259000 and 10259200), CC for Central Coast (Basin 11141280, 11143000, 11148900, 11224500 and 11253310) and BA for the Bay Area (Basin 11162500, 11176400 and 11180500). In

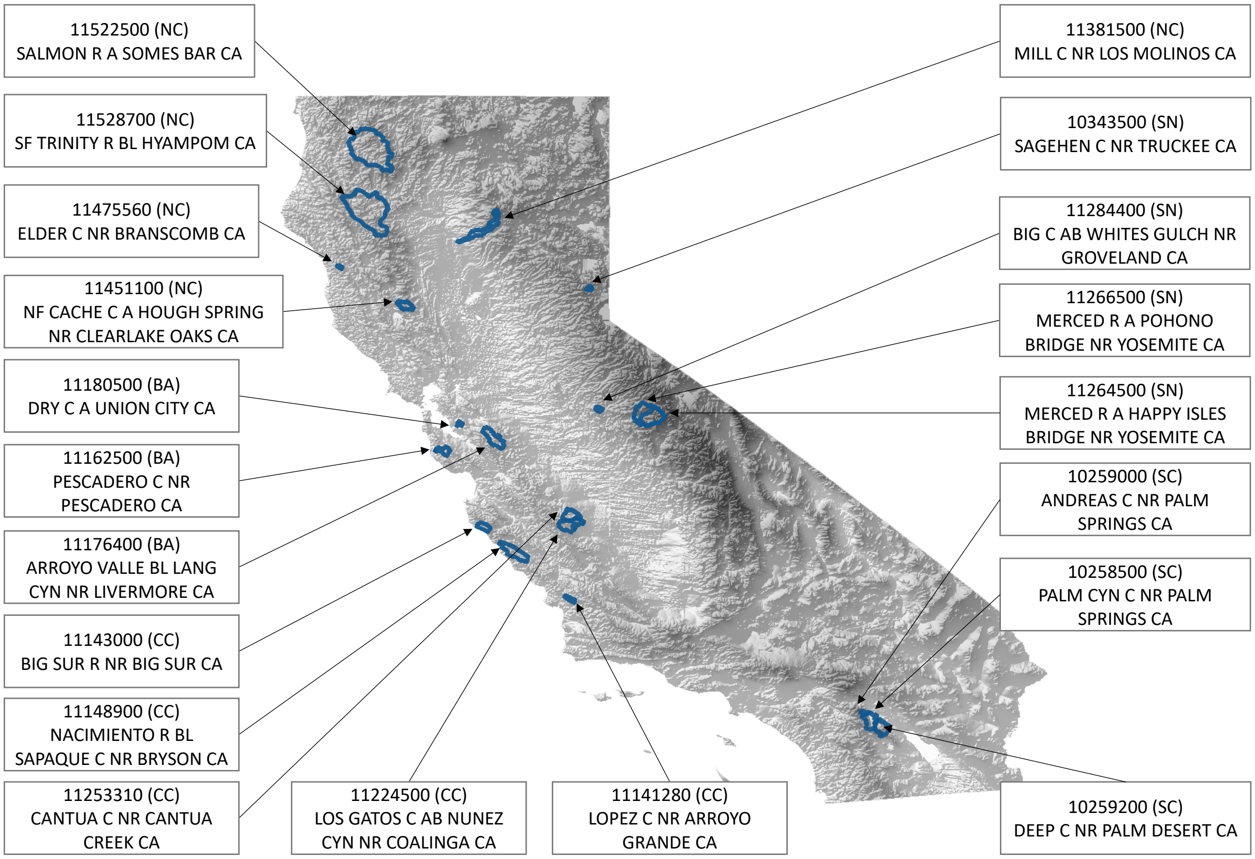
11522500 (NC)
SALMON R A SOMES BAR CA
11528700 (NC)
SF TRINITY R BL HYAMPOM CA
11475560 (NC)
ELDER C NR BRANSCOMB CA
11451100 (NC)
NF CACHE C A HOUGH SPRING NR CLEARLAKE OAKS CA
11180500 (BA)
DRY C A UNION CITY CA
11162500 (BA)
PESCADERO C NR PESCADERO CA
11176400 (BA)
ARROYO VALLE BL LANG CYN NR LIVERMORE CA
11143000 (CC)
BIG SUR R NR BIG SUR CA
11148900 (CC)
NACIMIENTO R BL SAPAQUE C NR BRYSON CA
11253310 (CC)
CANTUA C NR CANTUA CREEK CA
11224500 (CC)
LOS GATOS C AB NUNEZ CYN NR COALINGA CA
11141280 (CC)
LOPEZ C NR ARROYO GRANDE CA
11381500 (NC)
MILL C NR LOS MOLINOS CA
10343500 (SN)
SAGEHEN C NR TRUCKEE CA
11284400 (SN)
BIG C AB WHITES GULCH NR GROVELAND CA
11266500 (SN)
MERCED R A POHONO BRIDGE NR YOSEMITE CA
11264500 (SN)
MERCED R A HAPPY ISLES BRIDGE NR YOSEMITE CA
10259000 (SC)
ANDREAS C NR PALM SPRINGS CA
10258500 (SC)
PALM CYN C NR PALM SPRINGS CA
10259200 (SC)
DEEP C NR PALM DESERT CA

Figure 2.1: A topographic plot of California and the 20 watershed regions considered in this study.

our model, streamflow is normalized by the basin area to avoid discrepancies in the magnitude of the streamflow. The data period is from January 1st, 1980 through December 31st, 2014. In total, we select 10000 daily samples for training (approximately 27 years) and leave the remainder of the dataset for testing. These training samples are consecutive from the beginning of the time series.

2.2.2 LOCA Downscaled Meteorology

For future streamflow projection, the Localized Constructed Analogs (LOCA) dataset (Pierce et al., 2014) is employed. This dataset provides the three necessary input variables for this study, namely precipitation, solar radiation, and near-surface temperature. LOCA is a downscaled dataset ensemble with 6 kilometer resolution over North America from central Mexico through Southern Canada. Among all available LOCA datasets, we downselect four global climate model for this study, which are HadGEM2-ES, CNRM-CM5, CanESM2, and MIROC5 under RCP8.5. These models agree with the four models chosen by California's Climate Action Team Research Working Group as priority models for research contributing to California's Fourth Climate Change Assessment (Pierce et al., 2018). The climatology of these models can be described as warm/dry (HadGEM2-ES), cool/wet (CNRM-CM5), and average (CanESM2). Finally, MIROC5 was selected because it is the most unlike the other three. Since all the basins have irregular shapes, TempestRemap (Ullrich and Taylor, 2015; Ullrich et al., 2016) is used to conservatively regrid the LOCA data to obtain basin-mean forcing data. Because of the uncertainty from both the climate model output and the downscaling process, the historical LOCA data and CAMELS data have some significant disagreements, especially in the values of solar radiation. Specifically, LOCA tends to overestimate the solar radiation compared with NLDAS, (as seen in supplement table S1-S5). To avoid issues related to this systematic difference, the LOCA data was linearly transformed based on the historical forcing data to match the mean and variance of observations. The same transformation was also applied on the projection forcing data. Specifically, for a given daily input X_{LOCA}, either historical or projection, we denote the transformed value as X_{trans}, where μ and σ represent the corresponding mean and standard deviation from the historical period:

$$X_{\mathrm{trans}} = \frac{X_{\mathrm{LOCA}} - \mu_{\mathrm{LOCA\_hist}}}{\sigma_{\mathrm{LOCA\_hist}}} \times \sigma_{\mathrm{NLDAS\_hist}} + \mu_{\mathrm{NLDAS\_hist}} \tag{2.2}$$

(The values of μ and σ for NLDAS and the climate model ensemble can be found in table S1-S5 in supplements.)

Although LOCA provides historical daily atmospheric forcing data, it is not suitable for model training since it is generated from several climate models via an 'analogue-based' statistical downscaling method. The climate models produce a simulated climatology which is only constrained to the real world through prescribed atmospheric greenhouse gas concentrations, so there is effectively no relationship between LOCA and observed gage-based streamflow measurements. This is also the reason why we only analyze the climatology of flow rate in section 2.4, and do not directly compare the time series of streamflow.

2.2.3 Model Predictors and Target

As mentioned earlier, the input variables (predictors) for our streamflow models are precipitation, temperature and solar radiation. By default, the input time window size is set it to 365 days (although this is explored later in the text). In general, the length of the input time window needs to be long enough to capture the relevant physical relationships between input variables and streamflow. For each of our ML models, the target variable is streamflow on the last day in the time window. In other words, our objective is to determine the function f in the following equation:

$$Q_t = f(P_{t-N+1}, P_{t-N+2}, \ldots, P_t; T_{t-N+1}, T_{t-N+2}, \ldots, T_t; S_{t-N+1}, S_{t-N+2}, \ldots, S_t) \qquad (2.3)$$

where Q denotes streamflow, P the precipitation, T the temperature, and S the solar radiation. Note that this equation is only provided for the readers to better understand the relationship between streamflow and the independent quantities. The actual functional relationship will vary based on the model architecture. The subscript denotes the corresponding daily value for that particular quantity, and N denotes the input time window size. The input and output variables are all normalized before feeding them into the models via

$$X_i = \frac{x_i - \mu_x}{\sigma(x)}. \qquad (2.4)$$

Here X_i and x_i are the i^{th} normalized and original variable, and μ and σ stand for mean and standard deviation of that variable. With the normalized variables having zero mean and unit variance, the specific units and range of the inputs will not influence the model. In turn, the normalization procedure is expected to improve the model performance (e.g., (Shanker et al., 1996)).

2.2.4 Machine Learning Models

Four machine learning model architectures (ANN, LSTM, GRU, TCNN) have been investigated and compared with a baseline linear regression model. The detailed architectures are documented in Chapter 1. Stacked LSTM and GRU models are tested to compare against the stacked TCNN architecture. One-layer LSTM and GRU models are also benchmarked as they are common choices from previous studies. For the predictive simulations, model performance is quantified by the Nash-Sutcliffe model efficiency (NSE) coefficient (Nash and Sutcliffe, 1970), which is defined as:

$$\text{NSE} = 1 - \frac{\sum(Q_m^t - Q_o^t)^2}{\sum(Q_o^t - \bar{Q_o})^2} \tag{2.5}$$

where Q_m^t denotes the predicted flow at time t, Q_o^t the observed flow at time t, and $\bar{Q_o}$ the mean observed flow. Here the observed quantities refer to output streamflow from USGS gauge stations. Larger NSE values indicate better performance. Since the NSE is proportional to the square of the difference between model and observations, it tends to put greater emphasis on high flow periods. To maximize NSE, we set $1 - \text{NSE}$ as the loss function for our models – that is, the quantity to be minimized during training process. For each model, training is performed separately on each basin but with the same model architecture.

Before training these networks, we first need to set the hyperparameters, which are tuning factors in the model architectures and training process. Common hyperparameters include the number of layers, optimizer and number of epochs: The number of layers is important to the specific model architecture; the optimizer refers to the gradient descent algorithm used in the training process; and the number of epochs refers to the number of times that the model is trained on the entire training set. The Adam optimizer is used with 0.0005 as the learning rate. We trained each model for 150 epochs with the batch size set to 512. These training configurations are set based on the training loss function, which ensures the loss decreases and stabilizes at a low value. Although the hyperparameters are important for overall model performance (Bergstra and Bengio, 2012), in this work we hold the optimizer and the number of epochs the same for all models. This study does not investigate differences that may arise through more fine tuning of these hyperparameters for specific models – indeed a comprehensive investigation of the optimal hyperparameters for each model is beyond our current computational capability.

2.2.5 Ensemble Runs

Unlike the linear model, which has an exact analytical solution, all the neural networks use gradient-based method to optimize the loss function. Since the networks allow for local minima, different initial weights can potentially produce different models with different performance. Thus one needs to be careful to avoid drawing conclusions on the relative performance of each model that are merely a byproduct of the initial weights. In order to eliminate this effect, we run each model 15 times to get an ensemble distribution of NSE values. Thus our results and conclusions are based on the statistical distribution of model performance across the ensemble.

Throughout this study we make use of boxplots for assessing comparative performance between ensembles. As shown in (Krzywinski and Altman, 2014), comparative performance is intuitive from the boxplot – namely, if the median for one model is above the interquartile range of another, we are confident that it is the better model. However, if the median from the second model lies within the interquartile range of the first model, performance could be the result of randomness in the training process, making it difficult to determine the better model.

2.3 Results

In this section we first compare the various ML models discussed in section 2.2.4 to demonstrate the competitive performance of the TCNN. The TCNN is then examined in light of stability under extreme forcing, its sensitivity to choice of input variables across basins, and sensitivity to time window size. A physical interpretation of the observed model sensitivity is also discussed here.

2.3.1 Model Intercomparison

Figure 2.2 shows the ensemble prediction results for each basin among the four ML models, plus the linear regression model. The linear regression model performs the worst among available models in almost all basins, in testament to the nonlinearity of the prediction problem. The ANN model tends to achieve a higher NSE value than the linear regression model for almost all basins, but in terms of NSE the ANN is still inferior to the recurrent networks and the TCNN, especially for basins where

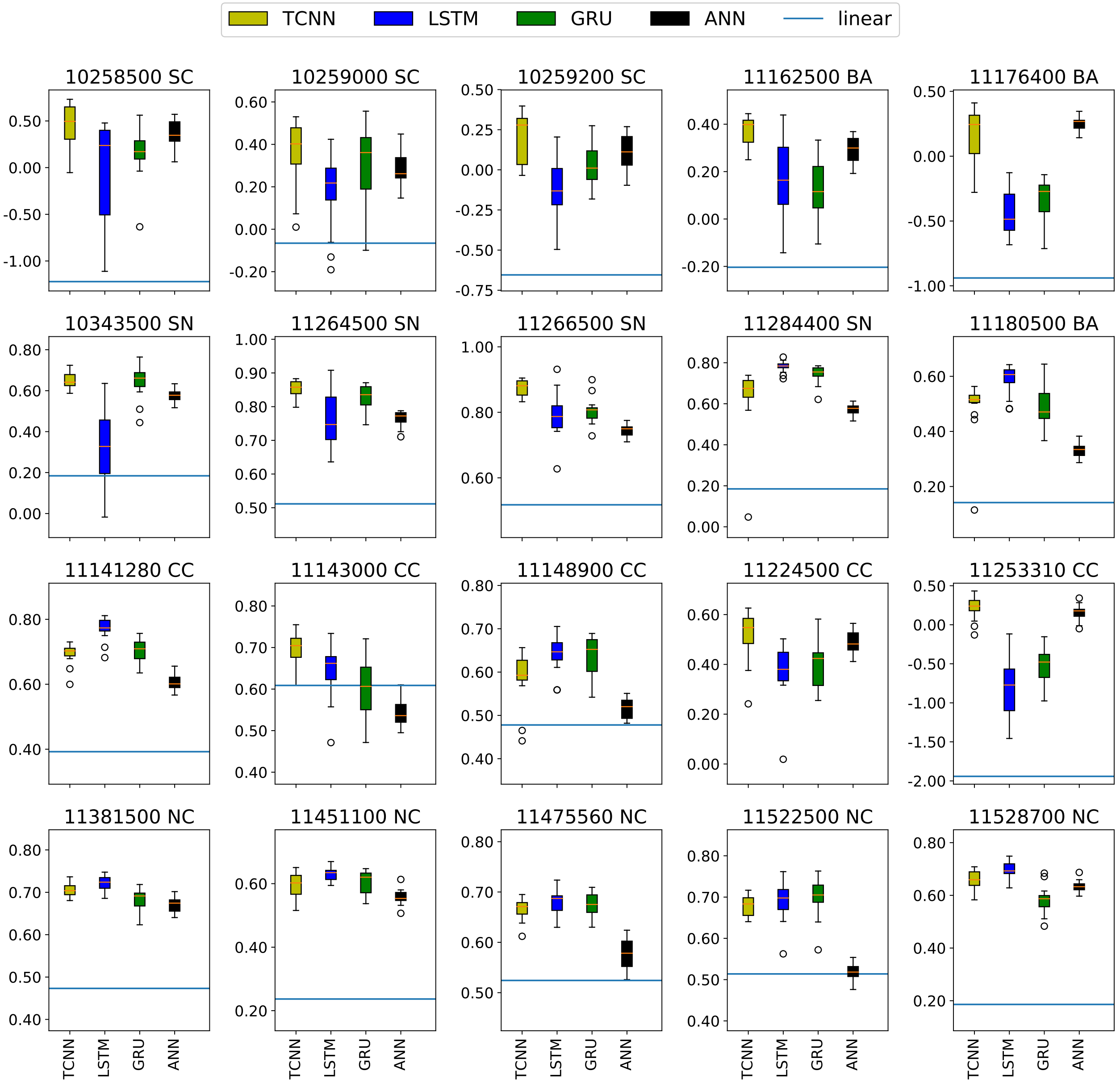

Figure 2.2: Ensemble prediction comparison for all basins with different models. The boxplots denote results over each ensemble of 15 model runs for the ML models. The straight line denotes the linear regression result. For basin 11224500 CC the linear regression model produced a NSE of -2.47.

the NSE values for the recurrent networks and the TCNN are over 0.6, such as 11475560(NC) and 11522500(NC). In these basins, the relatively low NSE scores from the ANN indicate that there are some temporal features that the ANN cannot capture but which are important for streamflow prediction. Nonetheless, for some basins, the ANN outperforms the LSTM. There are two possible reasons this may occur. Firstly, it could be that temporal features are not important for these basins, a hypothesis that is supported by the observation that the ANN tends to also be better than GRU (e.g. basins 11176400(BA) and 11224500(CC)). On the other hand, the TCNN doesn't have a recurrent architecture so it can effectively ignore the temporal features and mimic the ANN. This could suggest that LSTMs may not be as generalizable as TCNNs. Another possible reason is that the LSTM hyperparameter set is suboptimal for these basins – assessing this possibility may require a more comprehensive basin-dependent hyperparameter search.

Among models with temporal features (TCNN, LSTM, GRU), the TCNN exhibits the best average performance. The average NSE over all basins and all ensemble runs is 0.40 for LSTM, 0.44 for GRU and 0.55 for TCNN. The average NSE value for the best run over all basins is 0.58 for LSTM, 0.58 for GRU and 0.65 for TCNN. For those basins where LSTM and GRU achieve the highest NSE values, the performance of the TCNN model is competitive – for example, in basins 11141280(CC) and 11284400(SN) the NSE values among the different neural networks are all higher than 0.5. For basins where neural networks do not perform well, such as basins 10259200(SC), 11176400(BA) and 11253310(CC), the TCNN is nonetheless the best among the different neural networks. Notably, the recurrent networks can achieve high NSE values for some basins while performing poorly for other basins. That is, their performance varies substantially among different basins. The TCNN, however, is more stable among all basins: The standard deviation of the ensembles over all basins is 0.47 for LSTM, 0.38 for GRU and 0.23 for TCNN. Although the choice of hyperparameters is important in these results, the wider spread in the NSE value indicates that for streamflow prediction the recurrent networks have more local minima over the optimization space, and consequently must be trained many times to find a globally optimal configuration.

The stacked recurrent networks here are chosen to compare with the stacked TCNN model. A one-layer LSTM model, such as the one used in Kratzert et al. (2019a), is also investigated. We employ 256 hidden units for the LSTM to match Kratzert et al. (2019a). Similarly a one-layer GRU model with 256 hidden units is also compared. With this configuration, the average NSE over all basins and all ensemble runs does improve to 0.47 for the one-layer LSTM, but degrades to 0.34 for the one-layer GRU. The standard deviation is 0.40 and 0.74 for the one-layer LSTM and GRU, respectively. When comparing the average NSE for the best run, the one-layer LSTM and GRU achieve 0.61 and 0.56, respectively. We again tested another one-layer LSTM model with 370 hidden units so as to match the number of free parameters within the TCNN model. The average NSE over all ensemble runs is 0.44, and 0.58 for the best run. The standard deviation is 0.43. A comprehensive comparison can be found in table 2.1. Figure S1 shows the ensemble prediction comparison of the TCNN model with one-layer recurrent networks. Compared with the one-layer recurrent networks, the TCNN model still exhibits slightly better performance with lower variation.

Table 2.1: Mean and standard deviation for ensemble prediction comparison with different models. The number in parentheses denotes the number of hidden units.

	Mean NSE for all ensemble runs	Mean NSE for the best run	Standard deviation
TCNN	0.55	0.65	0.23
Stacked LSTM	0.40	0.58	0.47
Stacked GRU	0.44	0.58	0.38
One-layer GRU (256)	0.34	0.56	0.74
One-layer LSTM (256)	0.47	0.61	0.40
One-layer LSTM (370)	0.44	0.58	0.43

Besides evaluating the NSE value for the whole prediction period, we also examined the model performance for high flow and low flow days. High flow (low flow) days are defined as days when the observed flow rate is higher (lower) than the 95^{th} (5^{th}) percentile over all days. Since the low flow series for some basins is zero throughout, NSE cannot be used to assess performance. Instead, we use mean squared error (MSE) to quantify the performance, given by

$$\mathrm{MSE} = \sum (Q_m^t - Q_o^t)^2. \tag{2.6}$$

The MSE spread for all basins over the ensemble can be found in Figure S2 and S3. Whereas the TCNN tends to perform well during high flow periods, the LSTM does exhibit better performance in low flow periods. For the high flow period, when average MSE is compared over all the ensemble runs, the TCNN achieves the best performance on 12 basins compared with 4 basins for LSTM. When comparing the minimum MSE for the high flow period, the TCNN is the best model for 10 basins compared with 6 basins for LSTM. For the low flow period, when using the average MSE, TCNN is the best model for 2 basins compared with 10 basins for LSTM; when assessing minimum MSE value, the LSTM is supperior in 16 basins. The reason for this behavior is likely a simple consequence of the chosen hyperparameters of the model; further optimization will likely result in incremental improvements to both the TCNN and LSTM. Notably, the purpose for the comparisons in this section are not to show the TCNN is better than the LSTM, since such a proof would require us to effectively test all possible architectures and hyperparameters. Instead, these results demonstrate that the TCNN can achieve comparable performance to other commonly used models.

In addition to assessing model performance, training time also merits comparison among the different models. The average training time for one basin with the ANN on a single RTX 2080Ti is 11 seconds, 77 seconds for TCNN, 149 seconds for stacked GRU, and 150 seconds for stacked LSTM. For the one-layer LSTM model, it takes 220 seconds for 256 hidden units and 380 seconds for 370 hidden units. Hence, for this particular configuration the TCNN model is the fastest among the models with temporal features, although only by a factor of two.

Based on the results presented in this section, TCNN is chosen as our candidate network for prediction and projection of streamflow. The following section now focusses assessing and explaining the performance of the TCNN.

2.3.2 Model Stability under Extreme Climate Forcing

One of the biggest challenges for ML projection is concept drift (also known as non-stationarity). Under climate change, it is widely accepted that the statistical properties of the input predictors and output streamflow will change through time. Although surface temperatures are expected to increase almost everywhere, in parts of California these increases are also accompanied by an increase in total precipitation of about 1.2% per decade (Ullrich et al., 2018). It is further expected that the input variance will increase in conjunction with more frequent extreme precipitation and temperature events (Swain et al., 2018). However, because the TCNN model is trained on historical data, the end-of-century inputs may incur extrapolation, which has the potential to produce unphysical results such as negative flow. To test whether the TCNN model is able to produce physically reasonable results even when inputs are not within the range of the training data, an idealized test is devised to stress the model far beyond the long-term range of possible inputs. Specifically, the model was executed with quadruple precipitation and a temperature increase of 5 degrees Celsius from the training set. Only one simulation was performed for each basin, using the TCNN model with highest NSE value from the ensemble run.

The extreme scenario investigated here is unrealistic even in light of climate change. However, if the ML model were not stable, this extreme scenario, far outside the realm of the training data, should cause the model to "blow up" or generate negative flow rates. However, if our model can still produce acceptable results under such an extreme scenario, we have greater confidence that it will generate reasonable projection results under the RCP8.5 scenario. Results from a single representative basin are depicted in Figure 2.3. Although only one basin in shown here, the results are analogous in other basins (not shown). As expected, the projected streamflow is generally much larger than historical, with much higher flood peaks. In addition, the high flow period is longer under this test as a result of precipitation accumulation, and low flow periods produce consistently higher streamflow. The regression line from the scatter plot is $Q_p = 4.001 \times Q_h$ ($R^2 = 0.74$), where Q_p is projection streamflow and Q_h is the historical streamflow. Thus the 4× increase in precipitation produces approximately a 4× increase in streamflow. However, this simple linear factor appears to underestimate flows on the low flow days and overestimate flows during high flow days, again indicative of nonlinearity in the streamflow dynamics.

2.3.3 Model Sensitivity to Input Variables

As discussed earlier, the input variables for our full model are precipitation, temperature, and solar radiation. Although input fields beyond precipitation can improve model performance by capturing significant physical relationships, they also increase the complexity of the model, potentially leading to a wider spread among trained models. To test the importance of these variables for streamflow

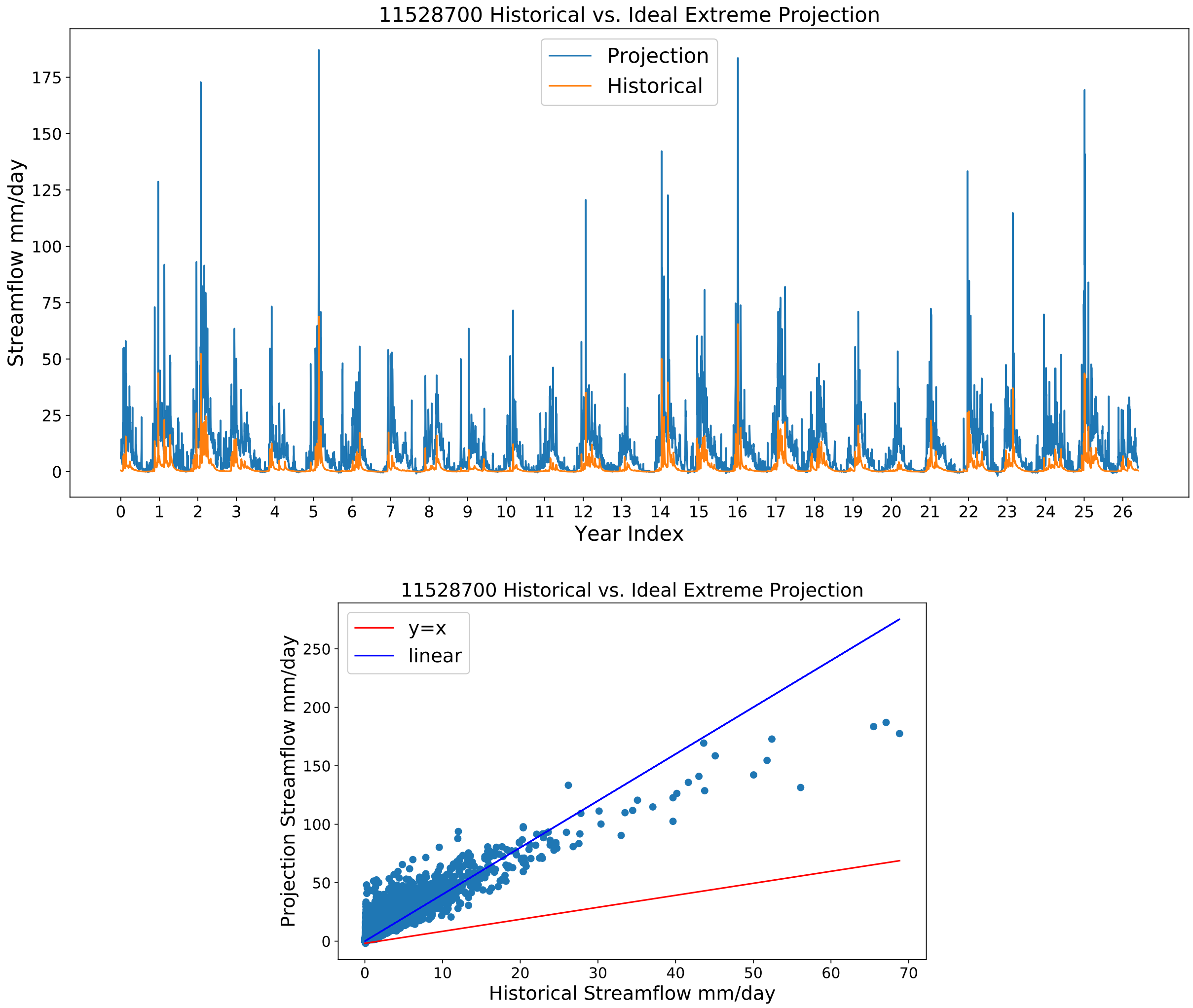

Figure 2.3: A depiction of the streamflow response under idealized extreme forcing showing (top) time series of flow and (bottom) historical streamflow vs. projected streamflow.

prediction three reduced models were compared, consisting of precipitation solely (p), precipitation and temperature (pt), and precipitation and solar radiation (ps). When comparing the performance of reduced models and the full model, the 15-model ensemble was again used to avoid noise from the initial state.

The overall performance of ps and pt models is again assessed using box plots of NSE values. Figure 2.4 shows the result of the ensemble comparison. It is apparent that for some basins temperature boosts predictability, while for others solar radiation is more important. There are only three basins where the best pst model is better than the best ps or pt model (11264500(SN), 11266500(SN), and 11381500(NC)), and in each of these cases the improvement with all three variables is modest. In each basin, the dominant variable does reflect the geographic features of the basin. Basins where temperature significantly improves performance are 10343500 (SN), 11264500 (SN), 11266500 (SN), 11451100 (NC), 11522500 (NC), 11176400 (BA) and 11224500 (CC) which include three Sierra Nevada basins, two Northern California basins, and one Bay Area basin. Basins where solar radiation improves performance are 10258500 (SC), 10259000 (SC), 11143000 (CC), 11253310 (CC), 11180500 (BA), 11284400 (SN), and 11475560 (NC) – except for the last two, these are located in coastal areas or in the inland desert of Southern California. These results suggest that, to a close approximately, we can divide the basins into three categories using these reduced models: those where temperature is important (generally in mountainous regions), those where solar radiation is important (generally near coastlines), and those where temperature and solar radiation offer no significant benefit to predictability.

The physical explanation underlying the performance of the reduced models is related to the climatological properties of these different basins. For instance, in the basins of the Sierra Nevadas and Northern California, accumulation and melt of wintertime snowpack generally plays an important role in driving streamflow. However, the inclusion of temperature in these mountainous regions does not necessarily guarantee a performance improvement. For instance, temperature does not improve the model of 11528700 (NC), where snow is a major driver for streamflow; nonetheless, the inclusion of temperature also does not significantly degrade performance. Further, in basins where temperature improves performance we also generally see that inclusion of solar radiation

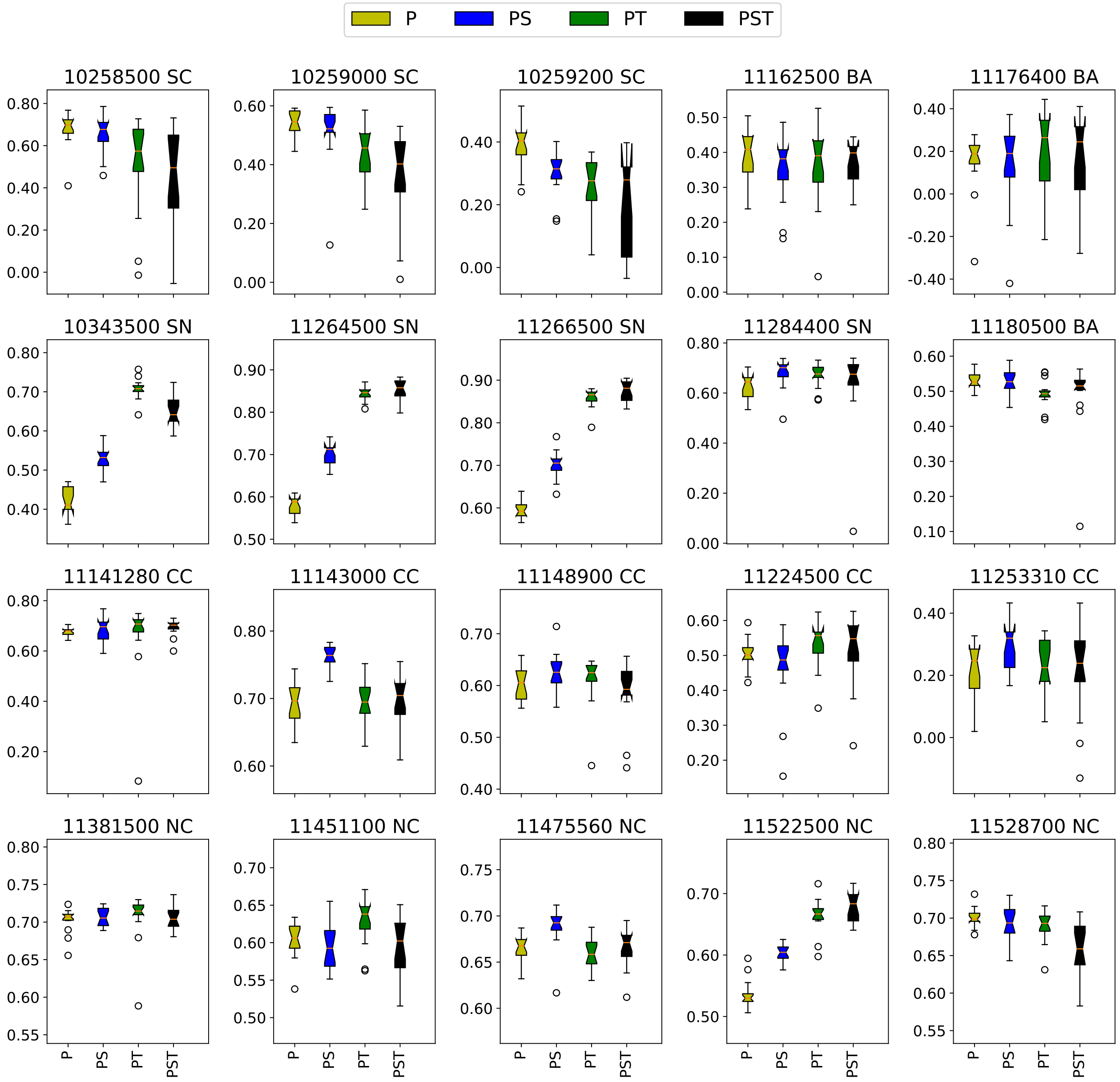

Figure 2.4: Ensemble prediction comparison for all basins with different reduced TCNN models.

does provide some improvement over the models only using precipitation – this suggests that the ML model is potentially identifying the relationship between solar radiation and temperature, or is instead using solar radiation to estimate snow melt rates.

The physical processes driving streamflow in the coastal basins are significantly different than those of the mountains. Namely, coastal basins do not experience significant temperature variations as a result of temperature regulation by the ocean. Further, because the ocean provides a ready source of moisture, air remains close to saturation. In accordance with the Penman-Monteith equation, evaporation from these basins will be driven primarily by radiative forcing, in agreement with our results. Among the central coast basins, the one exception that shows improved performance with temperature, but no significant improvement from solar radiation is 11224500 (CC). Although this basin is on the Central Coast, it is far from the coastline and so subject to larger temperature swings and lower relative humidity. The relatively high-altitude coastal ranges in this basin do produce occasional snow accumulation, but it is unlikely that snow dynamics plays a role here.

For those basins where inclusion of solar radiation and temperature produce worse model performance (i.e. the three Southern California basins), we hypothesize that the ML model is either identifying non-existent physical relationships between these variables and streamflow in the training data, or that the increased model complexity is making it more difficult for the model to converge to an optimal configuration. The truth is likely a combination of both of these factors, as for all three SC basins the "best performing" pst model is not significantly worse than the median p-only model, but is clearly worse than the best p-only model.

In conclusion, the reduced models explored here are helpful for giving insight into the processes that are most relevant for each basin, and thus the relevant causative relationships. Here snowpack dynamics and coastal meteorology have emerged as two obvious geographical features important for determining model behavior. Given this behavior agrees with our physical understanding of the system, we have further evidence to suggest that the models are behaving credibly.

2.3.4 Model Sensitivity to Time Window Size

The input time window size is an important hyperparameter for our model, and one that is intrinsically connected to the physical processes driving streamflow. However, a time window that is too large can reduce model performance and slow training time. Some past studies set the time window size based on the results from a purely statistical analysis of autocorrelation or partial correlation (Peng et al., 2017; Yaseen et al., 2016). In this study, we estimate the time window size from an understanding of the physical properties of each region. For streamflow prediction and projection, the response time for precipitation, groundwater and snowpack can range from several hours to months, and a proper time window size should capture all necessary features and avoid redundant information. The seasonality of the streamflow varies regionally and depends on the climatic characteristics and the contribution of snow/ice, and anthropologenic interventions. An investigation of monthly global steamflow (Dettinger and Diaz, 2000) indicated that lags between the peak precipitation and peak steamflow peaks up to 11 months, while 0-3 months was the typical value. In this study we explore 100, 180 and 365 days as different window sizes. The 365-day window corresponds to an entire water year, and so should capture all potential physical processes except for long-term withdrawals or variations in groundwater. The 100-day window captures a typical season length and the 180-day window is in between these two. Figure 2.5 shows the ensemble performance results comparing models with different time window sizes.

What stands out in Figure 2.5 is the monotonic tendencies in most basins. There are increasing tendencies with the time window size for basins 11224500 (CC), 10343500 (SN), 11264500 (SN), and 11266500 (SN), while 11162500 (BA), 11176400 (BA), 11253310 (CC), 11475560 (NC) and 11528700 (NC) show decreasing tendencies. An increasing tendency implies the presence of slow processes governing streamflow, whereas a decreasing tendency implies upstream processes are fast and there is no significant benefit in using a larger window size. In fact, we can again classify basins into two categories by their monotonic tendencies. Similar with the previous interpretation of different predictors, these results are likely to be related to physical factors, especially snowpack – particularly because of its long response time. In general, the basins with increasing tendencies

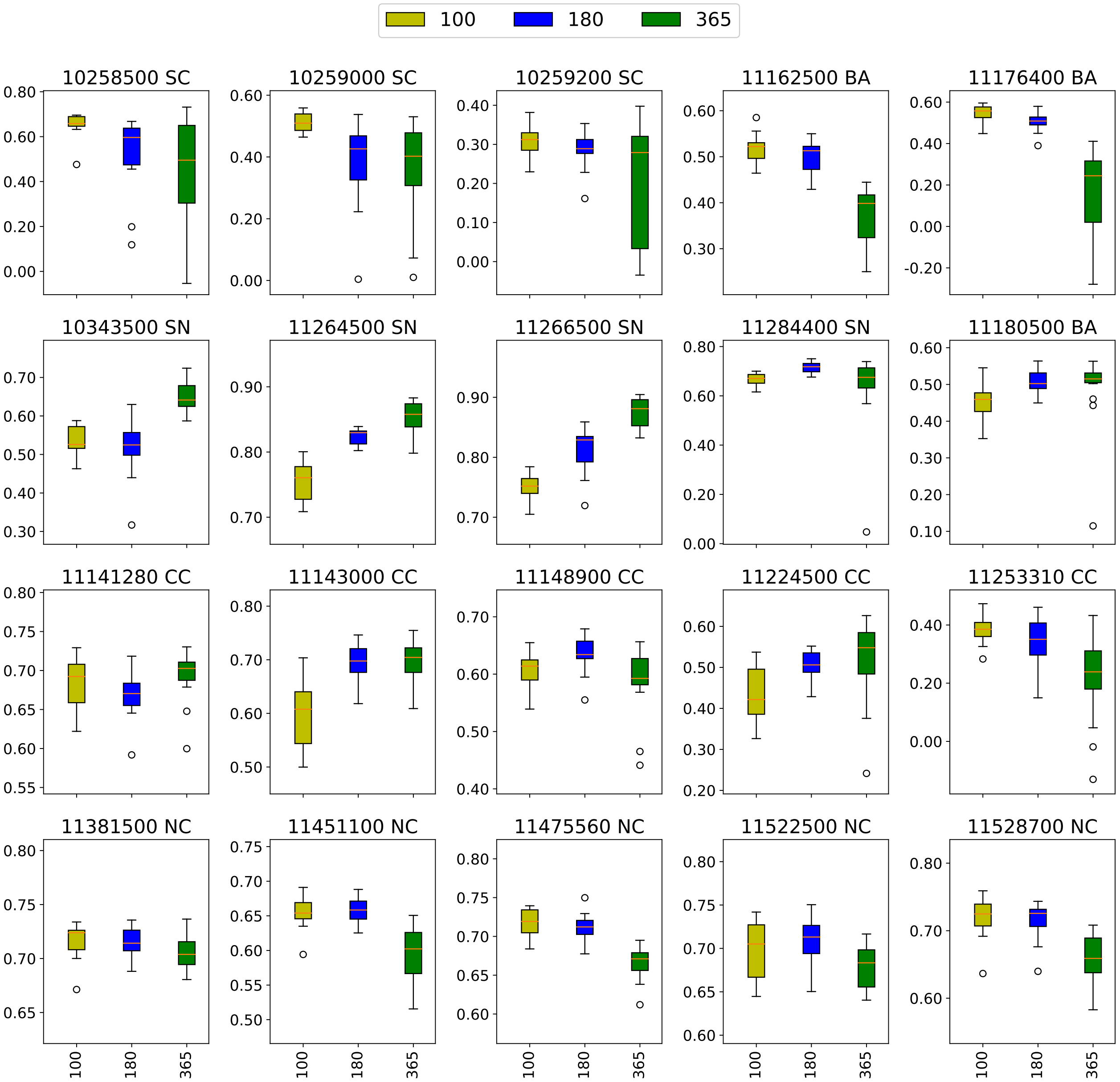

Figure 2.5: Ensemble prediction comparison for all basins with different window sizes.

are in mountainous area like Sierra Nevada and the Coastal Ranges while basins with decreasing tendencies are in Northern California, the Bay Area and the Central Coast, which are closer to the Pacific. Mountainous areas tend to have more snowpack due to their higher elevation and thus streamflow there is more likely influenced by snowpack. For coastal areas, snowpack does not play a role in streamflow dynamics, and since the temperature is more stable relative to inland areas, the impact from snowpack will also be weaker than that in inland basins. Therefore, snowpack should be the primary factor driving the direction of the tendency. Another factor not explored here that may affect the tendency is the groundwater response time – this may play a role in central coast basins such as 11143000 (CC) and 11224500 (CC), which respond positively to increased time window size.

2.4 Projected Streamflow

The best models from the ensemble run for each basin are now employed with remapped and rescaled LOCA data to produce our projection dataset. As described in section 2.2.2, we first apply TempestRemap to obtain mean forcing data for irregular basins from the gridded LOCA product. Then both future and historical forcing from LOCA are rescaled (bias corrected) based on the historical observations before being used to drive the ML model. Table S6, S7 and S8 show the mean daily precipitation, temperature and solar radiation from NLDAS and the four climate models employed. Figures S4, S5 and S6 also show the climatological daily mean of these variables. Generally CanESM2, CNRMCM5 and HadGEM2ES suggest a future wetter climate with more precipitation, while MIROC5 tends to produce similar or less precipitation for these basins. Essentially all the basins are projected to experience higher daily temperatures, but the change in solar radiation is small. Figures 2.6 and 2.7 show climatological daily streamflow with historical and under the future projection with RCP8.5 forcing. The daily streamflow projection dataset produced in this manner is available at (Duan et al., 2020b) with the units of millimeters per day. Within the database, each file has the name as the format of ‘nnnnnnnn-model-scenario.csv.’ The first eight digits are HUC8 identifiers for each basin, followed by the climate model name, and then the scenario (either ‘hist’ or ‘RCP8.5’).

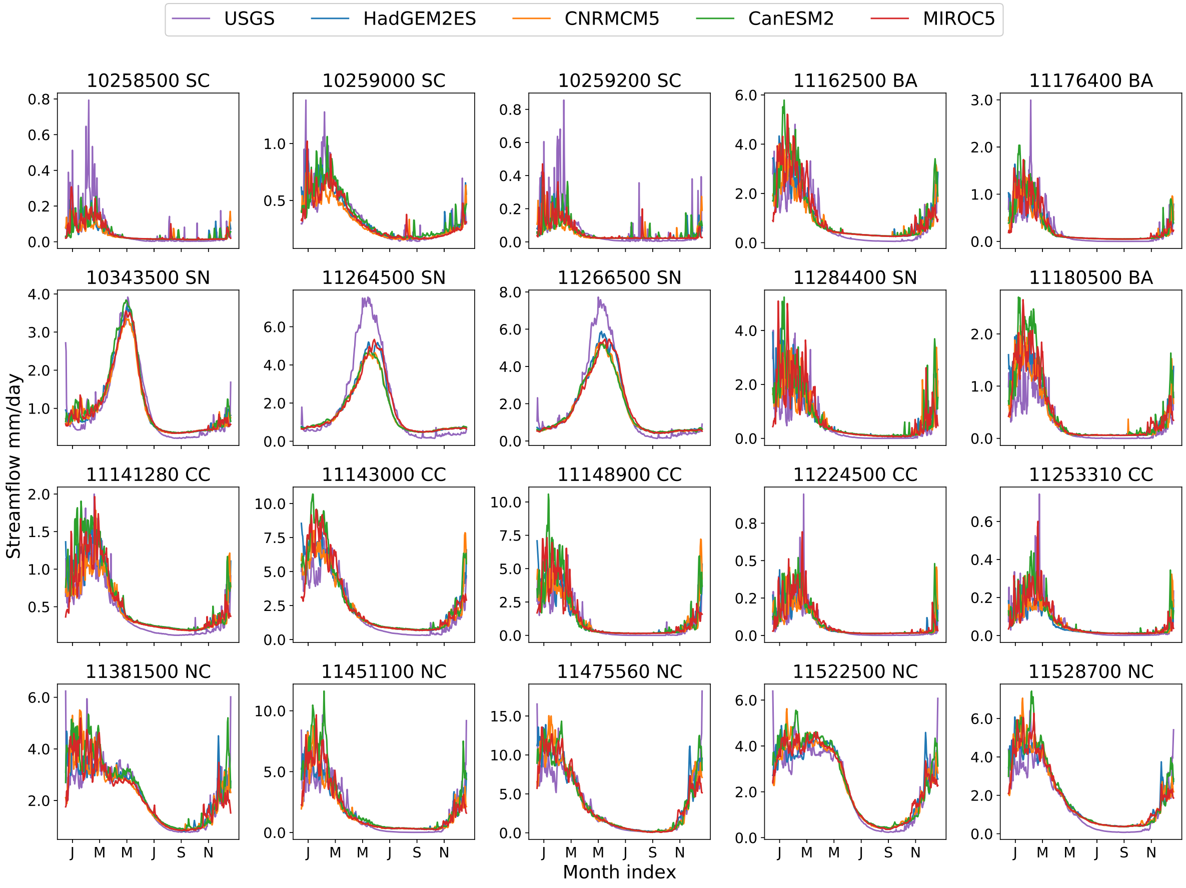

Figure 2.6: Historical climatological daily streamflow from USGS and four climate models.

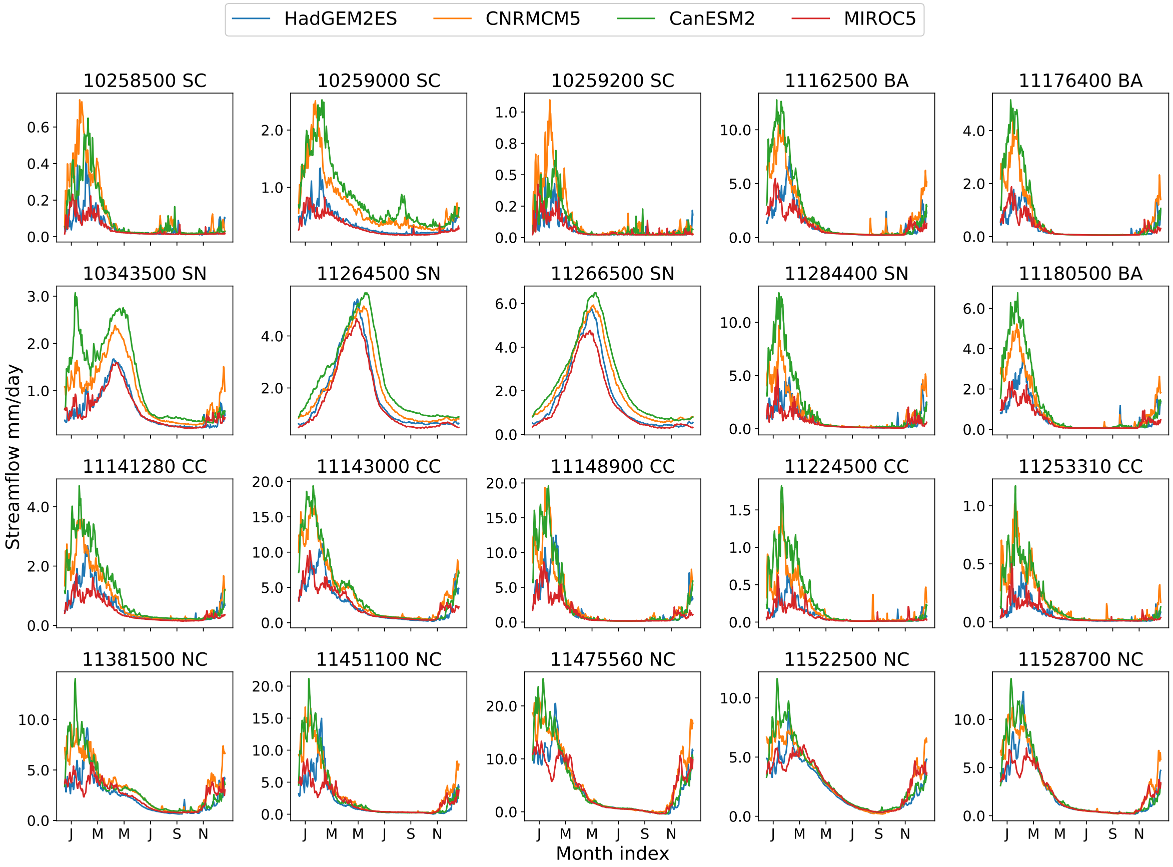

Figure 2.7: Projection climatological daily streamflow from four climate models.

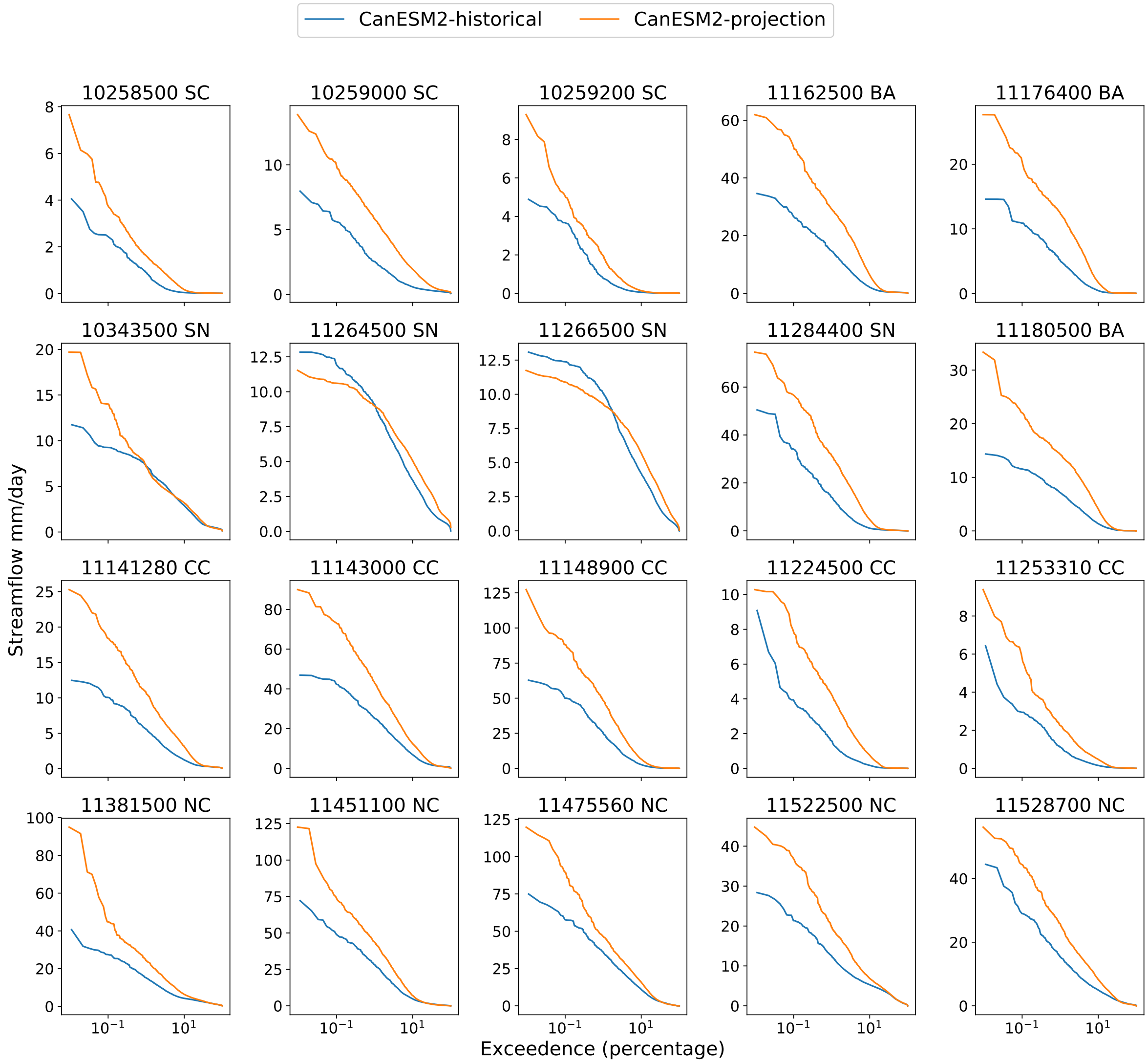

Figure 2.8: Flow duration curve with CanESM2 forcing over both historical and future (projection) periods.

2.4.1 Analysis of the Projected Streamflow

Since the historical forcing from different climate models are corrected to match observations (as discussed in section 2.2.2), historical streamflow exhibits nearly the same pattern and magnitude with forcings from different climate models (Figure 2.6). Compared with USGS observation, the flows tend to match fairly well except in a few SC and SN basins, where a clear magnitude difference at the flow peak emerges. For 10258500(SC) and 10259200(SC), even with the NLDAS forcing data the TCNN underestimates the peak, so we can conclude that the TCNN simply does not identify a relationship between forcing and streamflow during these high flow events. Looking at the SN basins, Figure 2.2 shows that the TCNN model achieves an NSE score around 0.9 for 11264500(SN) and 11266500(SN), so in this case the differences are likely due to differences between the forcing from NLDAS vs LOCA. Namely, we can deduce that for these basins the Gaussian bias correction (2.2) still produces a forcing which is still somewhat inconsistent with historical forcing. For 11264500(SN) and 11266500(SN) the primary source of this error appears to be wintertime and springtime temperatures, which are intimately connected to precipitation phase and snowpack melt rate; when the LOCA temperatures and radiation are replaced with NLDAS temperatures and radiation (while retaining the LOCA precipitation) the correct streamflow curves are recovered (Figure S7).

To assess the magnitude of future change, we examine the projected flow duration curve (FDC) versus the historical FDC from the same climate model. Figure 2.8, S8, S9 and S10 show the projected future and historical FDCs with four different climate models. When the projected streamflow curve is above the historical curve, the ML model indicates that higher streamflow rates become more probable. It is perhaps not surprising that since precipitation increases across almost all basins, almost all of the basins show increasing streamflow. The projections also generally indicate that the peak flow rate will be higher, potentially indicative of an increased probability of flooding (although the degree to which this is possible is a subject for future investigation). Note that the multimodel CMIP5 ensemble does produce some disagreement: For instance, under the MIROC5 projection, the FDC curves for historical and projection match closely for the most

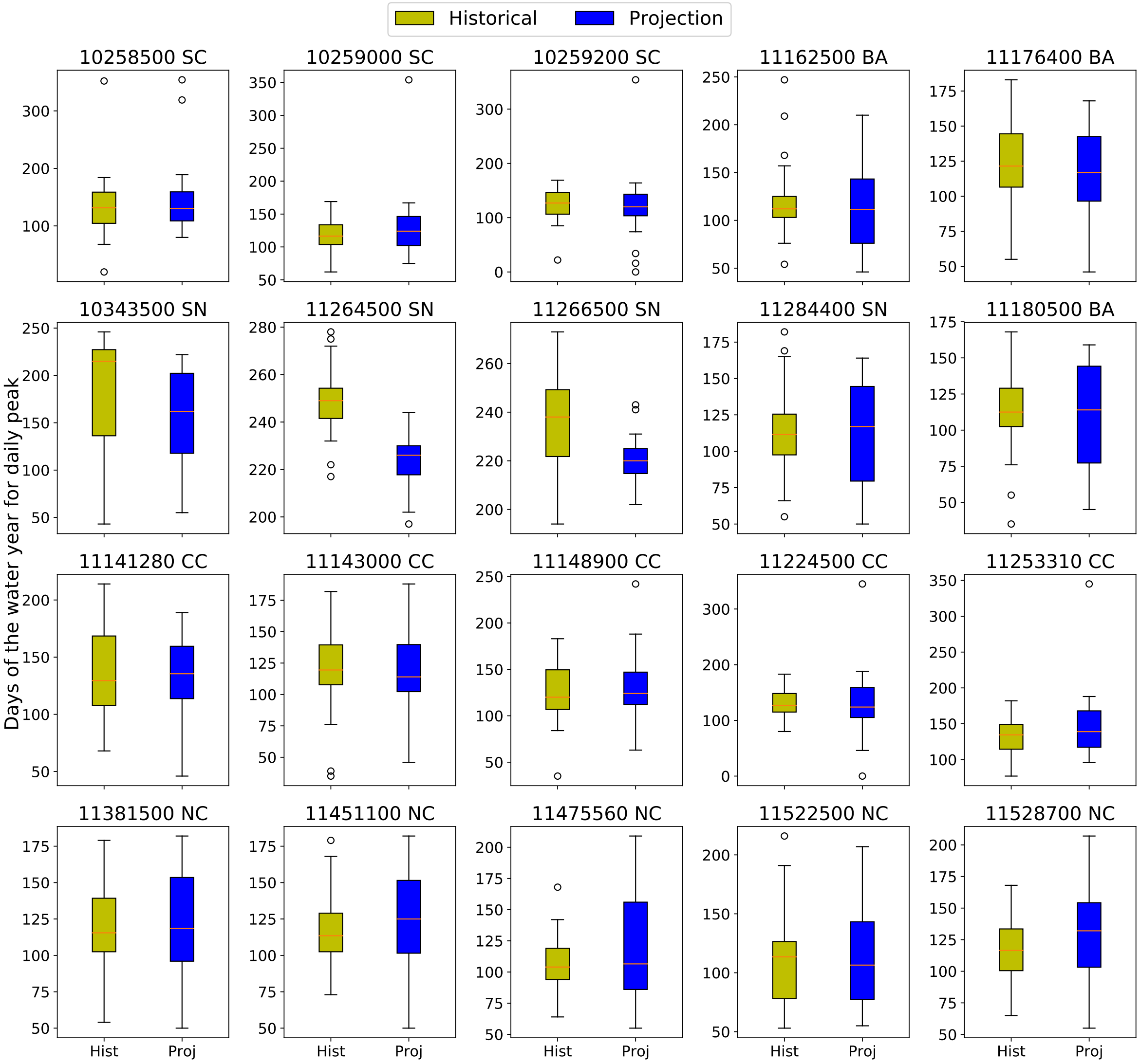

Figure 2.9: Day of peak flow for each basin with MIROC5 forcing.

basins.As noted earlier, the MIROC5 model is considered the most unlike the other CMIP5 models in this investigation, tending to produce precipitation amounts that are relatively constant over time.

Although most basins see an increase in flow rate, basins 10343500 (SN), 11264500 (SN) and 11266500 (SN) are notable exceptions. For these three basins, the future FDC curves are sometimes below the historical curves (this is even more obvious with MIROC5 forcing). For basins 11264500 (SN) and 1266500 (SN) lower flow rates become more probable but the maximum flow rate decreases. These three basins are all in the Sierra Nevada area – 10343500 (SN) in the Tahoe National Forest and the other two in Yosemite. Examining Figure 2.6 and 2.7, these three basins exhibit significant differences in the character of their flow compared with other basins. Namely, the climatological streamflow for these basins shows a peak in late Spring and Summer, while other basins are peaked in the winter season. Since we have shown earlier that streamflow in these basins are driven by snow dynamics, differences in streamflow are likely due to the impact of a slow snowmelt process. Notably, this is in accord with our previous discussion in 2.3.3 and 2.3.4 where these basins are temperature dominant and benefit from longer time window sizes. These projection results lend further evidence to the claim that streamflow in these basins is highly dependent on snowmelt.

The change in the peak flow timing for each basin was also investigated. The peak time is defined as the day of maximal flow rate for the year, measured in days since the beginning of a water year (set to October 1^{st} in our study). Figure 2.9 shows the peak time for each basin in historical and projection years with MIROC5 forcing. Peak timing figures with forcings from other climate models can be found in Figure S11, S12 and S13 in supplements. Although there is generally no significant change in peak timing for most basins, the Sierra Nevada basins are again outliers. Namely, there is a statistically significant shift to earlier peak times in these snowpack-dominated basins. Although it is not always the case for all the climate models, the projected lead of peak time associated with decrease of streamflow in the future again captures the unique hydrology dynamics in the Sierra Nevadas.

2.4.2 Understanding Nonlinearity in the Projection

To better understand the nonlinearity of the streamflow response to forcing under climate change, we consider a decomposition of the response according to its predictors. Specifically, the impact of precipitation alone on the projected streamflow can be isolated by holding the temperature and solar radiation at historical values while using the future projected precipitation. An analogous approach can then be employed for temperature and solar radiation. By then subtracting the historical streamflow time series from each of these streamflow projections, we obtain ΔQ_p, ΔQ_t, ΔQ_s, the change in streamflow from precipitation alone, temperature alone, and solar radiation alone. These are contrasted against ΔQ_{pts}, which denotes the change in streamflow from all three factors. From the first-order Taylor series expansion we then have

$$\begin{aligned}\Delta Q_{pts} =& \Delta Q_p + \Delta Q_t + \Delta Q_s + r \\ =& \Delta Q_{linear} + r\end{aligned} \tag{2.7}$$

for some residual r that captures the influence of high-order terms. The linear response is defined as summation of three individual responses. To reduce noise from daily variations in streamflow, the monthly averaged streamflow is used for comparison. In Figure 2.10 we plot ΔQ_{pts} versus ΔQ_{linear}, with the R^2 value in the title. A fully linear response would be expected to lay along the $y = x$ line.

As seen in Figure 2.10, almost all basins show a nearly linear response to the input variables, except for basin 10343500(SN), 11264500 (SN) and 11266400(SN) – all in the Sierra Nevada mountains. From our discussion in section 2.3.3 and 2.3.4, these SN basins are temperature dominated and require a longer time window size to correctly capture streamflow, indicating the interplay between precipitation and temperature in governing snow processes.

2.5 Conclusions and Future Work

In this study, a general temporal convolutional neural network for streamflow projection in California has been designed and analyzed. Causal convolution is used to maintain physical cau-

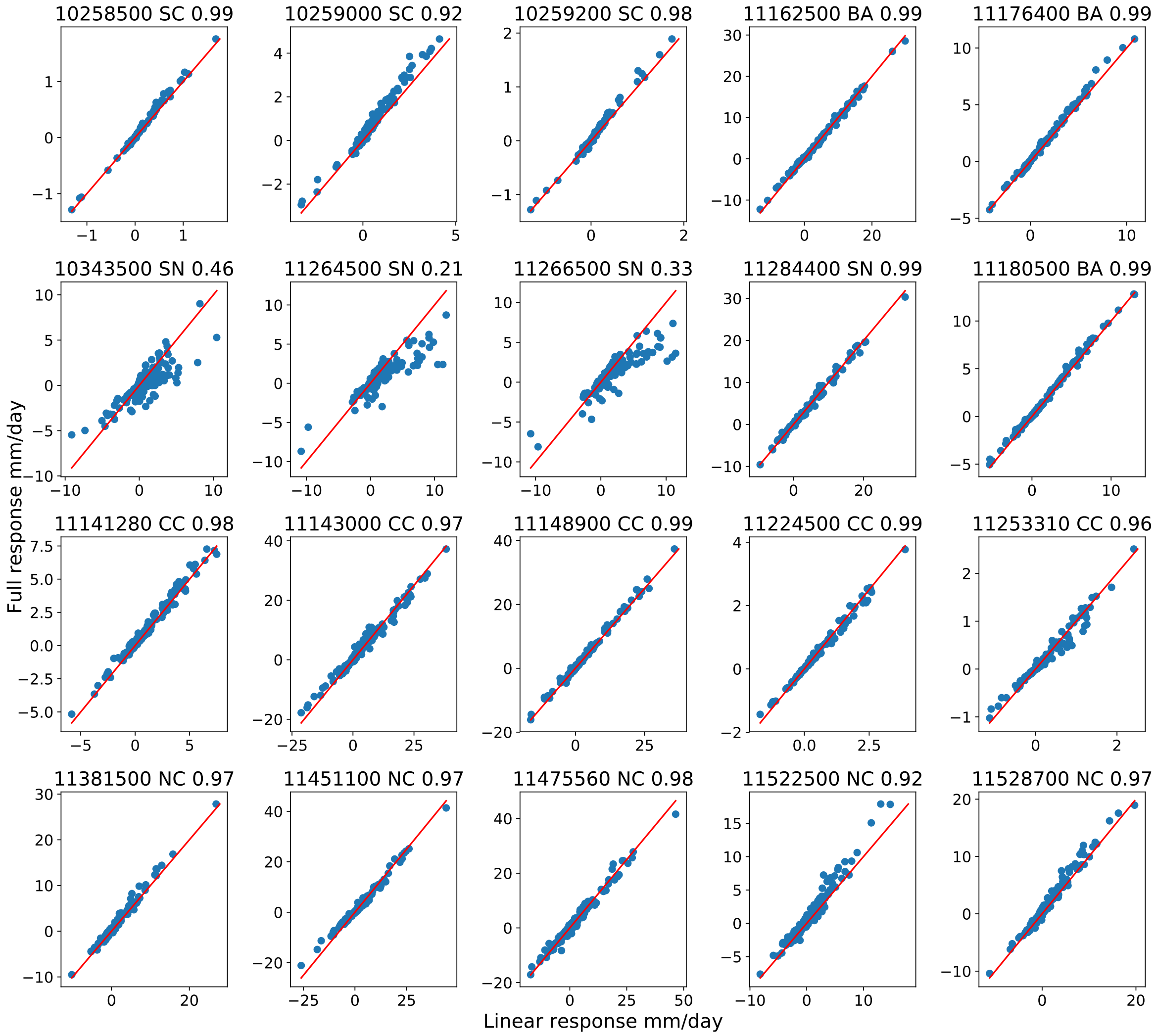

Figure 2.10: Full response and linear response for all basins with CanESM2 forcing.

sation. The input consists of precipitation, temperature and solar radiation over a particular past window size. In prediction mode, the TCNN model is compared with other commonly used ML models based on ensemble performance so as to eliminate random effects from initializing the training. The results of this intercomparison indicate there are some important temporal features that ANNs struggle to capture, in contrast to TCNNs and other recurrent neural networks (LSTMs and GRUs). Compared with other recurrent networks, the TCNN model is faster and more stable under training. Overall, the TCNN produces better agreement both on average and in the high-flow regime, whereas the LSTM was better in the low-flow regime. Like these other networks, the TCNN model can also be generalized to other basins while maintaining the same architecture.

To demonstrate model stability under extreme forcing, an idealized test with quadruple precipitation and 5 Celsius higher temperature is implemented to verify whether the model produces reasonable results when tested with data outside the training regime. A qualitative analysis and linear regression of projected streamflow against historical precipitation suggests our model produces physically acceptable results for projection.

We have also observed that the TCNN model can build different functional relationships for different basins, as demonstrated through the examination of reduced models, models with different time window sizes, and the nonlinear response of the model to input variables. With this understanding of the "under the hood" workings of the ML model, we can distinguish different geographic features across basins. This classification ability suggests our model can simulate physical processes with causal convolution as a constraint. In regions where snowpack is relevant, we conclude that temperature should be included as a model covariate; whereas in coastal regions, solar radiation should be included. Including both variables was not observed to significantly improve model performance in any basin. Also, in regions where snowpack is relevant, a longer time window size is desirable for model performance (here we tested a 365-day window), whereas in other regions a shorter time window of 100-days produced better results.

Under the RCP8.5 scenario, the nonlinearity of the streamflow response was examined by decomposing the response into three modes by the predictors. By inspecting the linear response and full response, we observed that most basins exhibit a linear response from precipitation, temperature and solar radiation, except for the basins in Sierra Nevada. The nonlinearity is likely associated with snowpack, which is a physical feature that is sensitive to both precipitation and temperature.

Model results for future projections and historical hindcasts were compared to understand the changing character of the streamflow. Generally streamflow in most basins increases through the end of the century, except for the Sierra Nevada basins. Peak flow time remained statistically indistinguishable among most basins, except the Sierra Nevada basins which showed a shift to earlier dates under some models. These results further indicate that the snow dynamics in the Sierra Nevada is important for correctly capturing streamflow in these basins.

The idealized test here mainly deals with the problem of model stability under extrapolation. In terms of ensuring the model produces physically plausible results under extreme forcings, we need to compare with a physically based model with the same extreme forcing. This problem has been saved for our future work. Also, to better understand the ML model and ensure its credibility for producing future projections, we intend to next cross-validate our projection datasets with a physically-based model over the same time period. Model credibility can also be enhanced through alternative designs that explicitly include physically-based conservation laws. For instance, subsurface flow or evaporation are not produced as outputs, and so validation of the water budget is impossible. With a more complicated design, ML models could predict streamflow, evaporation and groundwater, and be constrained via an appropriate physically-based conservation law. Such constraints would further enable physical interpretation of the model results. Finally, we wish to determine if the TCNN can be used to interpolate predictors to higher temporal resolution, for use (for instance) in physically-based models. The ML model could also be used to examine model performance when the strict causation is relaxed (namely, if future streamflow could provide a better estimate of present streamflow).

2.6 Supplements

Table S1: NLDAS mean and standard deviation for the historical period.

	precip_NLDAS_mean	precip_NLDAS_std	temp_NLDAS_mean	temp_NLDAS_std	solar_NLDAS_mean	solar_NLDAS_std
10258500	0.60	2.84	15.47	7.66	242.90	82.87
10259000	0.57	2.73	17.35	7.79	240.52	81.79
10259200	0.51	2.74	16.51	7.96	242.87	82.38
10343500	2.44	7.64	5.77	7.56	217.66	92.51
11141280	1.72	6.86	14.80	4.61	232.04	92.13
11143000	2.62	9.47	11.45	3.69	229.82	94.60
11148900	1.67	6.40	15.28	6.10	238.76	96.47
11162500	2.48	8.14	13.42	4.49	220.45	97.05
11176400	1.38	4.56	14.02	6.74	224.41	100.89
11180500	1.32	4.45	15.51	4.97	219.63	100.03
11224500	1.07	3.99	15.42	7.47	228.94	97.08
11253310	0.84	3.08	16.46	7.53	228.87	98.32
11264500	2.74	8.78	3.30	7.23	221.99	86.61
11266500	2.84	9.06	4.36	7.23	221.75	87.00
11284400	2.76	8.89	13.97	7.36	219.07	97.54
11381500	3.69	10.05	10.51	7.43	210.12	101.55
11451100	2.47	7.84	12.14	6.87	217.48	104.22
11475560	4.71	13.40	11.82	5.56	211.88	106.17
11522500	3.76	10.14	8.82	6.78	205.54	104.73
11528700	3.44	9.25	10.71	7.23	208.84	105.25

Table S2: CanESM2 model mean and standard deviation for the historical period.

	precip_LOCA_mean	precip_LOCA_std	temp_LOCA_mean	temp_LOCA_std	solar_LOCA_mean	solar_LOCA_std
10258500	0.95	4.05	16.00	7.47	313.82	105.63
10259000	1.07	4.22	14.63	7.29	313.15	107.86
10259200	0.81	4.42	15.71	7.84	316.66	105.23
10343500	2.79	8.45	4.33	7.67	270.30	128.21
11141280	1.80	6.88	12.25	3.61	308.46	116.67
11143000	2.88	9.50	10.11	3.81	302.58	120.67
11148900	1.74	6.07	12.50	4.80	308.88	119.49
11162500	2.63	8.04	12.05	3.89	288.10	120.16
11176400	1.34	4.00	14.06	6.57	300.08	125.18
11180500	1.48	4.58	14.32	4.58	289.12	119.94
11224500	1.41	5.50	13.95	6.99	305.30	119.58
11253310	1.10	4.17	14.90	7.13	302.28	118.56
11264500	3.33	10.43	2.67	7.61	260.03	120.30
11266500	3.33	10.32	3.70	7.64	267.55	121.82
11284400	2.54	8.05	13.16	7.30	294.63	128.43
11381500	3.85	9.79	9.86	7.29	273.39	134.15
11451100	3.04	8.83	11.02	6.85	287.48	131.36
11475560	6.40	16.45	8.56	4.55	279.04	133.81
11522500	4.26	10.36	7.41	7.09	262.34	133.77
11528700	4.22	11.05	8.82	7.15	273.29	132.45

Table S3: CNRMCM5 model mean and standard deviation for the historical period.

	precip_LOCA_mean	precip_LOCA_std	temp_LOCA_mean	temp_LOCA_std	solar_LOCA_mean	solar_LOCA_std
10258500	0.84	3.68	16.22	7.42	273.30	132.45
10259000	0.97	4.02	14.86	7.22	273.30	132.45
10259200	0.69	3.98	15.91	7.84	273.30	132.45
10343500	2.85	8.70	4.44	7.75	273.30	132.45
11141280	1.73	6.46	12.29	3.53	273.30	132.45
11143000	2.78	9.08	10.12	3.83	273.30	132.45
11148900	1.66	5.70	12.63	4.87	273.30	132.45
11162500	2.59	7.73	12.02	4.00	273.30	132.45
11176400	1.37	3.99	14.06	6.62	273.30	132.45
11180500	1.52	4.70	14.29	4.67	273.30	132.45
11224500	1.38	5.29	14.01	7.02	273.30	132.45
11253310	1.09	4.05	14.97	7.16	273.30	132.45
11264500	3.38	10.27	2.81	7.63	273.30	132.45
11266500	3.38	10.23	3.83	7.67	273.30	132.45
11284400	2.62	8.13	13.21	7.30	273.30	132.45
11381500	3.89	9.64	9.93	7.35	273.30	132.45
11451100	3.08	8.93	11.03	6.95	273.30	132.45
11475560	6.35	16.01	8.54	4.73	273.30	132.45
11522500	4.40	10.50	7.42	7.22	273.30	132.45
11528700	4.21	10.71	8.84	7.26	273.30	132.45

Table S4: HadGEM2ES model mean and standard deviation for the historical period.

	precip_LOCA_mean	precip_LOCA_std	temp_LOCA_mean	temp_LOCA_std	solar_LOCA_mean	solar_LOCA_std
10258500	0.89	3.78	16.02	7.35	273.30	132.45
10259000	1.04	4.17	14.65	7.17	273.30	132.45
10259200	0.71	3.79	15.74	7.72	273.30	132.45
10343500	2.81	8.65	4.09	7.53	273.30	132.45
11141280	1.92	6.95	11.86	3.43	273.30	132.45
11143000	2.96	9.51	9.64	3.68	273.30	132.45
11148900	1.82	6.19	11.87	4.55	273.30	132.45
11162500	2.77	8.13	11.66	3.90	273.30	132.45
11176400	1.42	4.12	13.62	6.46	273.30	132.45
11180500	1.58	4.71	13.93	4.49	273.30	132.45
11224500	1.47	5.55	13.61	6.88	273.30	132.45
11253310	1.16	4.24	14.59	7.06	273.30	132.45
11264500	3.44	10.55	2.44	7.55	273.30	132.45
11266500	3.44	10.43	3.47	7.57	273.30	132.45
11284400	2.67	8.12	12.85	7.14	273.30	132.45
11381500	3.87	9.62	9.61	7.21	273.30	132.45
11451100	3.15	8.80	10.70	6.85	273.30	132.45
11475560	6.61	16.42	8.27	4.61	273.30	132.45
11522500	4.29	10.33	7.23	7.07	273.30	132.45
11528700	4.19	10.61	8.58	7.15	273.30	132.45

Table S5: MIROC5 model mean and standard deviation for the historical period.

	precip_LOCA_mean	precip_LOCA_std	temp_LOCA_mean	temp_LOCA_std	solar_LOCA_mean	solar_LOCA_std
10258500	0.95	3.91	15.81	7.39	273.30	132.45
10259000	1.09	4.27	14.44	7.20	273.30	132.45
10259200	0.77	4.18	15.55	7.76	273.30	132.45
10343500	2.90	8.66	3.96	7.53	273.30	132.45
11141280	1.86	6.90	12.04	3.43	273.30	132.45
11143000	2.85	9.01	9.79	3.73	273.30	132.45
11148900	1.76	5.87	12.19	4.74	273.30	132.45
11162500	2.70	8.02	11.78	3.84	273.30	132.45
11176400	1.41	4.01	13.69	6.52	273.30	132.45
11180500	1.56	4.69	14.06	4.56	273.30	132.45
11224500	1.45	5.47	13.64	6.91	273.30	132.45
11253310	1.14	4.21	14.59	7.06	273.30	132.45
11264500	3.47	10.61	2.31	7.51	273.30	132.45
11266500	3.47	10.53	3.33	7.55	273.30	132.45
11284400	2.70	8.26	12.77	7.16	273.30	132.45
11381500	3.90	9.68	9.55	7.21	273.30	132.45
11451100	3.08	8.85	10.73	6.83	273.30	132.45
11475560	6.41	15.81	8.27	4.51	273.30	132.45
11522500	4.29	10.28	7.13	7.01	273.30	132.45
11528700	4.15	10.42	8.52	7.06	273.30	132.45

Table S6: Historical and projection comparison of mean precipitation.

	precip_NLDAS	precip_CanESM2	precip_CNRMCM5	precip_HadGEM2ES	precip_MIROC5
10258500	0.60	0.93	0.96	0.58	0.48
10259000	0.57	1.00	0.91	0.53	0.42
10259200	0.51	0.72	0.80	0.49	0.44
10343500	2.44	4.13	3.30	2.56	2.42
11141280	1.72	2.53	2.45	1.82	1.50
11143000	2.62	4.04	3.83	3.02	2.47
11148900	1.67	2.52	2.46	1.92	1.51
11162500	2.48	3.65	3.54	2.76	2.41
11176400	1.38	1.90	1.83	1.41	1.27
11180500	1.32	1.92	1.81	1.45	1.26
11224500	1.07	1.60	1.54	1.16	0.95
11253310	0.84	1.22	1.16	0.89	0.74
11264500	2.74	4.24	3.62	2.84	2.47
11266500	2.84	4.39	3.74	2.92	2.56
11284400	2.76	3.99	3.55	2.76	2.42
11381500	3.69	4.92	4.83	3.64	3.69
11451100	2.47	3.59	3.36	2.80	2.46
11475560	4.71	5.81	6.19	4.96	4.90
11522500	3.76	4.50	4.41	3.64	3.96
11528700	3.44	4.19	4.27	3.57	3.50

Table S7: Historical and projection comparison of mean temperature.

	temp_NLDAS	temp_CanESM2	temp_CNRMCM5	temp_HadGEM2ES	temp_MIROC5
10258500	15.47	15.73	14.94	15.90	15.52
10259000	17.35	18.40	17.58	18.59	18.22
10259200	16.51	16.96	16.21	17.13	16.73
10343500	5.77	12.17	11.80	12.77	12.15
11141280	14.80	15.65	14.93	16.40	15.25
11143000	11.45	13.33	12.58	13.86	12.93
11148900	15.28	17.25	16.24	18.33	16.69
11162500	13.42	14.50	13.63	14.77	13.82
11176400	14.02	13.71	12.94	14.08	13.30
11180500	15.51	15.03	14.26	15.30	14.38
11224500	15.42	16.69	16.01	17.35	16.49
11253310	16.46	16.94	16.25	17.55	16.76
11264500	3.30	9.44	9.20	9.99	9.67
11266500	4.36	10.00	9.75	10.54	10.21
11284400	13.97	15.45	15.16	16.01	15.50
11381500	10.51	13.49	12.99	14.03	13.07
11451100	12.14	14.82	14.01	15.02	14.25
11475560	11.82	15.39	14.32	15.32	14.58
11522500	8.82	13.23	12.61	13.53	12.54
11528700	10.71	14.67	13.91	14.80	13.98

Table S8: Historical and projection comparison of mean solar radiation.

	solar_NLDAS	solar_CanESM2	solar_CNRMCM5	solar_HadGEM2ES	solar_MIROC5
10258500	242.90	235.61	243.33	244.90	244.75
10259000	240.52	233.43	240.86	242.44	242.41
10259200	242.87	236.33	243.37	244.99	244.68
10343500	217.66	209.79	214.50	216.67	219.22
11141280	232.04	224.37	230.63	234.10	233.89
11143000	229.82	223.26	227.56	237.67	232.03
11148900	238.76	232.86	237.47	246.17	241.93
11162500	220.45	213.93	217.96	226.48	223.08
11176400	224.41	218.95	222.18	230.67	227.77
11180500	219.63	214.38	217.59	225.50	222.39
11224500	228.94	222.25	227.34	231.44	231.49
11253310	228.87	222.63	227.40	231.38	231.70
11264500	221.99	212.11	218.34	220.30	223.10
11266500	221.75	211.79	218.21	220.37	223.17
11284400	219.07	211.00	216.21	220.35	221.73
11381500	210.12	204.14	208.06	208.59	211.07
11451100	217.48	213.59	215.83	217.32	217.77
11475560	211.88	209.06	211.16	212.24	211.43
11522500	205.54	200.47	204.56	205.44	205.13
11528700	208.84	204.48	207.90	208.78	208.12

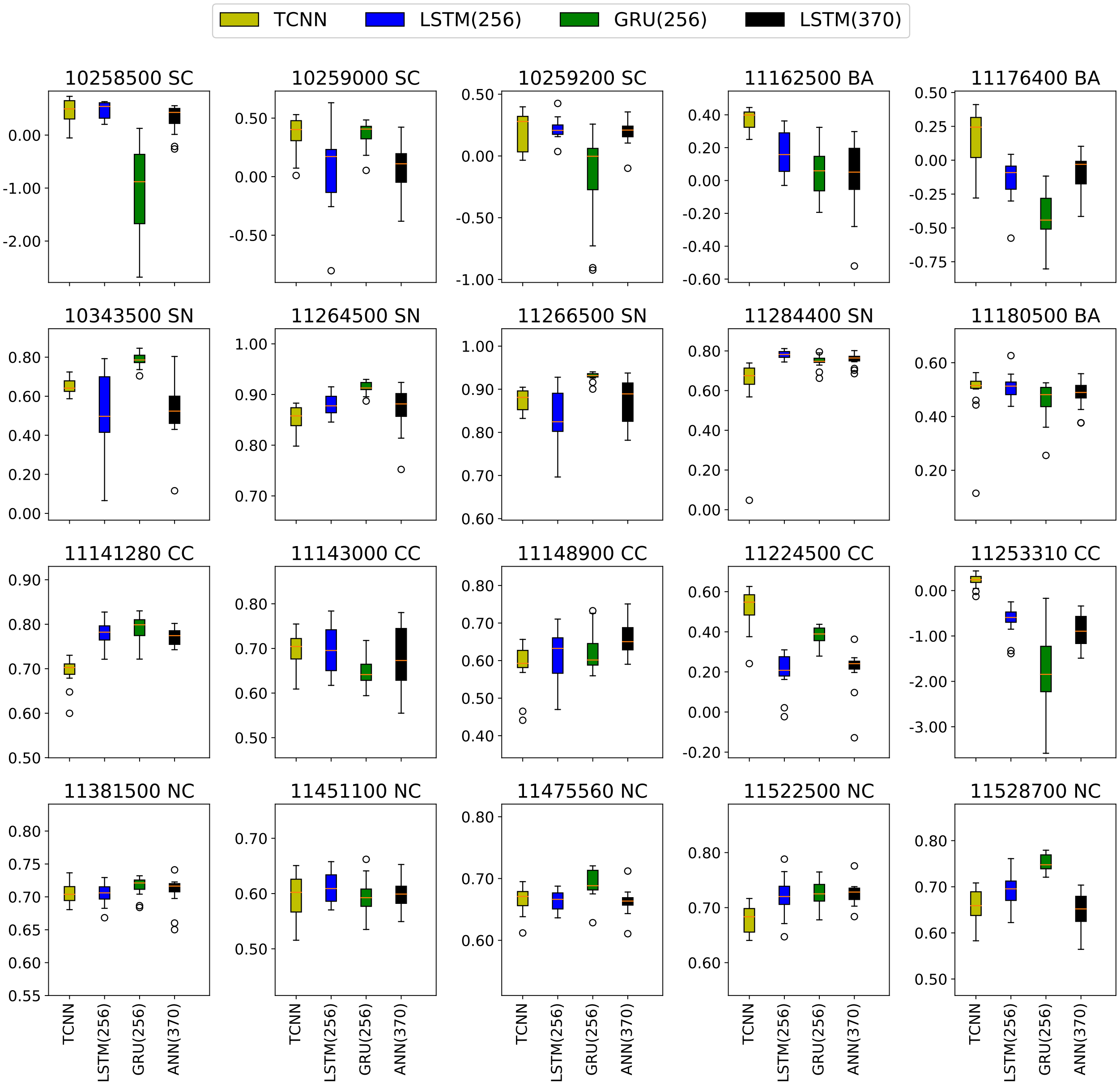

Figure S1: Ensemble prediction comparison for the one-layer recurrent networks and the TCNN.

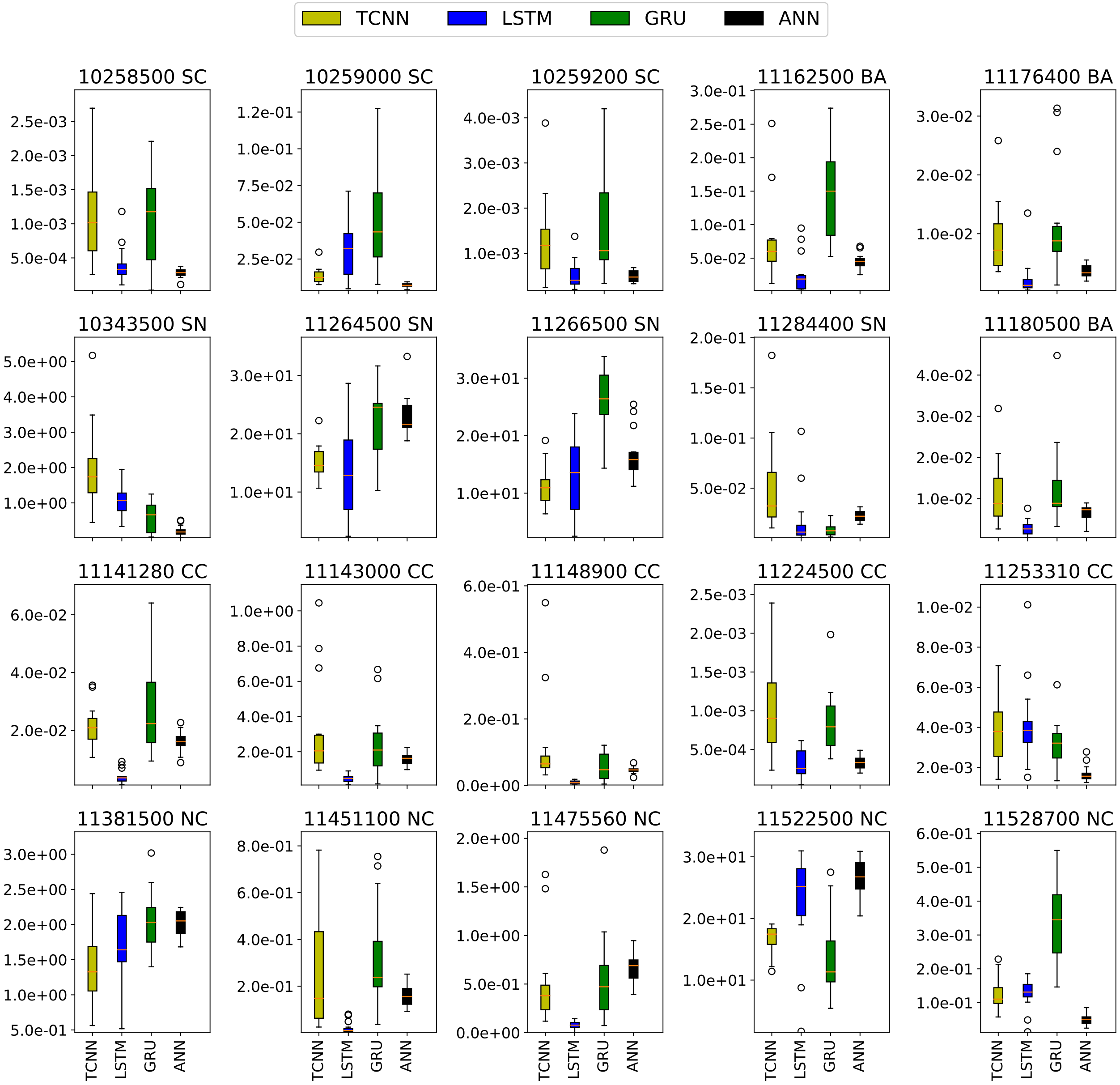

Figure S2: Ensemble performance of low flow period assessed using mean squared error (smaller values are better).

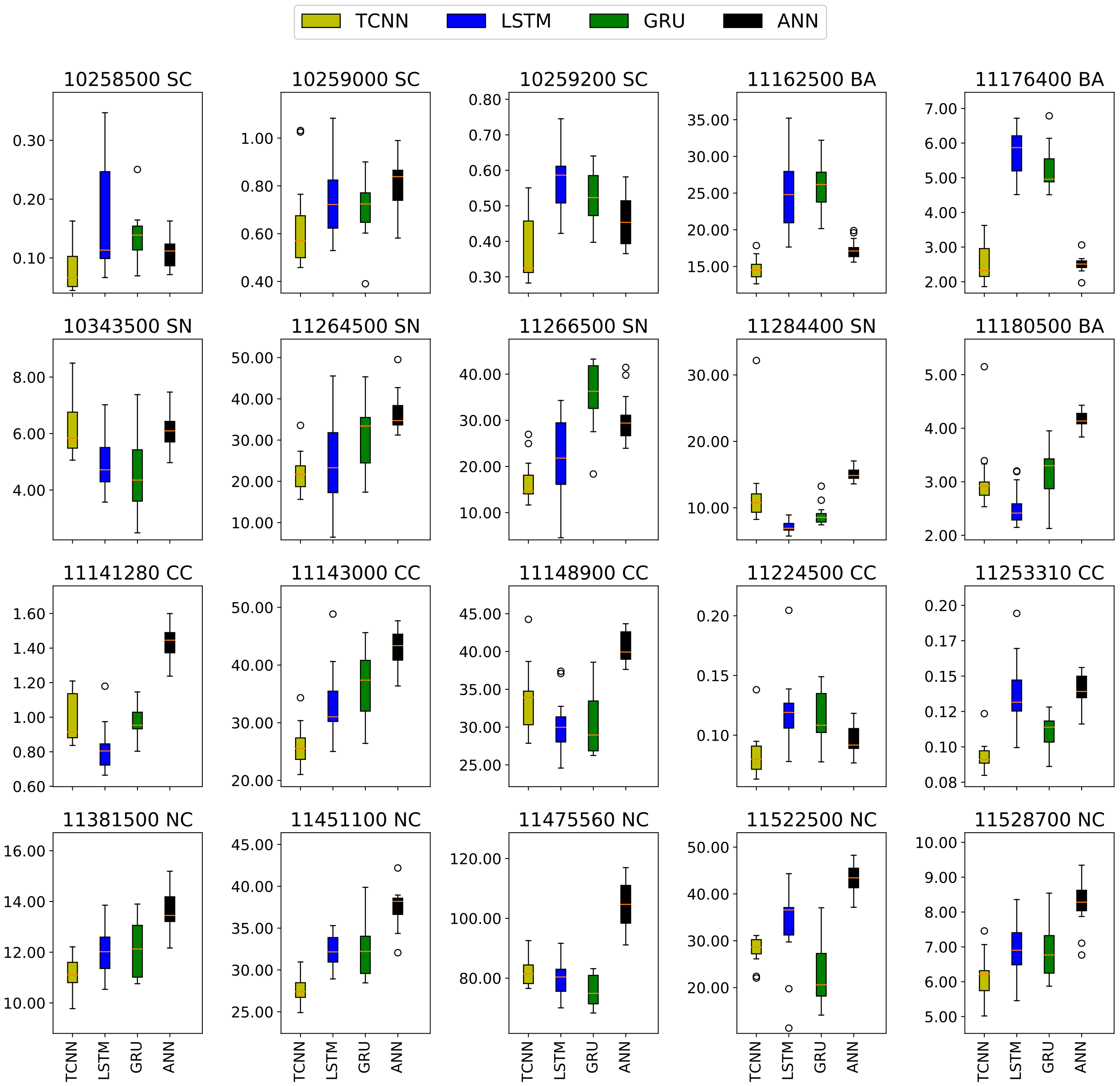

Figure S3: Ensemble performance of high flow period assessed using mean squared error (smaller values are better).

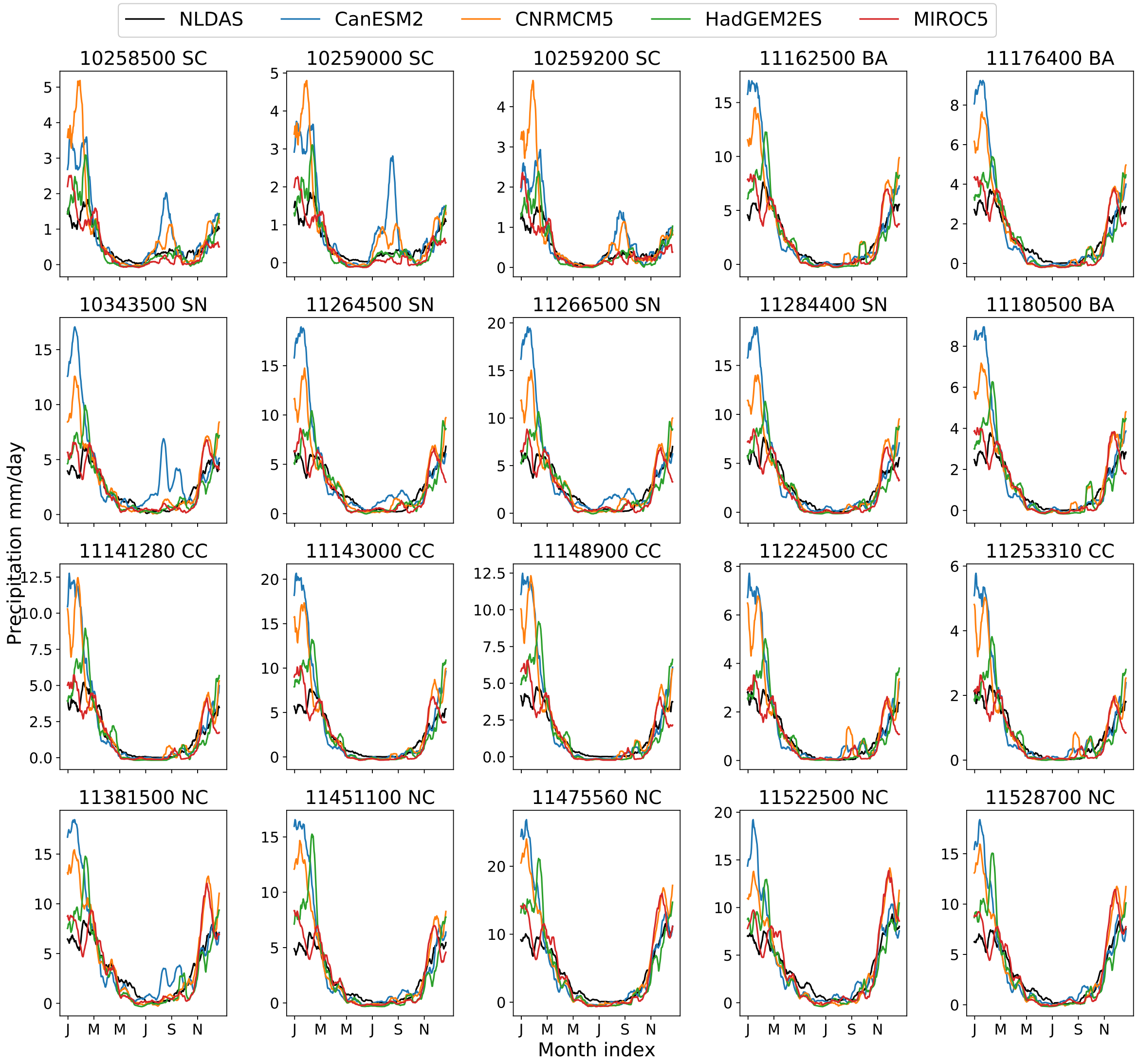

Figure S4: Comparison of precipitation from NLDAS and future forcing (in mm/day). To remove high frequency noise a 15-day running mean filter is applied.

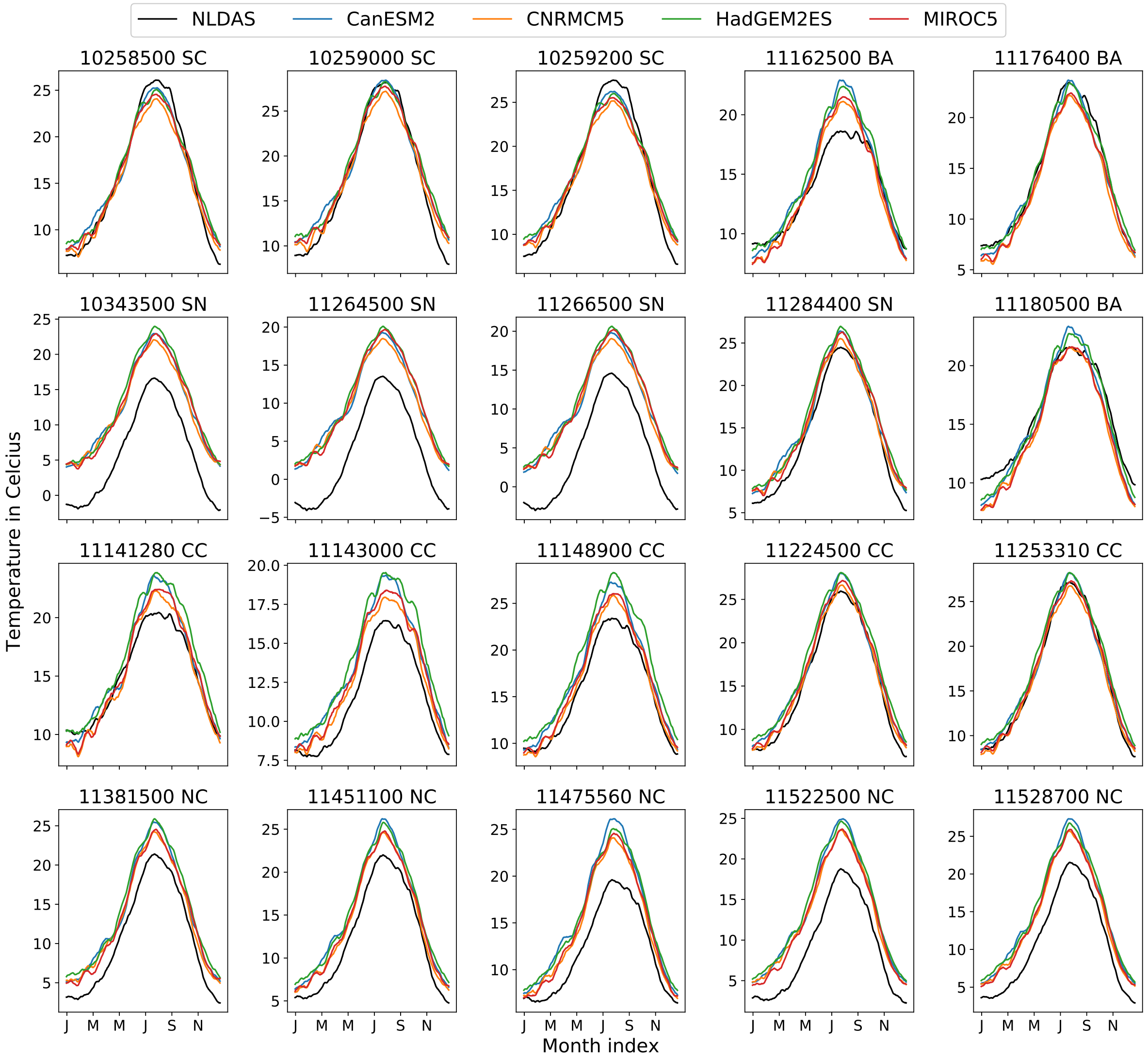

Figure S5: Comparison of temperature from NLDAS and future forcing (in degrees C). To remove high frequency noise a 15-day running mean filter is applied.

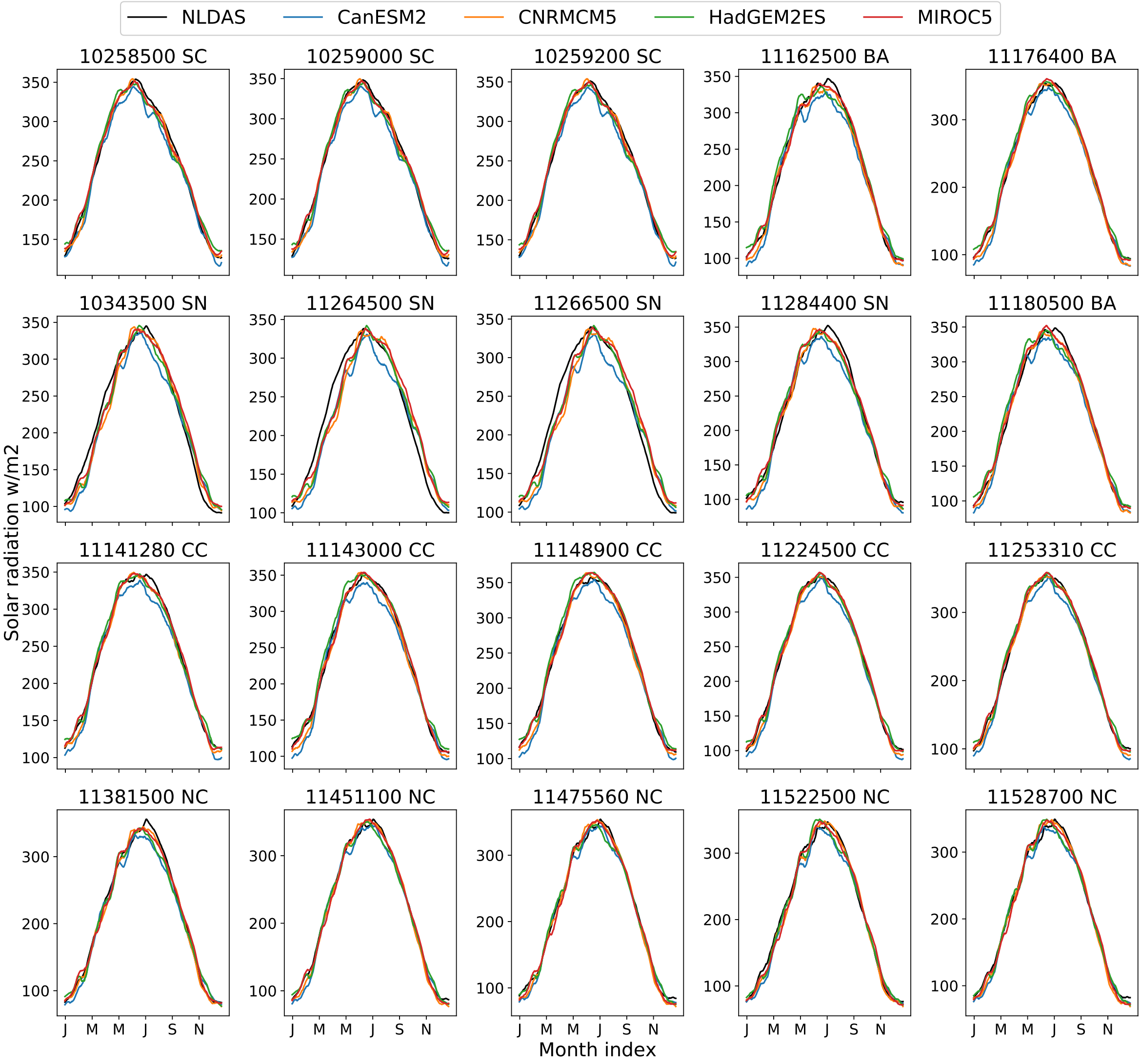

Figure S6: Comparison of solar radiation from NLDAS and future forcing (in W/m^2). To remove high frequency noise a 15-day running mean filter is applied.

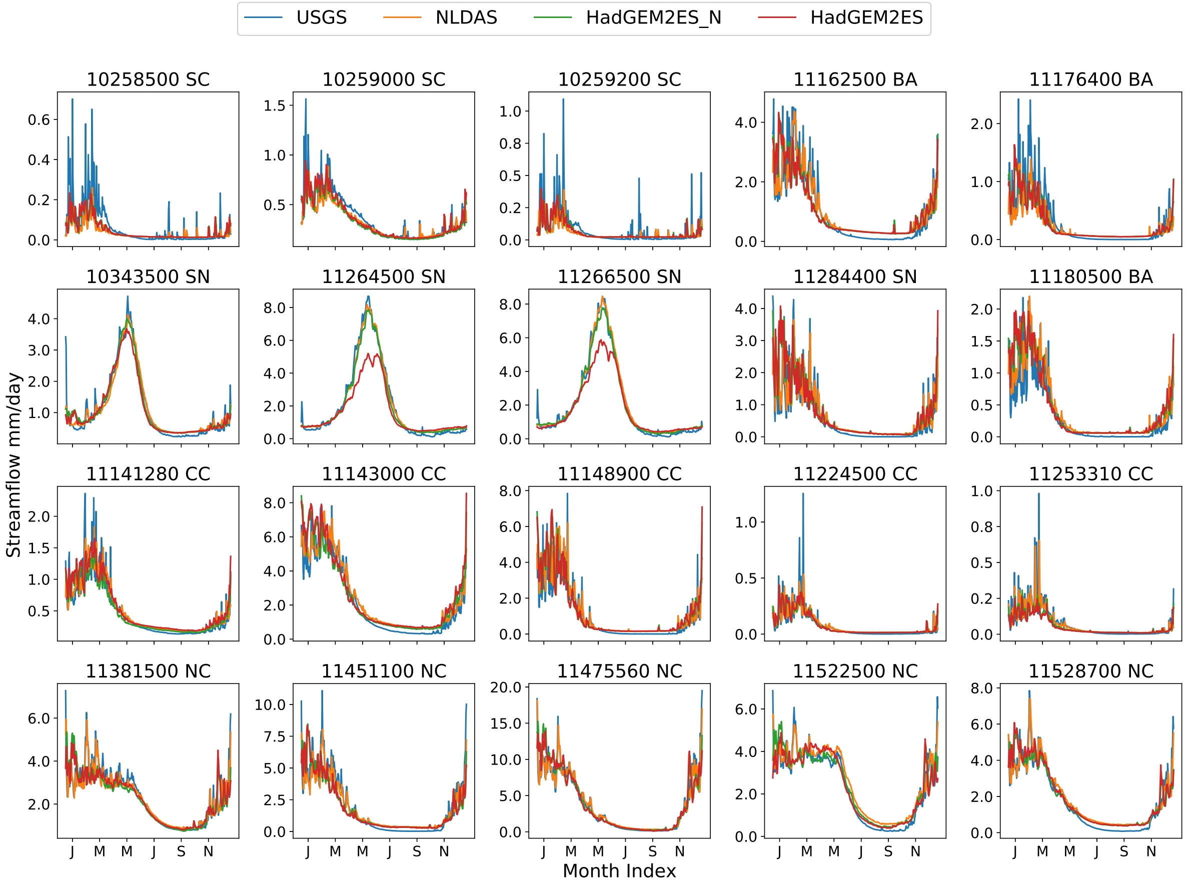

Figure S7: Climatological daily streamflow comparison for different forcing (HadGEM2ES_N represents forcings with LOCA (HadGEM2ES) precipitation and NLDAS temperature and solar radiation).

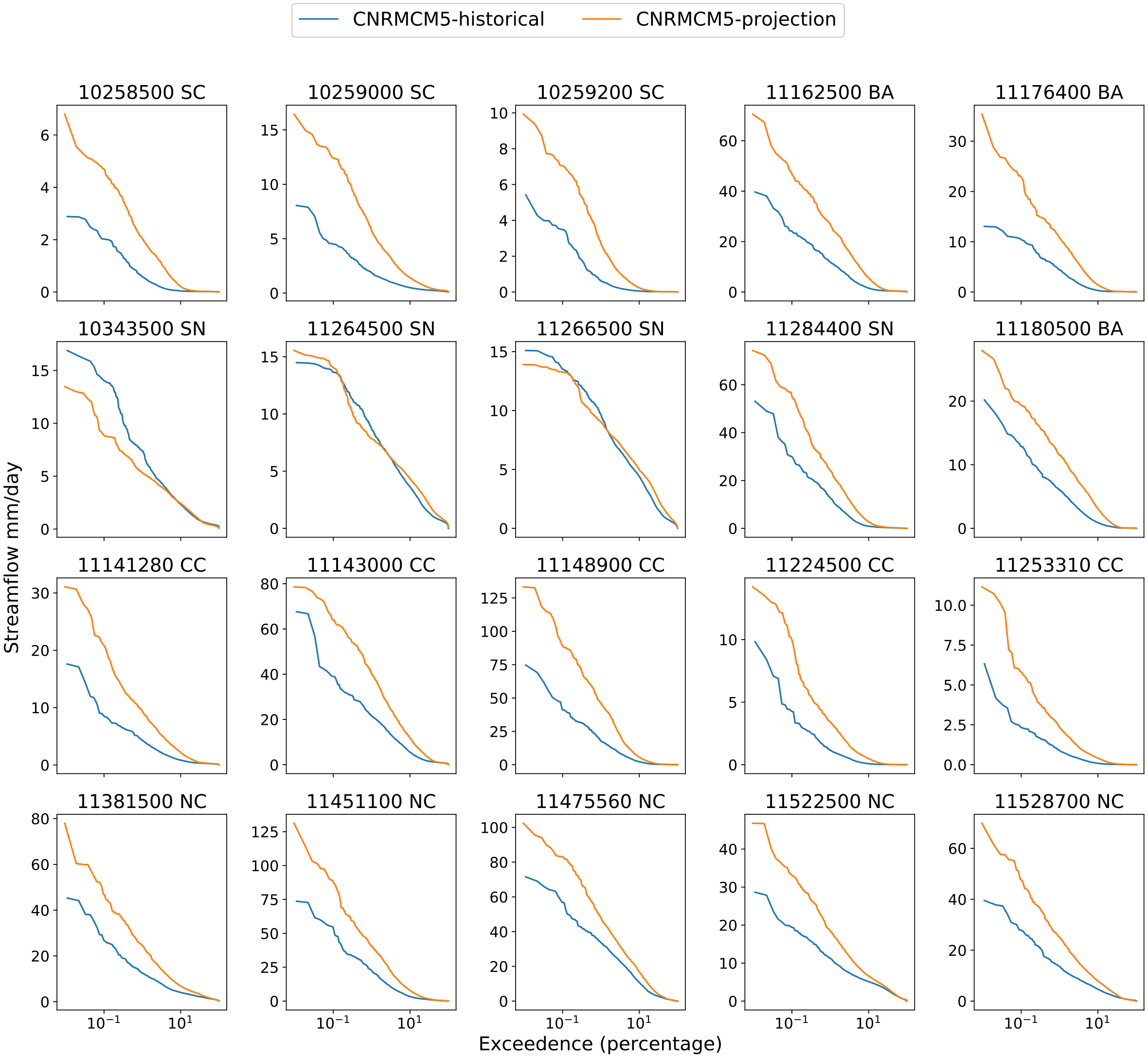

Figure S8: Flow duration curve with CNRMCM5 forcing over both historical and future (projection) periods.

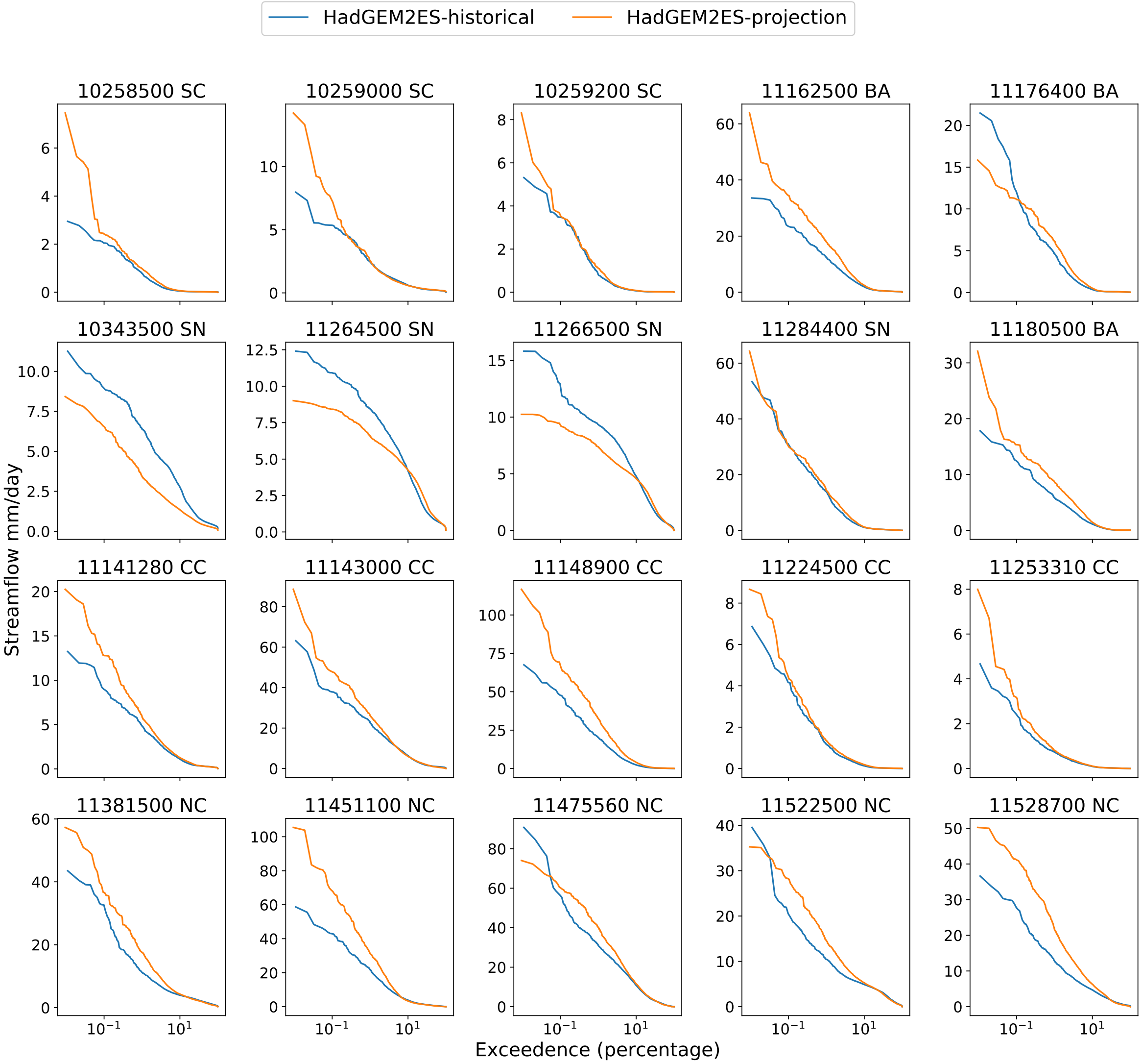

Figure S9: Flow duration curve with HadGEM2ES forcing over both historical and future (projection) periods.

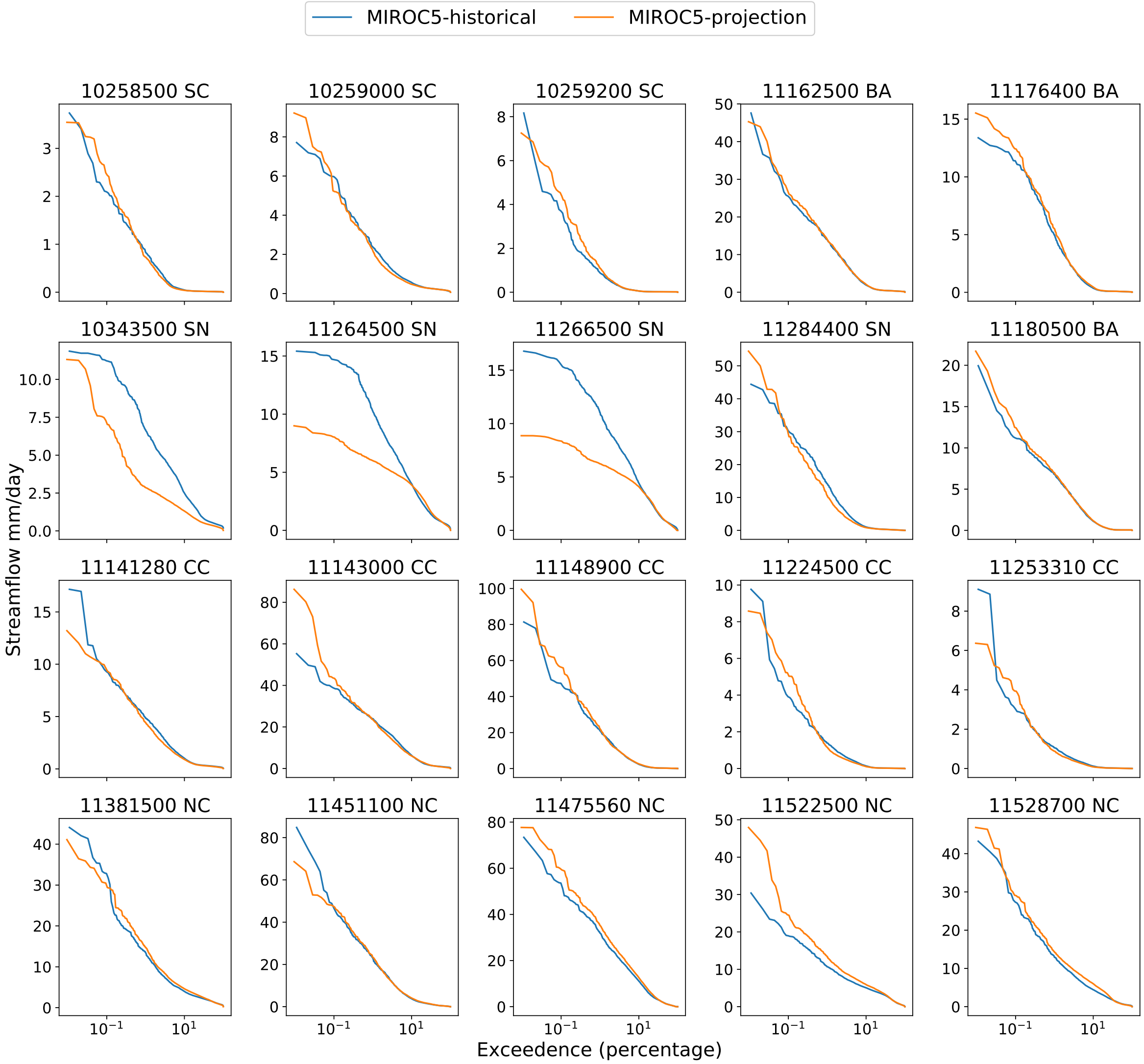

Figure S10: Flow duration curve with MIROC5 forcing over both historical and future (projection) periods.

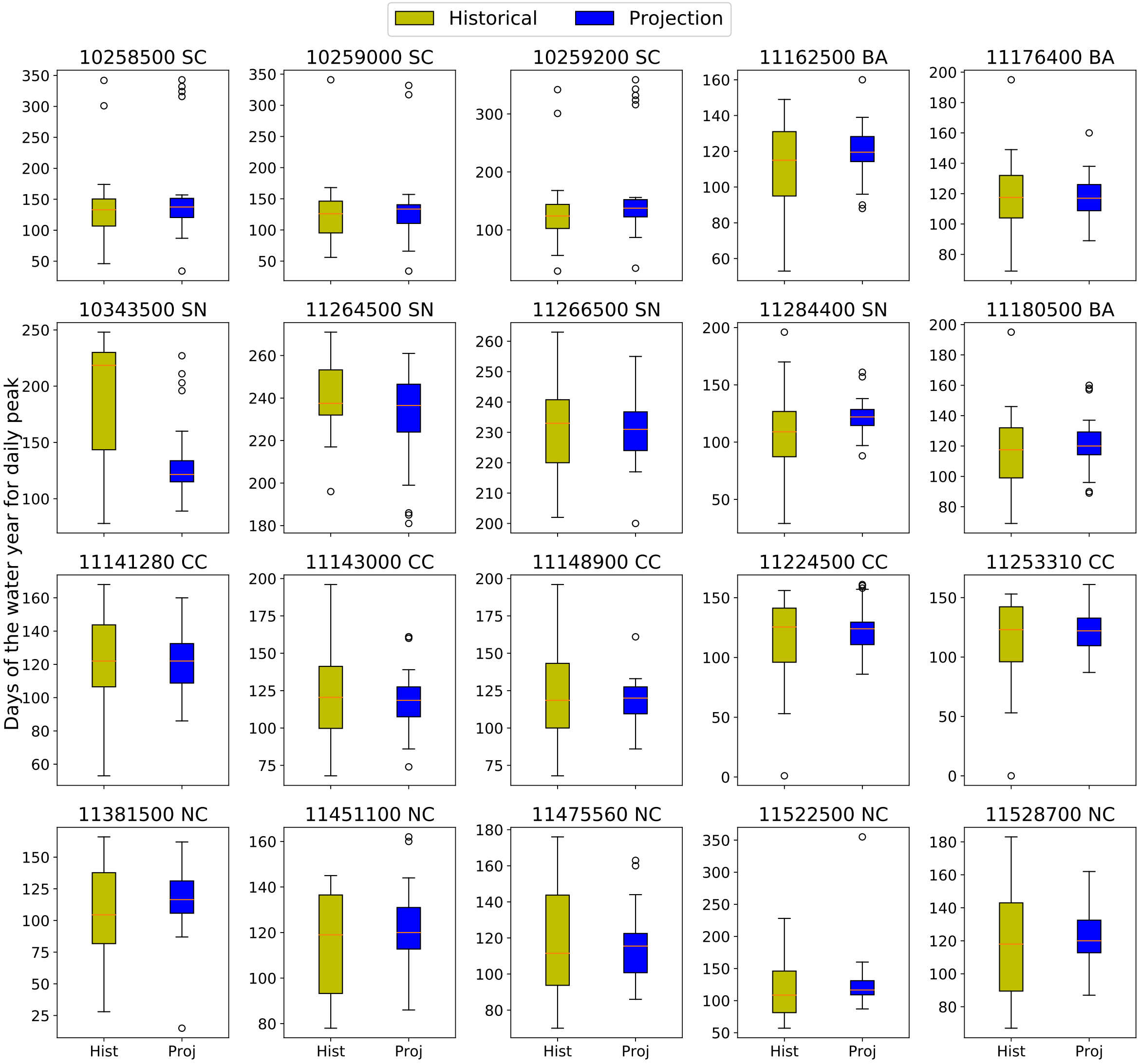

Figure S11: Day of peak flow for each basin with CanESM2 forcing.

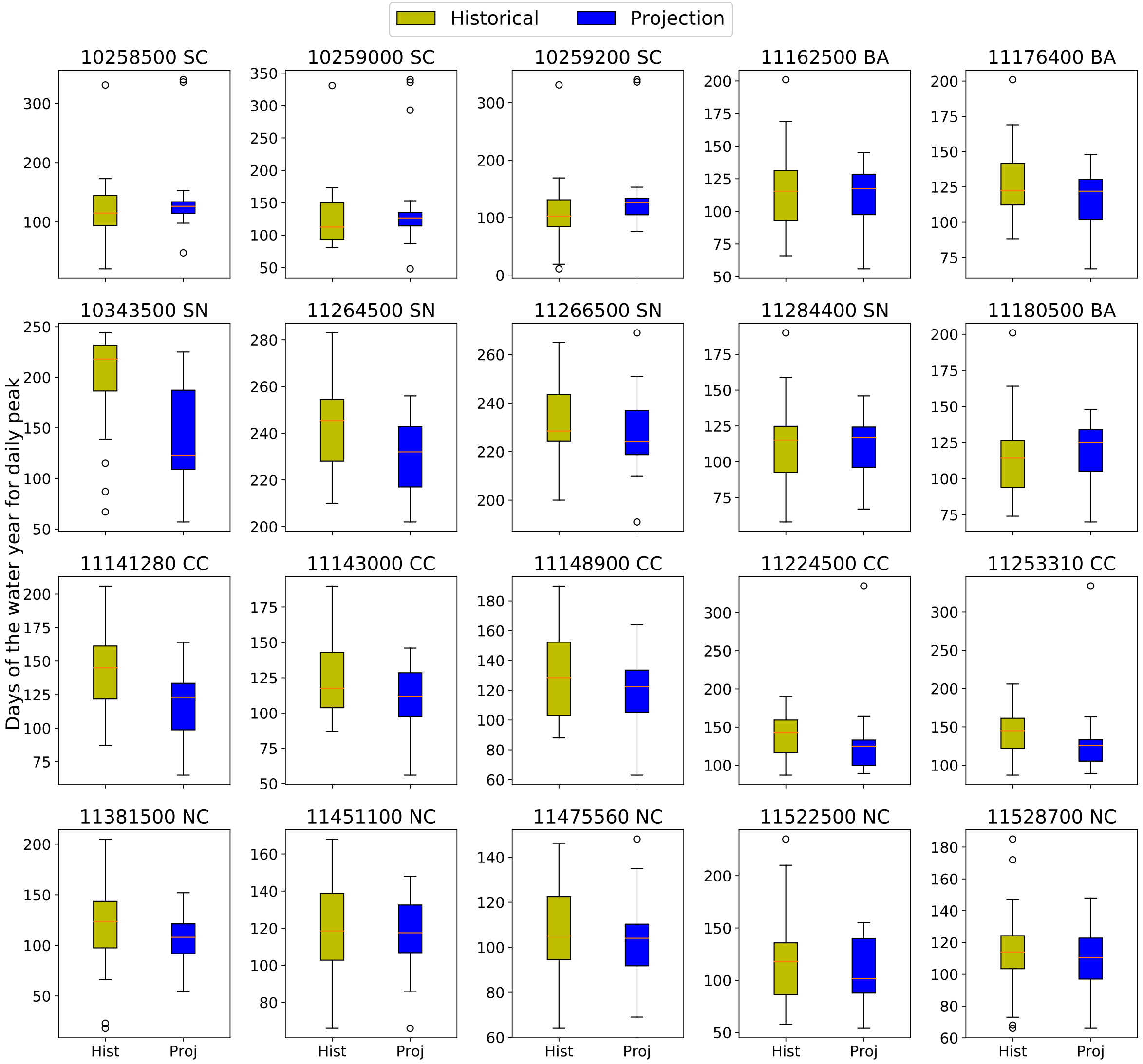

Figure S12: Day of peak flow for each basin with CNRMCM5 forcing.

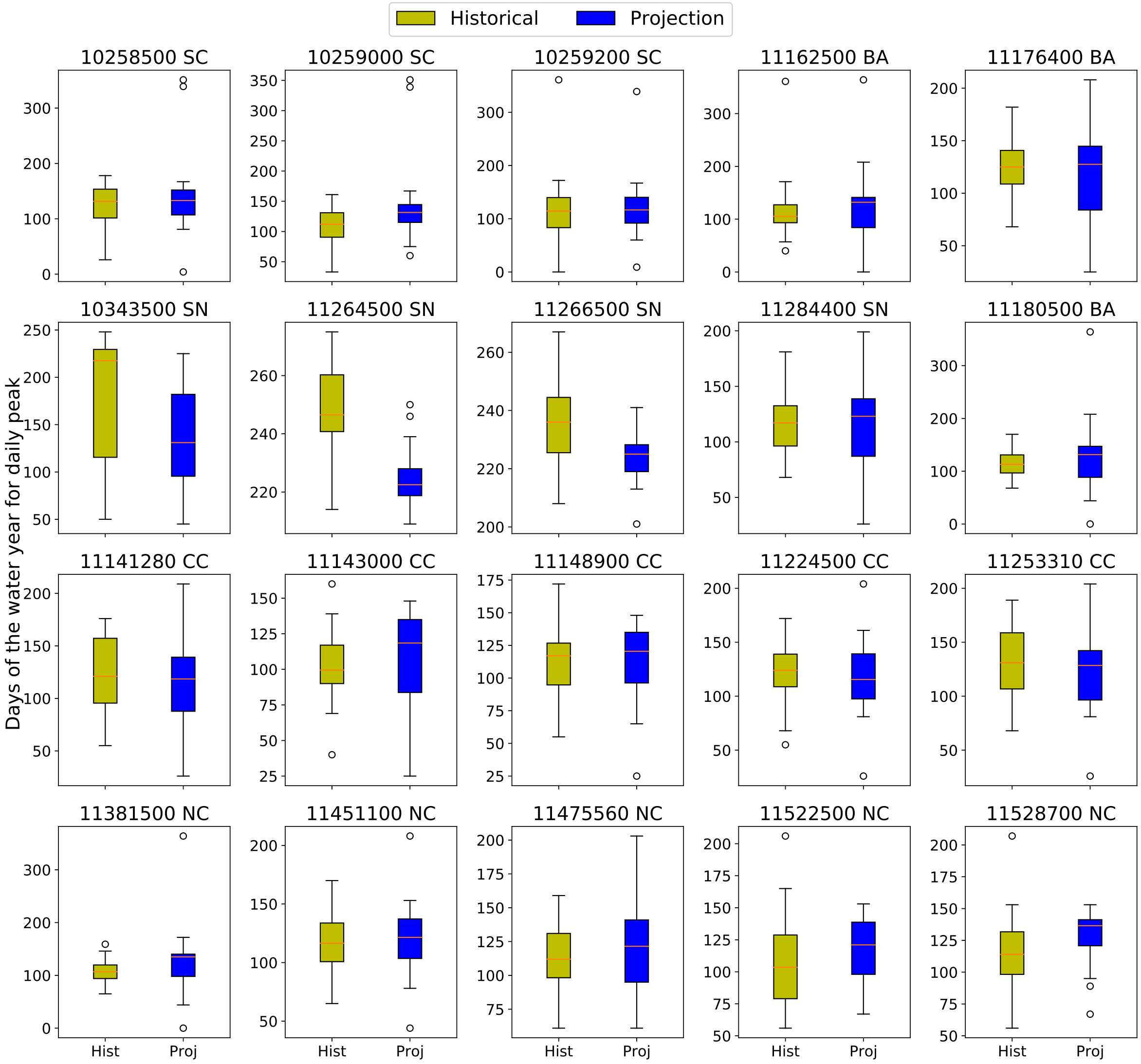

Figure S13: Day of peak flow for each basin with HadGEM2ES forcing.

Chapter 3 Snow Water Equivalent Prediction and Projection in Western U.S.

3.1 Background

Snowpack is a vital component in Earth's hrdro-climate system through its central role in the mountainous hydrologic cycle and influence on water supply to communities that reside downstream of mountains. It is especially true for watersheds in mid-to-high latitudes and at altitudes where streamflow is derived largely by snowmelt, which has been explored in previous chapter for streamflow and Berghuijs et al. (2019). Snow water equivalent (SWE) is a widely used measurement for snowpack studies. It equals to the water amount when the snowpack is completely melted, and it is often reported in mm of H_2O. It is the most commonly employed metric used by water managers to estimate water avialability in the snowpack. Climate change is expected to significantly reduce the maximum mountain SWE, which will change both streamflow and groundwater dynamics, and in turn posing a major challenge for water managers (Siirila-Woodburn et al., 2021). However, since SWE is highly correlated with elevation, and mountainous regions are known for rapid variations in elevation over short distances, mountain snowpack can be difficult to both measure and model. Thus, there is considerable value for both science and society in the development of novel methods for precisely estimating, predicting and projecting snowpack.

Lots of studies have been using ML models for SWE estimation. For instance, Snauffer et al. (2018) used an artificial neural network (ANN) model to estimate SWE from several reanalysis products. Their ML-generated SWE estimation exhibited better agreement with station observations, compared to those derived from a Variable Infiltration Capacity (VIC) hydrological model simulation. Odry et al. (2020) and Ntokas et al. (2021) designed an ANN model to predict SWE and demonstrated that their ML model outperformed the benchmark regression model. Their input variables included snow depth, temperature, accumulated precipitation and several indices such as the number of snow-free days and the number of layers in the snowpack. Random forest methods have also been adopted to bias correct gridded SWE products (King et al., 2020).

To date, ML-based SWE estimation has largely relied on inference or emulation of existing snow-related products, rather than accounting for physical processes that shape snow accumulation. However, recent work by Manepalli et al. (2019) used a conditional generative adversarial network (cGAN) to emulate VIC based estimates of SWE developed by Livneh et al. (2015). They formulated this task as an image-to-image translation problem, where the cGAN model translates gridded relationships between the input meteorological fields to the target SWE field without the needs of snow-related products. Although the cGAN model is demonstrably powerful, this type of image translation task doesn't allow time dependency to be incorporated in the model. Namely, it assumes the SWE at time t can be expressed as a function of meteorological variables at the concurrent time t. Under such an architecture, the model cannot capture temporal features from the input predictors, (i.e., the snow accumulation process is ignored), which is vital for time series prediction.

There have also been recent efforts to estimate SWE based on precipitation (P), temperature (T) and other factors that leverage physical causation and a process-based understanding of the system. That is, new DL models have modeled SWE as an accumulation process by relating SWE to a historic time series of meteorological variables, with the inputs from previous time steps:

$$\mathrm{SWE}_t = f(P_t, P_{t-1}, ..., P_{t-N+1}, T_t, T_{t-1}, ..., T_{t-N+1}, ...,) \tag{3.1}$$

where t denotes the time step and N is the length of the look-back window size. Using the above formula, Meyal et al. (2020) inputted precipitation, temperature, snow depth and SWE from previous days into a long-short term memory (LSTM) model for SWE prediction at five observational stations. They found that the LSTM model can capture the temporal features of snow accumulation and performed well at the selected stations.

Although ML and DL models can achieve satisfying results for historical SWE, models generally struggle with poor performance under extrapolation. Although the LSTM model in Meyal et al. (2020) performed well at the selected observational sites, it was not tested in out-of-sample areas, especially where the statistical properties of SWE accumulation are different from the training sites. This poses a major challenge, particularly if we want to generate a gridded SWE dataset with ML or DL models trained on in-situ observations. Given that in-situ estimates of SWE are generally located in those mountainous areas that are easily accessible and found at mid-elevation, they do not fully represent the areal heterogeneity of SWE at high-elevation or low-elevation that surround the stations. Therefore, a significant extrapolation problem may arise, particularly when applying the ML or DL models to low-elevation plains or valleys. This issue also makes it difficult to validate or calibrate process-based models, which in turn suggests a need for more observations at both low- and high-elevation. In the case of ML-based models, efforts to address the extrapolation problem include transformation of the output target for climate emulation or by evaluating model performance using (extreme) out-of-sample scenarios for streamflow projection (Beucler et al., 2021a; Duan et al., 2020a).

While many studies use ML and DL models for SWE estimation, few have investigated employing these models for long-term projections of SWE loss under climate change. For models with snow-related variables as inputs, such as snow depth, corresponding inputs from climate models are required; however, these inputs are usually provided at relatively coarse resolutions that do not accurately capture mountain landscape heterogeneity. Additionally, the use of models that incorporate information of SWE and snow depth from previous times, as in Meyal et al. (2020), is not practical for future SWE projections. Although one could use this model to integrate through time and substitute observational SWE and snow depth with predicted values, it is a computational expensive process and errors tend to accumulate over time, potentially resulting in a biased projection by the end of the snow season. Ideally, a projection model must be more heavily forced by external meteorological data instead of snow variables.

In our study, we first build three DL models based on equation (3.1), only using the meteorological forcings from 581 observational stations in the western United States (WUS). The model behavior and input sensitivity are subsequently analyzed using an explainable artificial intelligence

(XAI) method. With these trained DL models in hand, we then tackle the spatial extrapolation problem and generate a gridded SWE product over the Rocky Mountains. Finally, we use our model to examine climate change's impact on SWE, using LOCal Analogues (LOCA) down-scaled Coupled Model Intercomparison Project Phase 5 (CMIP5) forcing data. Section 3.2 describes the data sources and XAI method used in our study. Section 3.3 shows the DL model prediction results at observational stations and input sensitivity analysis. Our spatial extrapolation method for generating the gridded SWE product is presented in section 3.4. The SWE response to climate change is explored in section 3.5 and final conclusions are presented in section 3.6.

3.2 Data and Models

3.2.1 Datasets for DL Models

Snow Telemetry (SNOTEL) stations provide daily SWE measurements and are used as the prediction target for the ML model. From the 829 SNOTEL stations (including Alaska), we selected 581 stations across the WUS based on the number of available observations; specifically, the selection criterion was that a station had to provide at least one year of observations, without gaps, over the training period from 1980 to 1990. Meteorological inputs were derived from the gridMET dataset, including daily precipitation, maximum and minimum temperature, solar radiation, maximum and minimum relative humidity, specific humidity, vapor deficit and wind speed on 1/24th-degree grid (approximately 4-km) (Abatzoglou, 2013). Since SNOTEL stations are not located at gridMET grid points, meteorological data nearest to the SNOTEL location is used as the corresponding forcing. The input data are further augmented with static features, including latitude, longitude, elevation, diurnal anisotropic heat index (DAH) (Böhner and Antonić, 2009) and topographic solar radiation aspect index (TRASP) (Roberts and Cooper, 1989). The DAH and TRASP indices are used to account for local scale effects of solar radiation loading on the land surface and snowpack and have been used in Cristea et al. (2017) for snow modeling. The DAH index is a function of topographic slope and topographic aspect, calculated as

$$\mathrm{DAH} = \cos(\alpha_{\max} - \alpha) \times \arctan(\beta) \tag{3.2}$$

where $\alpha_{\max}$ is the aspect receiving the maximum amount of solar radiation (for the WUS, we use $\alpha_{\max} = 1.125\pi$, following Böhner and Antonić (2009)), α is the aspect (in radians), and β is the slope (also in radians). Note that the DAH index ranges between -1 and $+1$, with zero corresponding to flat terrain; for the WUS, DAH is largest on steep southwest-facing slopes that have higher afternoon solar radiation loading and lowest on steep north-facing slopes. The TRASP index is a function of topographic aspect only, namely,

$$\text{TRASP} = \frac{1}{2}\left[1 - \cos\left(\alpha - \frac{\pi}{6}\right)\right], \tag{3.3}$$

where again α is the aspect (in radians). TRASP accounts for daily solar radiation loading and ranges between 0 (for the coolest slopes) and +1 (for the warmest slopes). The calculations of both TRASP and DAH were made using the Parameter-elevation Regressions on Independent Slopes Model (PRISM) 800-meter topographic data (Daly et al., 2008). Similar to the meteorological forcings, the nearest grid cell to the SNOTEL station is used as the corresponding input to the DL model.

For purposes of constructing the DL model, the data are split into training, validation and testing sets. Specifically, we used 1980 Oct 1st to 1999 Sep 30th as the training time period, 1999 Oct 1st to 2008 Sep 30th for validation and 2008 Oct 1st to 2018 Sep 30th for testing. We calculate the mean $\overline{x}$ and standard deviation σ of both the input and output variables from all the SNOTEL stations in the training period to normalize the data via

$$X_{\text{normalized}} = \frac{x_i - \bar{x}}{\sigma}. \tag{3.4}$$

For the historical and future climate scenario projections, the Localized Constructed Analogs (Pierce et al., 2014, LOCA) dataset is used as our meteorological forcing. LOCA was applied to 32 global climate simulations from the CMIP5 ensemble and downscaled to 1/16th degree resolution over North America. We selected CESM-CAM5, CNRM-CM5, EC-EARTH, GFDL-ESM2M, HadGEM2-ES and MIROC5 to analyze the climate change impact on SWE over the Rocky Mountains. These climate models have been developed at different agencies across the globe, and represent a diversity of model subgrid parameter and structural differences that each provide largely independent projections of climate change impacts. A complete analysis of all LOCA models is practicable but not necessary given the similar projected snowpack patterns presented in the following section.

To compare our ML model performance to a modern process-based modeling data product, we employ the daily 4-km gridded SWE data from Zeng et al. (2018), which uses PRISM precipitation and temperature data and assimilates SNOTEL observations. This study partitioned rainfall and snowfall using daily 2-m air temperature thresholds, which were derived from station observations. When interpolating point measurements to other grids, instead of the absolute SWE measurements, they used the ratio of SWE observations over estimated net snowfall. The detailed methodology and analysis can be found in Broxton et al. (2016) and Zeng et al. (2018). This product is used as reference for the DL model SNOTEL predictions and extrapolation across the Rocky Mountains for the time period from 2008 to 2018. Since it has been developed at the University of Arizona and can be accessed through National Snow and Ice Data Center (NSIDC), this product is referred to as the NSIDC-UA dataset in the remainder of the paper.

3.2.2 Deep Learning Models

In this study we focus on deep learning models that can be employed for time series problems. Three different DL models (LSTM, TCNN and Attention) are investigated and compared, following the general framework depicted in Figure 3.1. Under this design, the temporal block extracts temporal features from the input data, while the dense layer generates a single step prediction. The DL models are trained to minimize an objective function (i.e., the loss function), which in this study is chosen to be the mean squared error (MSE). We set the number of training periods (epochs) to 50 and the optimization algorithm as Adam with a learning rate of 1e-4 (Kingma and Ba, 2014). The remainder of the hyperparameters associated with the model architecture are then determined by grid search and the searching candidates can be found in the supplement (Table S1). Although the hyperparameters are important for the model performance, they are not fine tuned because of the computational power required and because such tuning would not substantially improve our present

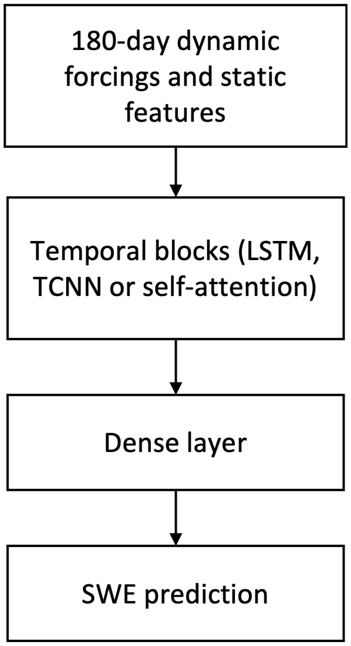
180-day dynamic forcings and static features
Temporal blocks (LSTM, TCNN or self-attention)
Dense layer
SWE prediction

Figure 3.1: Our framework for the ML models employed in this study.

results. The DL models are all developed with PyTorch (Paszke et al., 2019). The remainder of this section introduces each model and its corresponding model architecture.

3.2.3 Performance Metrics

Model performance is quantified using the Nash-Sutcliffe model efficiency coefficient (NSE), a widely used metric for hydrological model evaluation (Nash and Sutcliffe, 1970). It is defined via

$$\mathrm{NSE}(O_t, P_t) = 1 - \frac{\sum(O_t - P_t)^2}{\sum(\overline{O}_t - O_t)^2}, \tag{3.5}$$

where O and P denote observations and predictions, respectively. Index t denotes the time and $\overline{O}_t$ is the observation mean. NSE is in the range $(-\infty, 1]$, with larger values indicating better performance and a score of 1 indicating a perfect match between model and observations. In the context of regression, NSE is equivalent to the coefficient of determination (or R^2 score), a common metric for model accuracy assessment defined via

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}, \tag{3.6}$$

where RSS is the residual sum of squares and TSS is the total sum of squares. Therefore, in a regression model, RSS and TSS can be expressed as the numerator and denominator in the NSE formula and thus NSE is equivalent to R^2 score. It should be noted that an R^2 score is not the square of the correlation coefficient and it can be negative when RSS is larger than TSS.

We also assess the model performance with root mean squared error (RMSE) and mean absolute error (MAE), defined via

$$\mathrm{RMSE}(O_t, P_t) = \sqrt{\frac{\sum (O_t - P_t)^2}{n_{\mathrm{samples}}}}, \tag{3.7}$$

$$\mathrm{MAE}(O_t, P_t) = \frac{1}{n_{\mathrm{samples}}} \sum |O_t - P_t|. \tag{3.8}$$

where n_{samples} is the number of evaluated samples. RMSE and MAE are in the range $[0, +\infty)$ with lower values indicating a closer match, and a score of 0 indicating a perfect match between model and observations.

3.2.4 Feature Permutation, Sensitivity and Interpretation

Although ML models generate predictions, they are frequently referred to as 'black box' models since it is often unclear why and how the model produces its results. Recent advances in explainable AI (XAI) have enabled better interpretation of ML model results, especially in Earth system modeling (Barnes et al., 2020; Gagne II et al., 2019; McGovern et al., 2019; Toms et al., 2020).

In this study, we used a permutation-based method to quantify the relative importance of input variables in the DL models (Breiman, 1996). The permutation method evaluates the DL model through testing samples to obtain a baseline performance score. Then each feature is permuted to generate a shuffled dataset, and a new performance score is subsequently calculated. Details of the permutation method are given in section 3.3.2. The change in the performance score represents the importance of a given feature. A greater decrease in model skill corresponds to higher feature importance. Notably, this same approach has been applied in previous research addressing model interpretation (Gagne II et al., 2019). Although this method can reveal the relative importance of individual input features, the quantified performance is potentially confounded by correlation among input features. For example, a model that uses both mean and maximum daily temperature as input may see minimal performance loss from removal of either of these features while removal of both would be significant. Efforts to address correlation issues include the use of multi-pass permutation, as discussed in a review by McGovern et al. (2019). In this study, we permute both the training and the testing set and train a reduced model. By permuting the training set and retraining the model, the permuted variable is blocked and the reduced model only receives the information from the remaining non-permuted variables. The importance of the permuted variable will be quantified via a decreased ratio in the NSE value compared with the baseline score.

3.2.5 Ensemble Training

Since DL models generally use gradient-based optimization methods to update the weights, different initial weights are likely to generate different models. To reduce this effect, each model is trained 10 times with different initial weights to generate an ensemble of predictions. As shown in Wang et al. (2021) the use of ensembles can increase the prediction accuracy. In the following sections, all predictions are based on the ensemble mean from the DL models. The training time for each DL model is as follows with a single RTX 2080TI GPU: 5 hours to train the LSTM model, 10 hours for the TCNN model, and 26 hours for the Attention model. Although the training time varies among different DL models, all DL models only need to be trained once, and inference time is much shorter. For our application over the Rocky Mountains, it takes approximately 35 minutes to generate a 10 year prediction of SWE on the 168 × 108 grid covering the Rocky Mountains with parallel execution on a single RTX 2080TI GPU for either the LSTM and TCNN. The Attention model takes longer than the other DL models to generate a prediction – approximately 2.5 hours without parallel execution – and is limited by the GPU memory. Faster performance is anticipated using GPUs with larger memory. Given the much faster inference time, the training time is not

Table 3.1: Tabulated model performance for prediction of SWE at SNOTEL stations. The top table shows performance scores on dates when observed SWE is greater than zero, while the bottom table shows the whole evaluation period (from 2008 to 2018). The best scores for each metric are shown in bold font.

Nonzero SWE	Median NSE	Median MAE (in)	Median RMSE (in)
LSTM	**0.841**	**1.621**	**2.415**
TCNN	0.793	2.042	2.858
Attention	0.779	2.049	2.837
NSIDC-UA	0.755	1.959	2.875
Whole period	Median NSE	Median MAE (in)	Median RMSE (in)
LSTM	**0.912**	**0.962**	**1.832**
TCNN	0.879	1.236	2.209
Attention	0.874	1.187	2.106
NSIDC-UA	0.861	1.051	2.111

used for model selection or intercomparison.

3.3 Results for SNOTEL stations

3.3.1 SNOTEL Prediction Results

DL model performance is evaluated over the testing period (2008-10-01 to 2018-09-30) at all SNOTEL stations. Since SWE is zero at most SNOTEL stations in the summertime, the DL model is assessed when incorporating daily information for both the entire testing period and only those days in the testing period when SWE is nonzero. For each SNOTEL station's SWE prediction, we take the mean ensemble predictions for each DL model. Table 3.1 shows the median NSE across all SNOTEL stations. The LSTM has the highest median NSE value, followed by the TCNN and then the Attention model. The distribution of NSE values is shown in Figure 3.2. The LSTM cumulative distribution function (CDF) line lies above the TCNN and Attention CDF lines, showing there are more station SWE predictions with higher NSE values compared with TCNN and Attention. When only comparing the TCNN and Attention models, the TCNN has a slightly higher median performance, although the Attention model shows better overall performance.

In addition to intercomparing the DL models, we also compared prediction accuracy with the NSIDC-UA dataset over the same testing period, assessed using the nearest grid point to the

SNOTEL station. The median NSE value over the 581 SNOTEL stations for the whole period is 0.861 and 0.755 for the period with nonzero SWE. Under this metric, all DL models achieve better results than the NSIDC-UA dataset. Moreover, when comparing the distribution of NSE values in Figure 3.2, we see that the DL models have more stations with higher NSE values. As for MAE and MSE metrics, the UA dataset outperforms some of the DL models. It achieves a lower median MAE than TCNN and Attention for both the whole testing period and the nonzero-SWE period. In Figure 3.2, the NSIDC-UA dataset attains the lowest MSE and MAE across 20% to 30% of SNOTEL stations. Figure 3.3 illustrates the NSE distribution with respect to elevation. In general, all DL models perform better at higher-elevation. Although the NSIDC-UA dataset also generally shows improved performance with elevation, there is an anomalous dip in performance between 3026m and 3341m. These performance results are similar for maximum SWE in Figure 3.4. On the other hand, all DL models exhibit steadily increasing performance with the maximum SWE amount, while the NSIDC-UA performance drops when maximum SWE is larger than 63in. These results suggest that, for western US SNOTEL stations, DL models can generally produce comparable results with the UA dataset. The LSTM model provides the most accurate predictions across the three DL models assessed. A prediction example for a selected SNOTEL station can be found in Figure S1.

As DL models are purely data-driven, their strong performance in this context is perhaps unsurprising. Notably, they are not constrained by physical conservation laws, and so are able to internally learn and correct for biases in the observational data. Additionally, the NSIDC-UA dataset is an estimation of SWE values on a grid, while the DL models are optimized for prediction of SWE at the SNOTEL stations themselves. This interpolation error is investigated further when our DL models are extrapolated to the Rocky Mountains. We find that nearest-neighbor interpolation is likely a major factor in the weaker performance of the NSIDC-UA dataset at SNOTEL stations, and further find that the more interpolations that are performed, the worse the resulting NSE values. Details of this analysis can be found in the supplement.

Although different DL models yield different NSE values, the spatial distribution of their performance tends to be similar – that is, stations with lower (higher) NSE values in one model tend

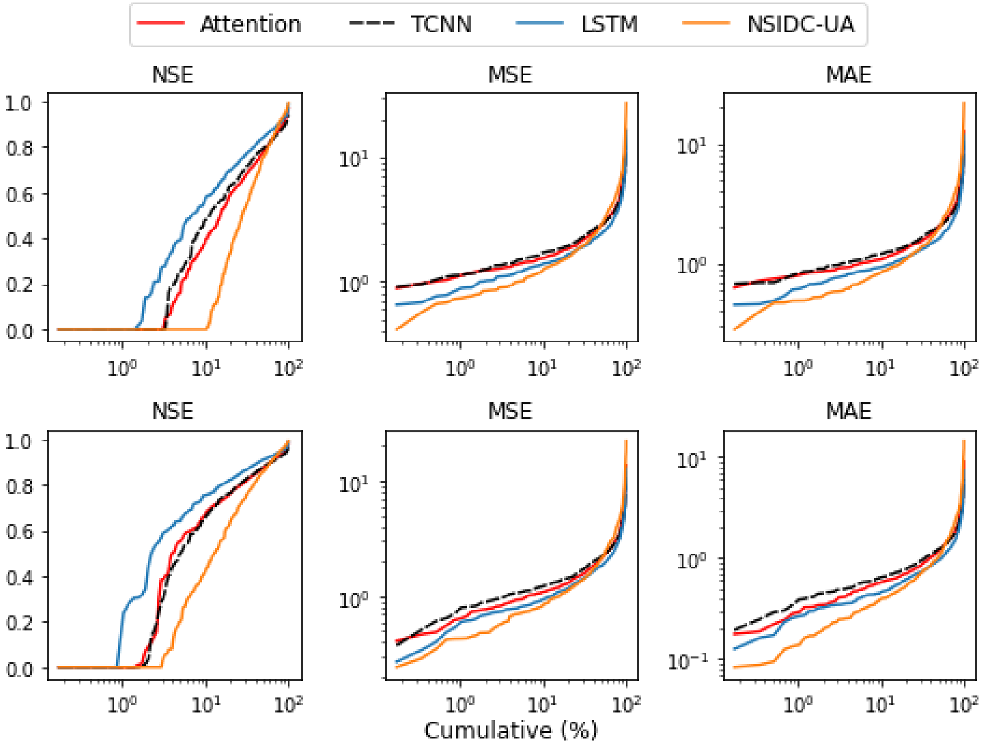

Figure 3.2: Cumulative distribution functions across three model performance metrics for predicting SWE across SNOTEL stations in the western US. NSE values are truncated at 0. The upper row shows model performance when only those SWE observations larger than zero are used. The lower row shows model performance for all zero and nonzero SWE observations.

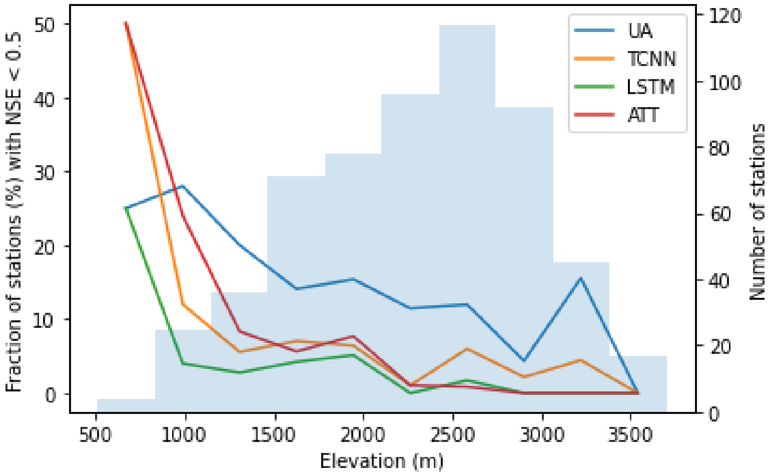

Figure 3.3: NSE values distribution with respect to elevation. The left y-axis denotes the fraction of stations with NSE values less than 0.5 for each elevation bin. The right y-axis shows the number of SNOTEL stations in each elevation bin.

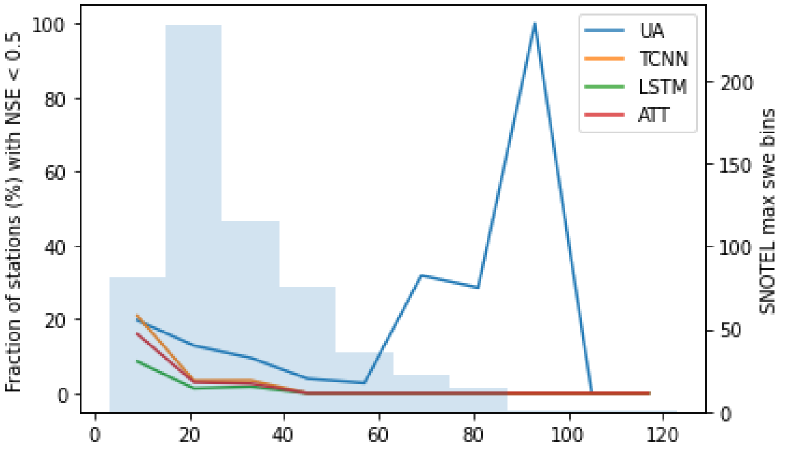

Figure 3.4: NSE values distribution with respect to maximum SWE measurement. The left y-axis denotes the fraction of stations with NSE values less than 0.5 for each SWE bin. The right y-axis shows the number of SNOTEL stations in each SWE bin.

to have lower (higher) NSE values in other models. As shown in Figure 3.5, all DL models exhibit relatively poor performance (i.e., negative NSE values) in Western Washington, Northern Nevada, Southern and Northwestern Oregon, and Northern Montana. Indeed, Figure S2 shows a strong correlation between NSE values across the DL models (Pearson correlation values of 0.945 and 0.818 for TCNN and Attention with respect to LSTM). This implies that the NSE value distributions are relatively uniform among stations in spite of differences in DL models. In general, these stations tend to have a lower maximum SWE than other stations, which is consistent with our earlier attribution of model performance (Figure 3.4).

3.3.2 Permutation-based Interpretation of Deep Learning Models

As the LSTM model produces the highest NSE values at SNOTEL stations, hereafter it will be the focus of our investigation. In this section the LSTM is studied using a permutation analysis. Permutation analysis is an explainable AI (XAI) technique that identifies those quantities that are most important for model performance. The permutation analysis consists of augmenting one input variable (to generate a new shuffled dataset) and training a new set of LSTM models. The importance of each input variable is quantified by assessing the ratio of the permuted LSTM model NSE prediction to the LSTM model baseline NSE prediction, which is trained and tested with the

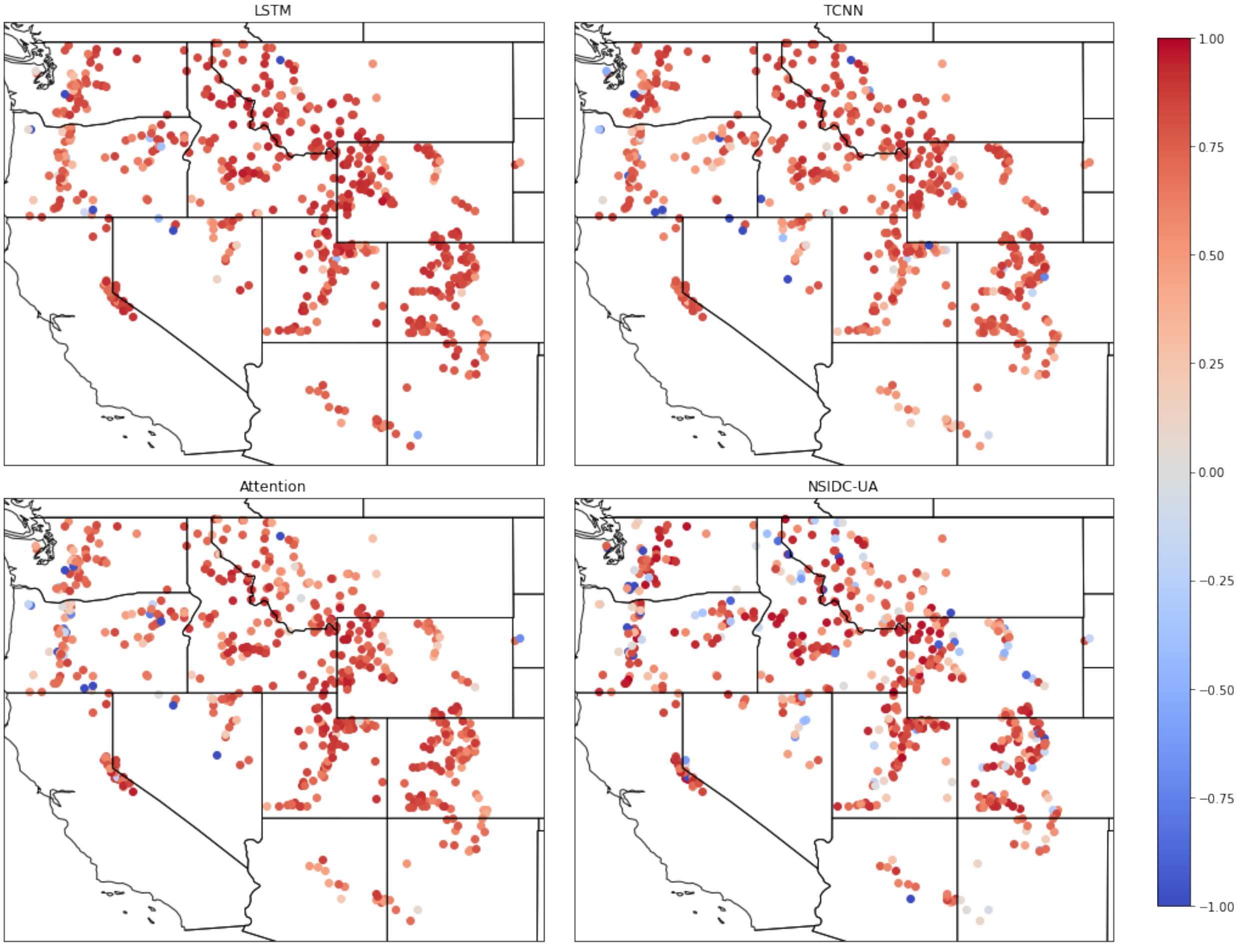

Figure 3.5: SWE prediction performance from the DL models and NSIDC-UA dataset. Dots represent individual SNOTEL stations, with the color of the dot representing the NSE value (truncated at -1).

non-permuted dataset. The permuted LSTM models are also trained 10 times to build an ensemble of predictions. To quantify model uncertainty, we use bootstrap sampling to provide results with a 90% confidence interval, as shown in Figure 3.6. For meteorological inputs, the time series from each grid point is used for re-sampling so that the statistical properties of these variables are preserved (i.e., only the time steps are shuffled). For static features the permutation is performed among all stations. In Figure 3.6, the green bars represent the performance decline for each static variable, the blue bars represent the performance decline from individual meteorological variables, and the orange bar represents the combined effect of all static variables.

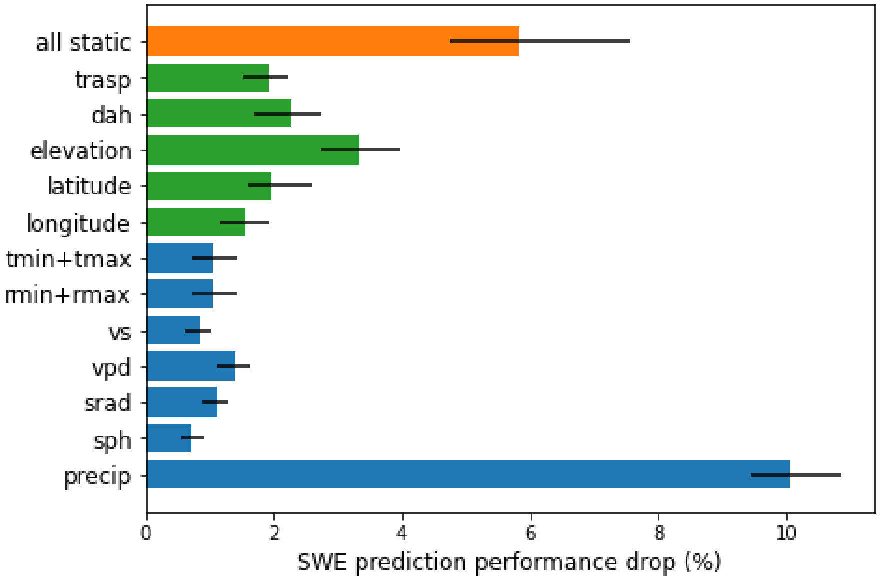

Figure 3.6: SWE prediction performance drop (%) quantified using the NSE values among the permuted estimates. Error bars represent the 90% confidence interval from bootstrap sampling. Precipitation is abbreviated to precip, sph stands for specific humidity, srad for solar radiation, vpd for vapor deficit, and vs for wind speed. Tmin, tmax, rmin, and rmax refer to minimum and maximum temperature and relative humidity.

The input variable with the most influence on SWE prediction was precipitation, followed by elevation, while the rest of the input variables had comparable influence. This result agrees with the intuition that precipitation provides water mass to build snowpack, and precipitation type is determined by temperature which is shaped by elevation via the lapse rate. We also investigated the importance of static inputs to the LSTM model. The combined effect of static features are critical, as the LSTM model accuracy would drop 6% without their inclusion, which is more than half the influence of precipitation. Although the meteorological variables are responsible for the temporal variability in SWE amount, the LSTM model gains valuable spatial information from the static features. These features are useful for modulating snowpack dynamics at each SNOTEL station. The utility of static variables was also reported in Kratzert et al. (2019b), where a LSTM was used to pridict streamflow.

Among all the static variables, there were three categories: location (latitude and longitude), aspect and slope (DAH and TRASP), and elevation. We combined each static variable into these categories during the permutation process to compare their relative importance. Relative to the baseline LSTM model with median NSE of 0.912, the LSTM model that did not include location information had the highest median NSE score (0.885), followed by aspect and slope (0.877), and elevation (0.869). This result again emphasizes the critical role that elevation plays in SWE prediction since it can determine the temperature and rain-snow partitioning of precipitation. Although temperature is also affected by latitude through differences in solar loading, the LSTM model benefited more from information related to aspect and slope, which have more localized effects on temperature.

It should be noted that the relative importance of variables other than precipitation can not be directly determined from the permutation test. This is especially true for temperature, since vapor deficit is a function of temperature and relative humidity. The temperature signals can be inferred from vapor deficit and relative humidity even if we permute temperature. To better compare their relative influence on SWE predictability, we instead built several reduced order LSTM models. In each reduced order model, precipitation and one of the other meteorological variables are used, and the remaining variables are permuted. The baseline for comparison was an LSTM model with only precipitation and the reference was a model with the full set of meteorological variables. The model using precipitation plus relative humidity was not included in this analysis because it did not converge to a reasonably performant model. As shown in Table 3.2, among the reduced order models, precipitation and wind speed give the lowest NSE value, although the use of wind speed and precipitation does improve the SWE prediction skill tremendously compared with the baseline model. The median NSE scores across the rest of the reduced order models are all above 0.8, and the combination of temperature and precipitation achieves the closest performance to the reference model. This indicates that vapor pressure deficit, solar radiation, and specific humidity contain influential information for SWE prediction, yet temperature is the most critical variable that influences SWE prediction besides precipitation.

To determine the best third variable in the model, five additional models were trained: each model consists of precipitation, temperature, and one other variable. The results are shown in Table 3.2. The model with precipitation, temperature and relative humidity attains the lowest NSE value, which is consistent with previous anomalous low performance with precipitation and relative humidity. The combination of precipitation, temperature and vapor deficit obtain a median NSE value of 0.888, which is identical to the model with only precipitation and temperature. This result suggests that the vapor deficit information is largely encoded in the precipitation and temperature fields. The inclusion of wind speed and specific humidity both increase the model performance, probably because wind speed and specific humidity cannot be inferred from precipitation and temperature. Adding these two variables indeed injects independent information to the LSTM model. The model with precipitation, temperature and solar radiation obtained the highest median NSE value of 0.893, or 97% of the full model performance. With far fewer input variables, the model with precipitation, temperature and solar radiation was capable of capturing the necessary temporal features for SWE prediction. This again highlights the important roles that these three variables have in affecting the dynamics of mountain snowpack. Additionally, this result suggests that good estimates of snowpack can be obtained from datasets providing these quantities, such as CAMELS (Addor et al., 2017).

Given that we used both minimum and maximum temperature as inputs into the LSTM model, we also apply the permutation analysis to investigate the importance of including the diurnal signal of temperature on SWE prediction versus daily average temperature. We did this by comparing the performance of a model with only average temperature versus a model with both minimum (night) and maximum (daytime) temperatures. Table 3.3 shows the influence of accounting for the range of diurnal temperature versus daily average conditions. Generally the minimum and

Table 3.2: A comparison of several reduced order LSTMs for predicting total SWE. Precipitation is abbreviated to Precip.

	First Quantile NSE	Median NSE	Third Quantile NSE
Precipitation Only	0.143	0.385	0.600
Precip+Wind Speed	0.427	0.612	0.734
Precip+Specific Humidity	0.727	0.834	0.898
Precip+Solar Radiation	0.720	0.834	0.906
Precip+Vapor Deficit	0.748	0.838	0.894
Precip+Temperature+Relative Humidity	0.812	0.887	0.931
Precip+Temperature	0.820	0.888	0.930
Precip+Temperature+Vapor Deficit	0.819	0.888	0.931
Precip+Temperature+Wind Speed	0.815	0.890	0.929
Precip+Temperature+Specific Humidity	0.820	0.892	0.932
Precip+Temperature+Solar Radiation	0.811	0.893	0.937
Full Model	0.847	0.912	0.942

Table 3.3: LSTM model performance in predicting SWE with daily average temperature versus diurnal temperature range.

	First Quantile NSE	Median NSE	Third Quantile NSE
Average temperature	0.802	0.884	0.924
Min & max temperature	0.820	0.888	0.930

maximum temperature can provide more information and achieve a higher NSE score, suggesting the inclusion of the diurnal signal influences the SWE prediction skill of the LSTM model. However the improvement from adding the diurnal signal is very limited relative to only including daily average temperature (0.4% difference for the median NSE score). This result is consistent with Kirkham et al. (2019) who showed that snow cover can minimize the diurnal surface temperature range which can, in turn, influence the daily change in SWE through influences on rain-snow partitioning in the accumulation season and above freezing conditions that lead to ablation of the snowpack.

3.4 Spatial extrapolation of SWE to the Rocky Mountains

Although the DL models are capable of predicting daily SWE at individual SNOTEL stations, we demonstrate here that these models can also be applied to out of sample areas (as is done with the process-based NSIDC-UA dataset), even when there are out-of-sample differences in the statistical

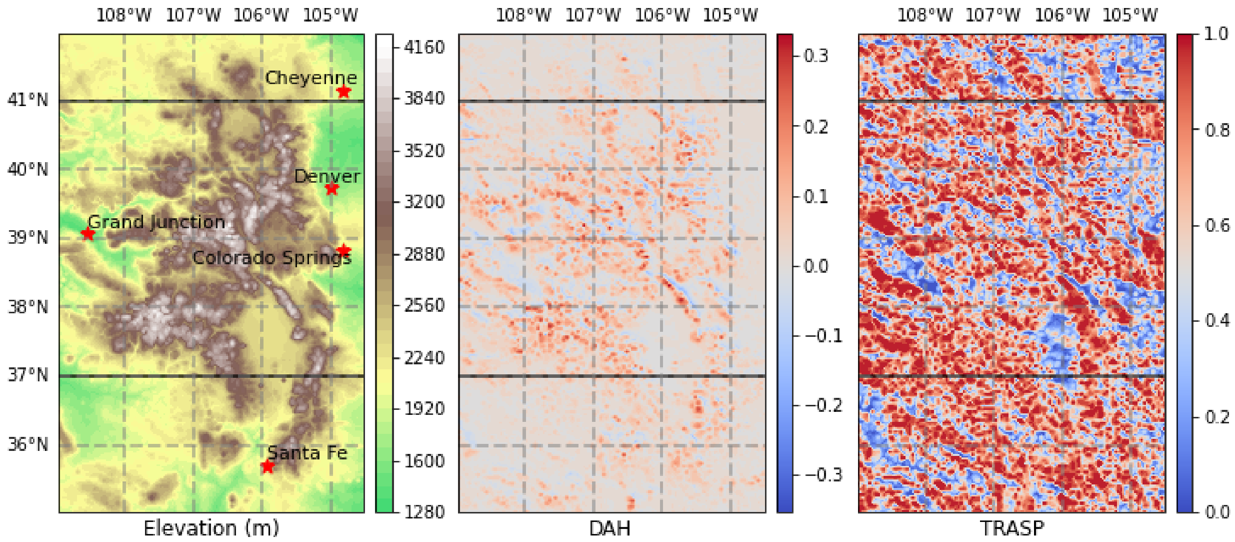

Figure 3.7: Estimates of elevation, DAH and TRASP over the Rocky Mountains derived from the 800m PRISM dataset.

properties of the DL models' input and output variables. The use of these models outside of their training sample is a common problem referred to within the machine learning community as concept drift or extrapolation (Tsymbal, 2004). As shown in Balestriero et al. (2021), for high-dimensional datasets, even if all the components of the vectors are in the training range, the high-dimensional input vector can still be out of the convex hull of the training set. In our case, extrapolation is expected to be even more common since some input variables are out of the training range (e.g., many grid points have elevations lower or higher than the lowest or highest SNOTEL station). Nonetheless, we demonstrate here that the DL models employed in this out-of-sample manner are still able to accurately predict SWE amount.

In this section, we use our DL models to generate a gridded SWE product with 4km grid spacing over the Rocky Mountains (Figure 3.7). The 4km grid spacing is inherited from the gridMET data, which is used for forcing, although a similar approach could be used to produce an even higher resolution product (e.g., one matching the 800m PRISM product). The simulation period spans 2008-10-01 to 2018-09-30, which is the same period as the SNOTEL testing set. The gridMET data is regridded to the NSIDC-UA gridpoint for better comparison. A nearest neighbor method is used for regridding: the nearest gridMET datapoint is assigned to each NSIDC-UA gridpoint as the

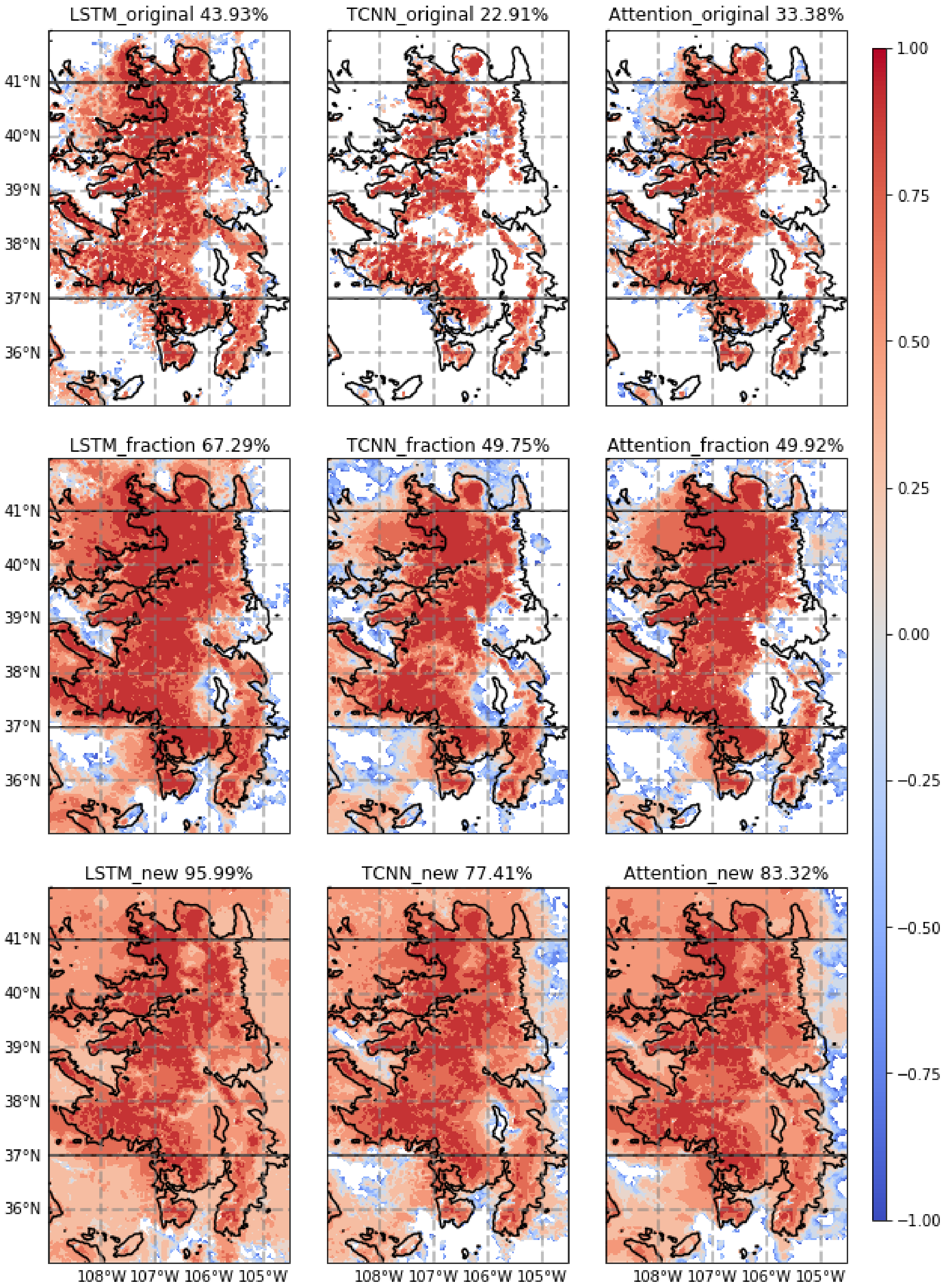

Figure 3.8: NSE values for DL model extrapolation estimates over the Rocky Mountains. The top row shows the original DL model SWE prediction versus NSIDC-UA. The middle row is the SWE fraction evaluation from the original models, computed via equation (3.10). The bottom row represents a new set of DL models that predict SWE fraction, computed via equation (3.12). NSE values below -1 are masked in all figure subpanels. The black line is the 2300 meter contour. The percentage value given in the title is the fraction of grid points with positive NSE values.

corresponding forcing. When applied over the Rocky Mountains, the gridMET forcing variables are normalized with the mean and standard deviation from the training SNOTEL stations (equation 3.4). The DL model SWE prediction is then transformed back to its original units with the same equation. The top row in Figure 3.8 shows the NSE values obtained from the DL-generated dataset (10 ensemble member mean) when using the NSIDC-UA dataset as ground truth across the Rocky Mountains. That is,

$$\mathrm{NSE}_{\mathrm{original}} = \mathrm{NSE}(\text{SWE-UA}, \text{SWE-DL}), \tag{3.9}$$

where SWE-UA denotes the grid point SWE from the NSIDC-UA dataset (taken as observations) and SWE-DL denotes the grid point SWE from the deep learning model. The DL model estimates largely agree with the process-based estimates in high-elevation areas, while performance is relatively poor in low-elevation areas (see Figure 3.7). However, this performance discrepancy in low-lying regions is partially attributed to greater sensitivity of NSE in regions with lower SWE. Further, as shown earlier, both DL models and NSIDC-UA exhibit relatively poor performance at lower elevations and when SWE amounts are low – indeed, the ground truth in these regions is poorly constrained given a dearth of relevant measurements. These factors make it difficult to quantify how much error may be attributed to out-of-sample application of the DL model.

Assuming NSIDC-UA is correct, the relatively poor performance of the model in low-elevation areas may be due to a biased magnitude estimation (i.e., too little or too much SWE) or because of errors in temporal dependency (i.e., too slow/rapid accumulation/melt). To mitigate issues related to magnitude estimation, we assess model performance using the fraction of maximum SWE (i.e., SWE/ max SWE), where the maximum is with respect to the historical/training period. That is, the NSE from snow fraction is calculated via

$$\begin{aligned}\mathrm{NSE}_{\mathrm{fraction}} &= \mathrm{NSE}(\text{SWE-UA}, \frac{\text{SWE-DL}}{\max \text{SWE-DL}} \times \max \text{SWE-UA}) \qquad (3.10)\\ &= \mathrm{NSE}\left(\frac{\text{SWE-UA}}{\max(\text{SWE-UA})}, \frac{\text{SWE-DL}}{\max(\text{SWE-DL})}\right). \qquad (3.11)\end{aligned}$$

Equations (3.10) and (3.11) are equivalent since the NSE value is unaffected when the predictions and observations are multiplied or divided by the same constant. Importantly, the fraction of maximum SWE is constrained to the range $[0, 1)$ at all grid points, regardless of elevation. This allows us to better equalize the importance of each grid point in the loss calculation; we hypothesize this choice will improve model performance in low elevation regions. By using the SWE fraction, differences in SWE magnitude between the NSIDC-UA dataset and DL model are mitigated and so the evaluation emphasizes the tendency of the SWE. The middle row in Figure 3.8 shows the assessment with SWE fraction. Under this metric, the DL model appears significantly better, with higher NSE values almost everywhere and a larger portion of positive NSEs. This difference indicates that while the DL models can capture the temporal dependence of SWE, magnitude biases are large over low-elevation areas.

Given the improvement in model performance when using SWE fraction, another set of DL models was trained on SNOTEL data to output SWE fraction, rather than the SWE itself (with predictions hereafter referred to as SWE-DL-FRAC). NSE values are then computed via

$$\begin{aligned} \mathrm{NSE}_{\mathrm{new}} =& \mathrm{NSE}(\text{SWE-UA}, \text{SWE-DL-FRAC} \times \max(\text{SWE-UA})) && (3.12) \\ =& \mathrm{NSE}\left(\frac{\text{SWE-UA}}{\max \text{SWE-UA}}, \text{SWE-DL-FRAC}\right). && (3.13) \end{aligned}$$

The bottom row in Figure 3.8 shows the NSE result when predicting SWE fraction directly from the DL models. When comparing the original DL models, which predicted absolute SWE, against the DL models trained to predict SWE fraction, there is a clear and significant improvement almost everywhere in the domain, but particularly in low elevation regions. Compared to the original models assessed with SWE fraction, this new set of models still attain better performance (as indicated by more grid points with positive NSE values). This result shows that normalization of maximum SWE is effective regardless of the DL model chosen. Among all the DL models, the LSTM-based model again provides the best overall SWE prediction, determined by the larger number of positive NSE values. Of course, to transform the fraction of maximum SWE back to a SWE value, the historical maximum SWE is needed within each grid cell. But since SNOTEL observations are unevenly distributed throughout the Rocky Mountains, we must rely on an alternative estimate of maximum SWE at each grid point; in this case, we use the historical maximum SWE values from the NSIDC-UA dataset at each grid point over the training period to estimate maximum SWE. An example of annual maximum SWE prediction from this new LSTM model is shown in Figure S3.

3.4.1 Interpolation Errors

As mentioned in section 3.3.1, some of the error in the NSIDC-UA dataset observed when benchmarking at SNOTEL sites may be attributed to interpolation errors from the nearest gridpoint method. With the LSTM extrapolation results, we further compared the grid predictions against in-situ observations. 105 SNOTEL stations inside the Rocky Mountain area are selected. Four predictions are evaluated here: in-situ LSTM predictions (LSTM-in-situ), LSTM SWE predictions (LSTM-extra-SWE), LSTM SWE fraction predictions (LSTM-extra-fraction) and NSIDC-UA estimations. LSTM-in-situ predictions are from the LSTM model that output SWE fraction. This is the new model trained with gridMET forcing that are closest to SNOTEL stations in previous section. With SWE fraction as output, it is transformed back with the historical maximum SWE at each station. LSTM-extra-fraction and LSTM-extra-SWE are both from the same LSTM model. Unlike the LSTM-in-situ predictions, the inputs to LSTM-extra-fraction and LSTM-extra-SWE are from gridMET points that are closest to the NSIDC-UA gridpoints, rather than the SNOTEL stations. Further, the LSTM-extra-SWE uses historical maximum SWE estimations from the NSIDC-UA to transform from SWE fraction to SWE. Among these four predictions, LSTM-in-situ is the only one using the SNOTEL station inputs, and the rest are all gridded products. Gridded predictions nearest to the SNOTEL stations are taken from LSTM-extra-SWE, LSTM-extra-fraction and NSIDC-UA, and assessed against SNOTEL observations. As shown in Table 3.4, the in-situ LSTM model has the highest NSE value, and lowest MAE and MSE. This again suggests that gridded estimations are not always representative of observational stations, and the estimations towards SNOTEL stations give the best accuracy. Among the gridded predictions, LSTM-extra-SWE is worse than NSIDC-UA and LSTM-extra-fraction with lower NSE values. NSIDC-UA and LSTM-extra-fraction have similar performance: NSIDC-UA is better over the nonzero-SWE period, while

Table 3.4: Tabulated model comparison for prediction of SWE at SNOTEL stations in the Rocky Mountains. The top table shows performance scores on dates when observed SWE is greater than zero, while the bottom table shows the whole evaluation period (from 2008 to 2018). The best scores for each metric are shown in bold font.

Nonzero SWE	Median NSE	Median MAE (in)	Median RMSE (in)
LSTM-in-situ	**0.798**	**1.759**	**2.440**
NSIDC-UA	0.775	1.775	2.537
LSTM-extra-fraction	0.761	-	-
LSTM-extra-SWE	0.692	2.183	2.948
Whole period	Median NSE	Median MAE (in)	Median RMSE (in)
LSTM-in-situ	**0.885**	**1.062**	**1.883**
NSIDC-UA	0.861	1.010	1.949
LSTM-extra-fraction	0.864	-	-
LSTM-extra-SWE	0.820	1.428	2.313
Inputs data			
LSTM-in-situ	gridMET nearest to SNOTEL stations, max SWE from SNOTEL stations		
LSTM-extra-fraction	gridMET nearest to NSIDC-UA grids		
LSTM-extra-SWE	gridMET nearest to NSIDC-UA grids, NSIDC-UA max SWE		

LSTM-extra-fraction outperforms when assessed in the whole testing period. It is worth noting that LSTM-in-situ, LSTM-extra-fraction and LSTM-extra-SWE are from the same model. The only difference is in their input data. LSTM-in-situ uses the nearest gridMET to the SNOTEL stations, which yields the best performance. As we mentioned before, the gridMET forcings for LSTM-extra-fraction are from the nearest gridpoints to the NSIDC-UA gridpoints. This regridded gridMET may not be representative of SNOTEL stations. As for LSTM-extra-SWE, similar with LSTM-extra-fraction, it suffers from the spatial mismatch between regridded gridMET and SNOTEL stations. Additionally, although we regridded gridMET to NSIDC-UA grids, the mismatch between these two datasets are not negligible, which would further aggravate the spatial error when it uses the historical maximum SWE for the transformation. In this case, with more nearest regridding operations, we would get lower NSE scores. This performance deterioration with nearest regridding proves the gap between in-situ observations and gridded datasets, and suggests the necessity of more advanced regridding algorithms.

3.5 Future projections with CMIP5/LOCA Data

Previous work by Rhoades et al. (2018c) suggests a dramatic decline in mountain snowpack across the Western US by the end of the 21st century. However, the grid spacing employed in this study is only 28km, which is largely unable to capture the most rugged topography of the Rocky Mountain region. As the DL models outputting SWE fraction demonstrate good performance over the Rocky Mountains, in this section we will employ these models for future projections of SWE in this region at high spatial resolution.

The LSTM model is selected for our study as it has consistently produced the best results among all DL models. As in Duan et al. (2020a), meteorological forcings for the DL model projections are provided by the LOCA dataset. Specifically, we use six of the downscaled climate simulations: CESM-CAM5, CNRM-CM5, EC-EARTH, GFDL-ESM2M, HadGEM2-ES, and MIROC5. LOCA also provides SWE projection estimates from the VIC model that use the same downscaled meteorological variables as forcings. These VIC-based projections were used as comparisons to our LSTM-based projections. The Representative Concentration Pathways (RCP) 8.5 future projections over the period 2071 to 2091 are used and compared with the historical simulated period of 1981 to 2001. So as to communicate more management-relevant climate change impacts related to the seasonal cycle of SWE, we focus our analysis on metrics from Rhoades et al. (2018a) and Rhoades et al. (2018b). Brief descriptions of these metrics are provided in Table 3.5. Since the extrapolation is only valid for the SWE fraction, we continue to use the historical maximum SWE from the NSIDC-UA dataset to transform back to absolute SWE. Although the NSIDC-UA dataset is necessary for the transformation, the metrics associated with the percentage of maximum SWE (i.e. SAD, SPD, CMD) can be calculated directly with SWE fractions. The changes in SAD, SPD, CMD and ASL, MSL, SSL are calculated as the the future values minus historical values. The snowpack peak change is shown in both SWE differences and SWE change ratio.

Historical simulation results from the LSTM model and LOCA-VIC are first compared with the NSIDC-UA dataset. The first row in Figure 3.9 shows the historical SSL, SAD, CMD, SPD and peak SWE from the NSIDC-UA dataset. Simulations from VIC and LSTM forced by CESM output

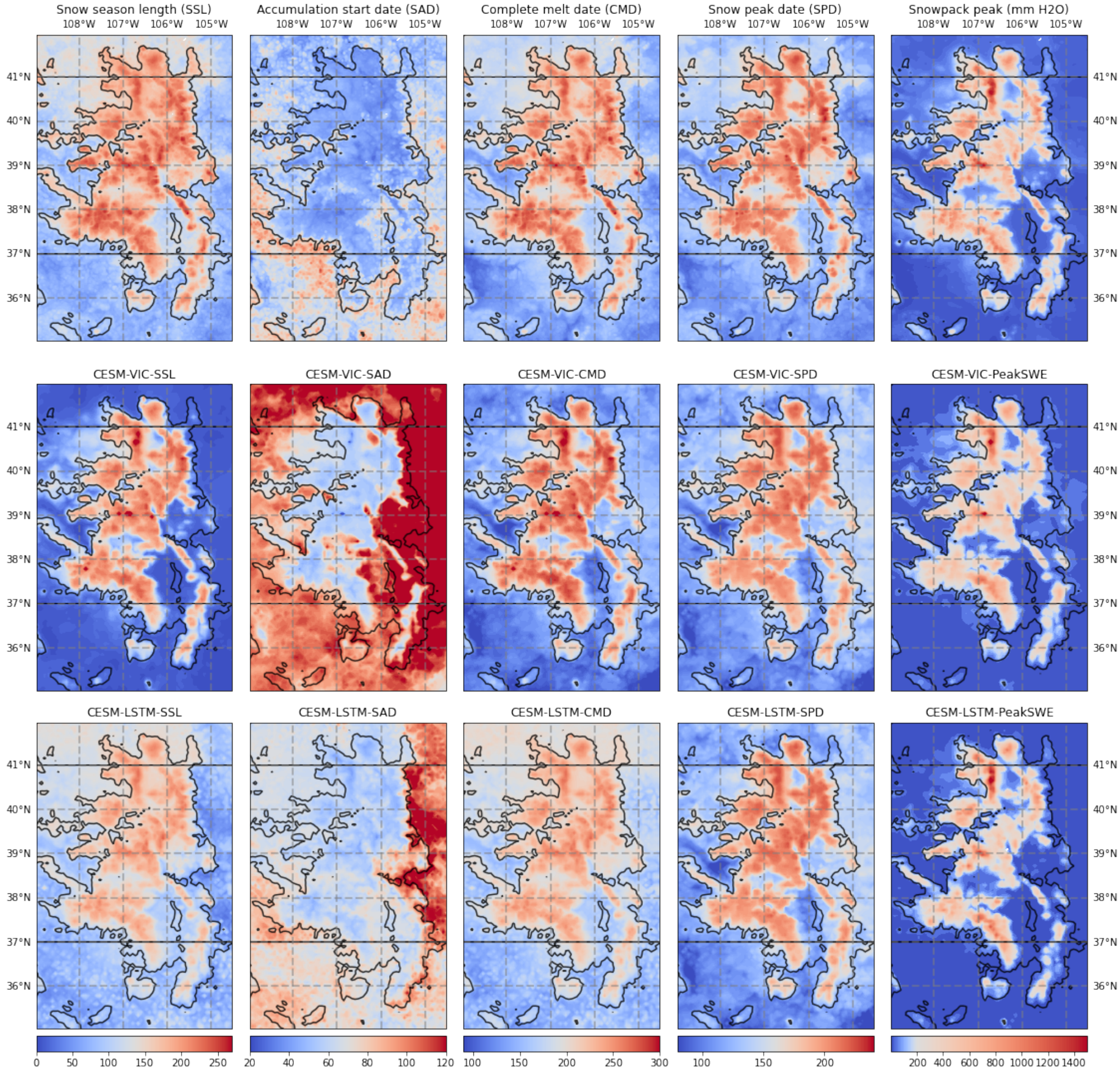

Figure 3.9: Historical snowpack metrics from the NSIDC-UA, CESM-VIC and CESM-LSTM simulations. The black line is the 2300 meter contour.

Table 3.5: Metrics used to assess projected changes in SWE. Days are given as integers, referenced to the beginning of the hydrological year on Oct 1st.

Metric	Description
Snowpack accumulation start date (SAD)	Day when SWE >10% of maximum SWE
Snowpack peak accumulation date (SPD)	Day of maximum SWE
Complete melt date (CMD)	Day when SWE<10% of maximum SWE
Snowpack peak	Maximum SWE
Length of the accumulation season (ASL)	Days from SAD to SPD
Length of the melt season (MSL)	Days from SPD to CMD
Length of the snow season (SSL)	Days from SAD to CMD

Table 3.6: Spatial correlations for historical simulations from LOCA-VIC and the LOCA-LSTM against the NSIDC-UA dataset.

	HadGEM	MIROC	CESM	EC	CNRM	GFDL
SSL						
LOCA-VIC	0.845	0.838	0.846	0.831	0.837	0.829
LOCA-LSTM	0.880	0.885	0.879	0.890	0.881	0.874
SAD						
LOCA-VIC	0.402	0.393	0.377	0.251	0.362	0.378
LOCA-LSTM	0.454	0.487	0.393	0.553	0.502	0.469
CMD						
LOCA-VIC	0.874	0.882	0.891	0.900	0.873	0.884
LOCA-LSTM	0.899	0.903	0.902	0.886	0.898	0.900
SPD						
LOCA-VIC	0.834	0.856	0.855	0.864	0.856	0.878
LOCA-LSTM	0.901	0.904	0.893	0.887	0.916	0.901
Snowpack Peak						
LOCA-VIC	0.956	0.957	0.959	0.954	0.959	0.958
LOCA-LSTM	0.972	0.972	0.977	0.966	0.972	0.971

are presented in the second and third rows. Plots for VIC and LSTM forced by other climate models are available in the supplements (Figure S4-S13). Table 3.6 lists the spatial correlation coefficients between the NSIDC-UA dataset, upsampled to the LOCA grid, and model simulations. Overall, most metrics from both the LOCA-VIC and LSTM-based simulation agree with the NSIDC-UA dataset: the high-elevation region has a longer SSL, earlier SAD, and later CMD and SPD than the plains. SSL, CMD and SPD have higher correlation coefficients, while SAD exhibits a larger bias overall difference, predominantly due to a delayed snow season in the eastern plain area. Moreover, the LSTM model has higher correlations with the NSIDC-UA dataset than the LOCA-VIC model.

Table 3.7: Spatial correlation between LOCA-VIC and LOCA-LSTM projections

	HadGEM	MIROC	CESM	EC	CNRM	GFDL
Historical SSL	0.836	0.838	0.830	0.843	0.843	0.821
Future SSL	0.896	0.888	0.906	0.913	0.914	0.841
Historical SAD	0.697	0.668	0.703	0.576	0.748	0.680
Future SAD	0.464	0.365	0.560	0.548	0.605	0.624
Historical CMD	0.827	0.852	0.854	0.888	0.830	0.824
Future CMD	0.893	0.866	0.933	0.913	0.921	0.869
Historical SPD	0.841	0.857	0.864	0.907	0.875	0.840
Future SPD	0.857	0.819	0.916	0.874	0.895	0.884

Notably, the LOCA-VIC model tends to underestimate snowpack in low-elevation areas, especially the Northern and Southwestern portions of the Rocky Mountains; in both of these regions, SSL and CMD are underestimated. Further, the spatial structure of SSL and CMD is significantly different in the LOCA-VIC simulations. Besides the comparison with the NSIDC-UA dataset, both the historical and future projections from the LSTM model and LOCA-VIC are compared in pairs, and their spatial correlations are shown in Table 3.7. For most snowpack metrics, high correlations are apparent between both models, except for the SAD. Given the relatively low correlation of SAD with the historical UA dataset in both the DL and process-based models, it is likely that SAD is overestimated over the eastern plains in the LSTM, although the LSTM does improve on the CESM-VIC results. Curiously, both process-based and ML-based streamflow models have similarly struggled with streamflow prediction in this region (e.g., Konapala et al., 2020), suggesting further study of the source of these biases is needed here.

LSTM projections of changes to SSL, SAD, CMD and snowpack peak (or maximum SWE), as a function of elevation, are depicted in Figure 3.10. Other metrics are plotted in the same manner and shown in Figure S14-S20. The solid lines in red (blue) indicate the median values for each elevation bin for the future (historical) periods. Shading around the solid lines indicates the range of values for that elevation bin. Spatial changes from GFDL, CESM and MIROC are depicted in Figure 3.11. Figures from other climate models are shown in Figure S21-S23. Regardless of climate model and in both historical and future periods, as elevation increases SSL increases, SAD occurs earlier in the year and CMD occurs later. When comparing changes in SAD (CMD), the

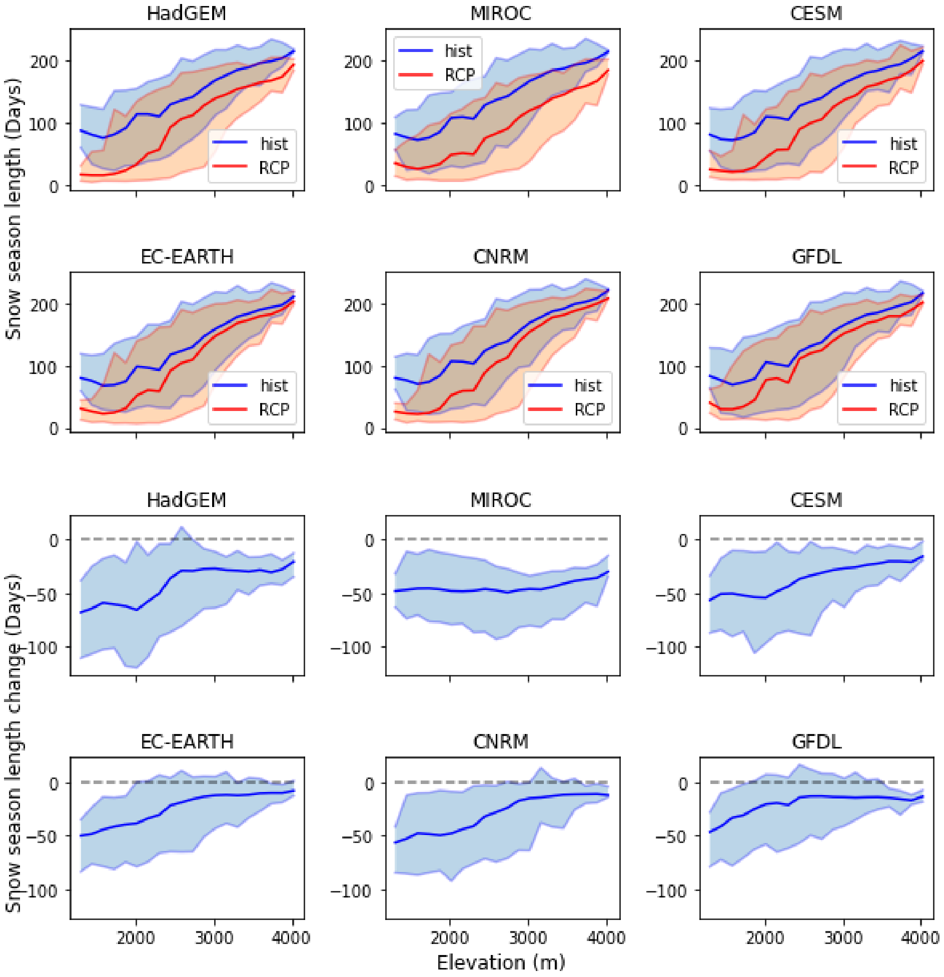

Figure 3.10: Elevation dependency of snow season length (SSL) in the Rocky Mountains. SSL change is calculated as the future (2071-2091) SSL minus the historical (1981-2001) SSL.

postponement (advancement) of the snow season dates at lower elevations is larger than at higher elevations. The symmetric responses in SAD and CMD truncates SSL, particularly over lower elevation areas. The height dependencies of SSL and its changes are consistent across different climate models, but the change magnitude varies. As shown in Figure 3.11, with relatively less decreases in SSL, GFDL produces a future projection of less snow loss, while MIROC represents a scenario of substantial snow loss with much less projected SSL. In addition to SSL, lower elevations also show a more pronounced response to climate change in terms of snowpack peak in Figure S16. The fractional decrease in the snowpack peak is over 50% at the lowest altitudes but tends towards zero at higher altitudes, as shown in Figure S17. This fractional decrease is comparable with what was reported in Rhoades et al. (2018c) and McCrary et al. (2022) for the Rockies. For elevation bands over 3000m, little decrease in the snowpack peak is found, and, in some climate models, there are even increases (i.e., projected peak SWE changes from GFDL and CESM in Figure 3.11). These increases in snowpack at high elevations also emerged in dynamical downscaling models (McCrary and Mearns, 2019; Wrzesien and Pavelsky, 2020) and pseudo-global-warming simulations (Rasmussen et al., 2014). Although snowpack peak provides information about the maximum SWE in each year, April 1st SWE is a traditionally employed metric of snowpack utilized by water managers to estimate summer water availability. With our LSTM-based projections, we further compared the change ratio of April 1st SWE over the historical and RCP8.5 periods. Table 3.8 lists the fractional decrease of April 1st SWE along with the decreases in SSL. Spatial mean SWE over the Rocky Mountains is assessed for the April 1st SWE, since the mean SWE represents the average available water stored in snowpack, and can be transformed to total water resources by multiplying the regional area. Across all selected climate models, the decreases in snowpack are consistent with lower projected April 1st SWE. This result suggests a drier summer in the future with less available water from snowmelt. Despite of consistency in the loss of SWE, different climate models exhibit a large variance in the fractional decrease, which varies from 3.9% to 48%, and only the MIROC5 model comes close to the 45% SWE loss reported in Li et al. (2017). This projection spread is accorded with Figure 3.11, which shows that the drop in peak SWE is more pronounced in MIROC5, but far more muted in GFDL-ESM2M. This is likely due to the differences

Table 3.8: Metrics related to historical and projected April 1st mean SWE.

	April 1st SWE Fractional Decrease (%)	Historical April 1st SWE (mm)	Projected April 1st SWE (mm)
HadGEM2-ES	20.79	62.58	49.57
MIROC5	48.14	62.32	32.32
EC-EARTH	7.43	58.54	54.19
CNRM-CM5	5.83	62.97	59.29
CESM-CAM5	21.30	67.99	53.51
GFDL-ESM2M	3.94	63.65	61.14

in precipitation and temperature from the climate models: a quick assessment of the climatology reveals GDFL-ESM2M produces more precipitation and cooler temperatures, which results in less SWE loss. The comparison of mean precipitation and temperature between GDFL-ESM2M and MIROC5 can be found in Figure S24.

3.6 Conclusions

Previous studies have investigated and demonstrated that DL models are useful for Earth system applications. In this work, we investigated three DL models for SWE prediction and projection over the Western US, with a focus on the Rocky Mountain region. The LSTM model is a fast and simple but useful DL model for time-series tasks with its natural ability to handle information from previous time steps. The TCNN, another DL model, mimics the temporal dependency with stacked 1-D CNN layers. Without inherent states like LSTM model, its performance is somewhat worse. Attention models are also promising DL methods and have become widespread in their use for time-series tasks, especially for natural language processing (NLP). In recent years, significant effort has been made to optimize the original Transformer architecture and make it more computationally and memory efficient (Lin et al., 2021; Tay et al., 2020). These variants of Transformer models could provide more choices for Earth system applications. Besides these typical DL models, there has been effort to combine different types of sequential layers or blocks in a hybrid model, as shown in Xu et al. (2020) and Chen et al. (2020). In this study, we were limited by our computational resources and so didn't test these hybrid forms of DL models, but plan to investigate them in future work.

Compared with more traditional process-based methods to estimate SWE (e.g., the NSIDC-UA SWE dataset and the hydrologic model VIC), we show that DL models can achieve higher

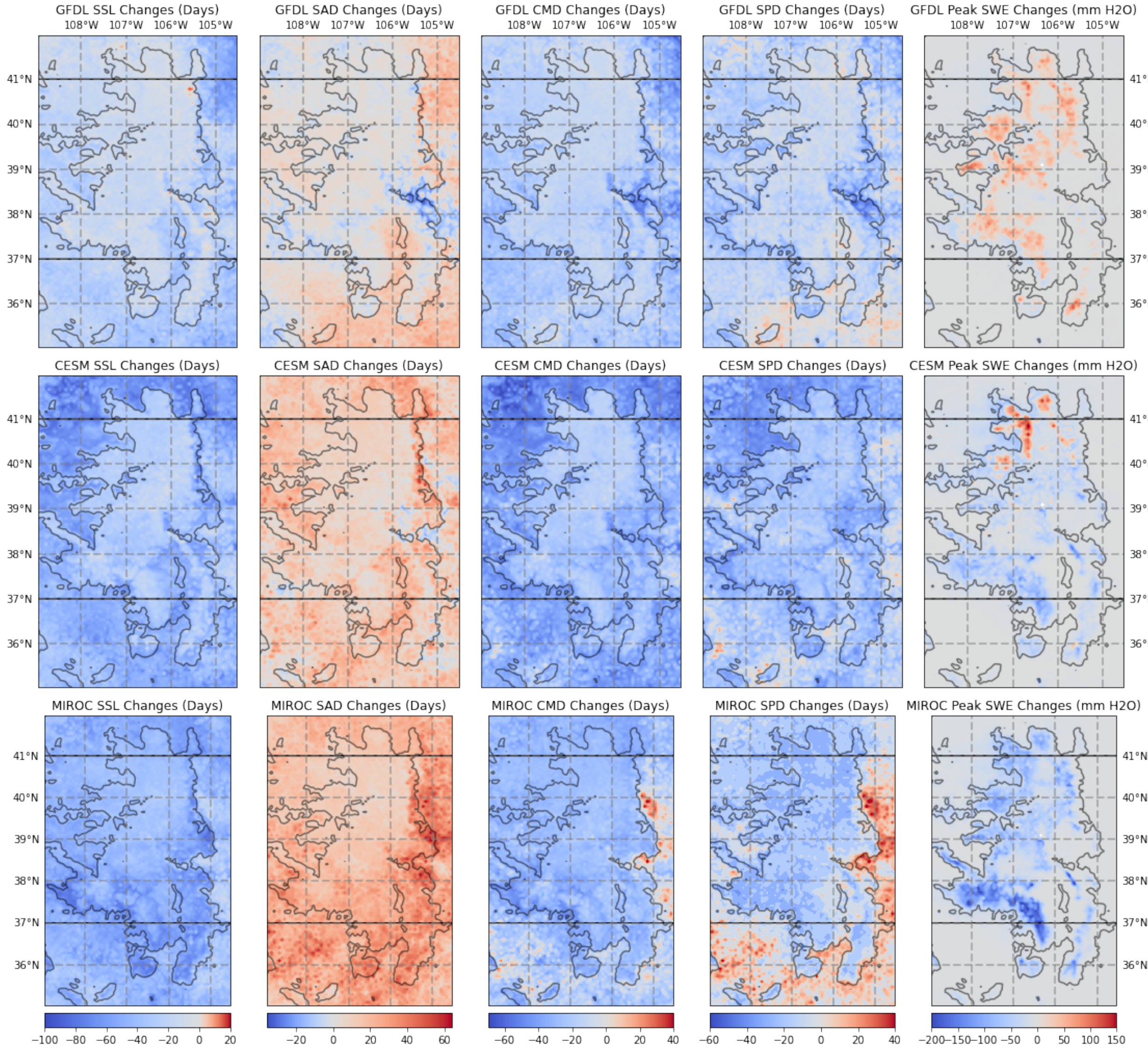

Figure 3.11: Future (2071-2091) minus historical (1981-2001) changes in snowpack metrics from GFDL (top), CESM (middle) and MIROC (bottom). Changes are calculated as projected metrics minus historical values. The black line is the 2300 meter contour.

accuracy (in terms of NSE), when estimating in-situ SNOTEL observations. With the acceleration of GPUs, the training time is acceptable and the inference is fast: within 35 minutes for a 10-year prediction on the 168 × 108 grid with parallel execution. Given the important role that SWE has on the mountainous hydrological cycle, DL models show promise for use in operational forecasting. The computational speed of DL models also enable the ability to provide an ensemble of SWE predictions through perturbations of the initial weights of the model, enabling probabilistic SWE prediction and projection.

We also show a permutation-based method that enhances DL model interpretation. Precipitation and elevation are the two dominant variables for SWE prediction, and this is consistent with our physical understanding of the snowpack dynamics. We caution that any conclusions drawn from this interpretation could be sensitive to strong correlations among input variables. In future work, we would like to examine methods that could eliminate these input correlations. For example, one could reconstruct a set of orthogonal input variables from the original inputs using principal component analysis. These orthogonal variables would contain the same information as the original inputs, which preserves the accuracy of the DL models, and the orthogonality would simplify the interpretation process. Nonetheless, the interpretation will be drawn from the reconstructed variables, which are linear combinations of original inputs and may not represent any real physical features.

Although in-situ estimates of SWE are useful for particular applications, spatiotemporally continuous SWE predictions and projections are needed for a wider range of applications. To this end, we apply the trained DL model to generate historical gridded SWE across the Rocky Mountains. A major constraint for our DL model is that most in-situ estimates of SWE are provided at mid-to-high elevations at discrete points throughout the Rocky Mountains. Therefore, the extrapolation problem for our DL model is particularly pronounced when we apply our model to a wider spatial area where the statistical properties learned from the in-situ measurements might not hold (e.g., lower elevations). Without additional training data, our extrapolation results prove that we can generalize the DL models by altering the prediction from an absolute SWE depth to its seasonality. With this transformation, the target prediction becomes an elevation-invariant quantity that can be generalized to low-elevation areas, an approach also used for climate model emulation in Beucler et al. (2021b). To overcome the extrapolation problem without any loss of information (or transformation), the DL models would either need more training data in low-elevation areas (e.g., satellite images) or incorporate physical constraints into their architectures (Kashinath et al., 2021).

Given the skill of the LSTM DL model in generating spatiotemporal predictions of historical SWE over the Rocky Mountains, we then leveraged the LSTM to provide end-of-century snowpack projections using the LOCA ensemble of downscaled climate model projections. We found that the LSTM SWE historical predictions and future projections were consistent across different LOCA climate model projections. Consistent with physical intuition, lower-elevations tended to have shorter ASL, MSL and SSL, and lower snowpack peak (or maximum SWE) in both historical and end-of-century scenarios. Comparing the future and historical period, the LSTM model suggests the ASL, MSL, SSL and snowpack peak will decrease over the Rocky Mountains, with lower-elevations more sensitive to climate change, exhibiting larger decreases in SSL and maximum SWE. Interestingly, an increase in the maximum SWE was found at higher-elevations for some climate model projections. The relationships between snowpack and other static features was also investigated. We show a strong correlation between DAH and snowpack, which is consistent with our permutation test that elevation and DAH are the two most important static features in estimating SWE across space.

A limitation of our study is that it mainly focuses on the use of data-driven models and does not incorporate physical constraints. One opportunity for future work would be to add mass balance into the model, as with the model described in Hoedt et al. (2021). These physical constraints could improve the physical interpretability of these models, as well. Additionally, for the extrapolation problem, we only used the NSIDC-UA dataset as a reference since we lack in-situ observations across the entire Rocky Mountains. To better understand sensitivities related to extrapolation, we can apply our model to the areas where we have other source of observations, such as the ASO Airborne Lidar measurements of SWE in California (Painter et al., 2018), and snow course measurements. Finally, it is clear that the mean squared error-based loss function employed in DL model training often underestimates extreme values. With the success of generative adversarial models in Earth system modelling (Manepalli et al., 2019; Pan et al., 2021), there is a possibility we will be able to obtain better estimates of SWE in the extreme using both sequential models and adversarial loss.

3.7 Supplements

Table S1: Hyperparameter search candidates for all DL models.

LSTM					
	Hidden states	64	128	256	512
TCNN					
	Blocks	4	5	6	
	Kernel size	7	9		
	Number of kernels	16	32	64	
Attention					
	Heads	8	16		
	Embedding size ratio	1	2		
	Attention layers	2	3	4	
	Forward dimension	16	32	64	

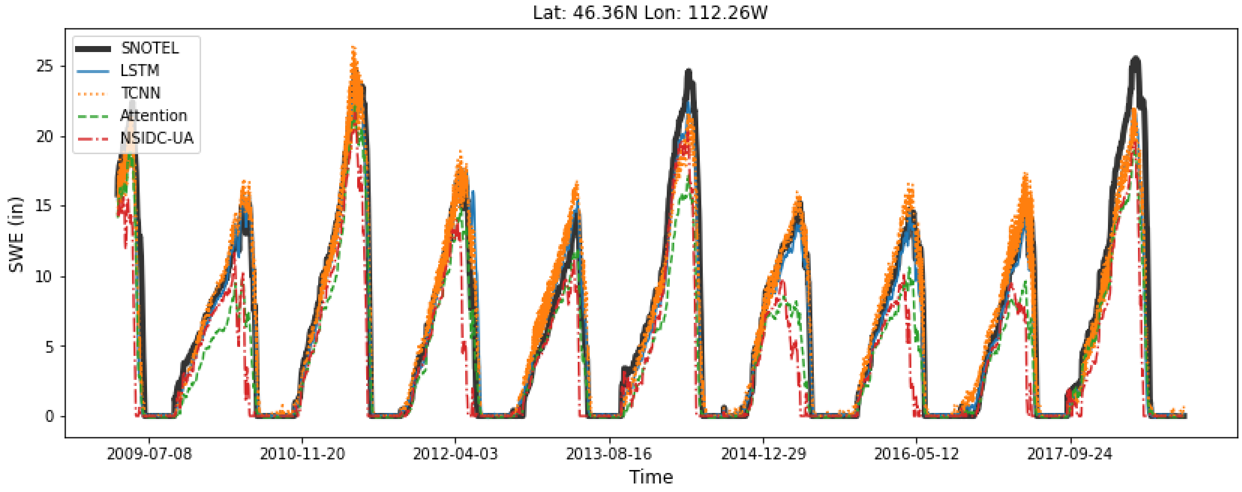

Figure S1: SWE prediction example for a selected SNOTEL station.

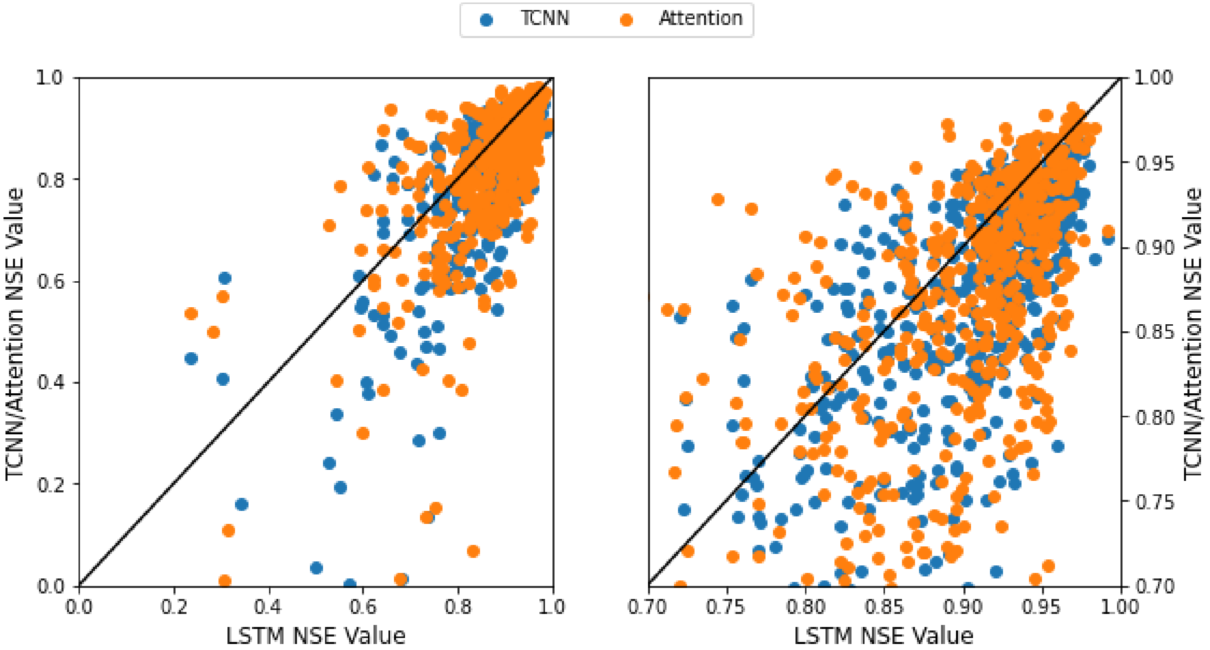

Figure S2: NSE values for the LSTM and TCNN/Attention models. The left figure shows NSE values in [0, 1], while the right figure magnifies the range [0.7, 1].

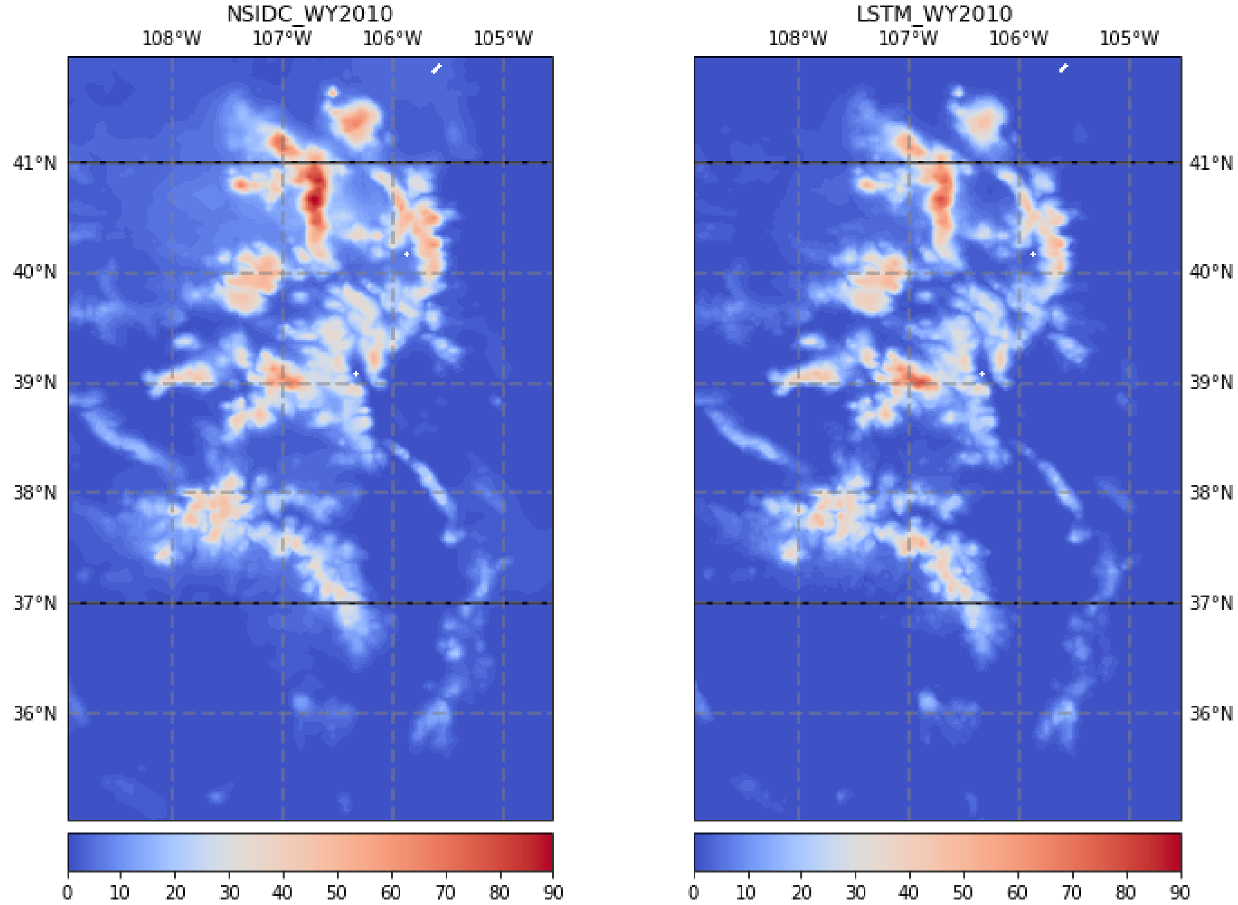

Figure S3: Extrapolation maximum SWE (in) for water year 2010. Left figure is from the NSIDC-UA dataset. Right figure is from the LSTM simulation.

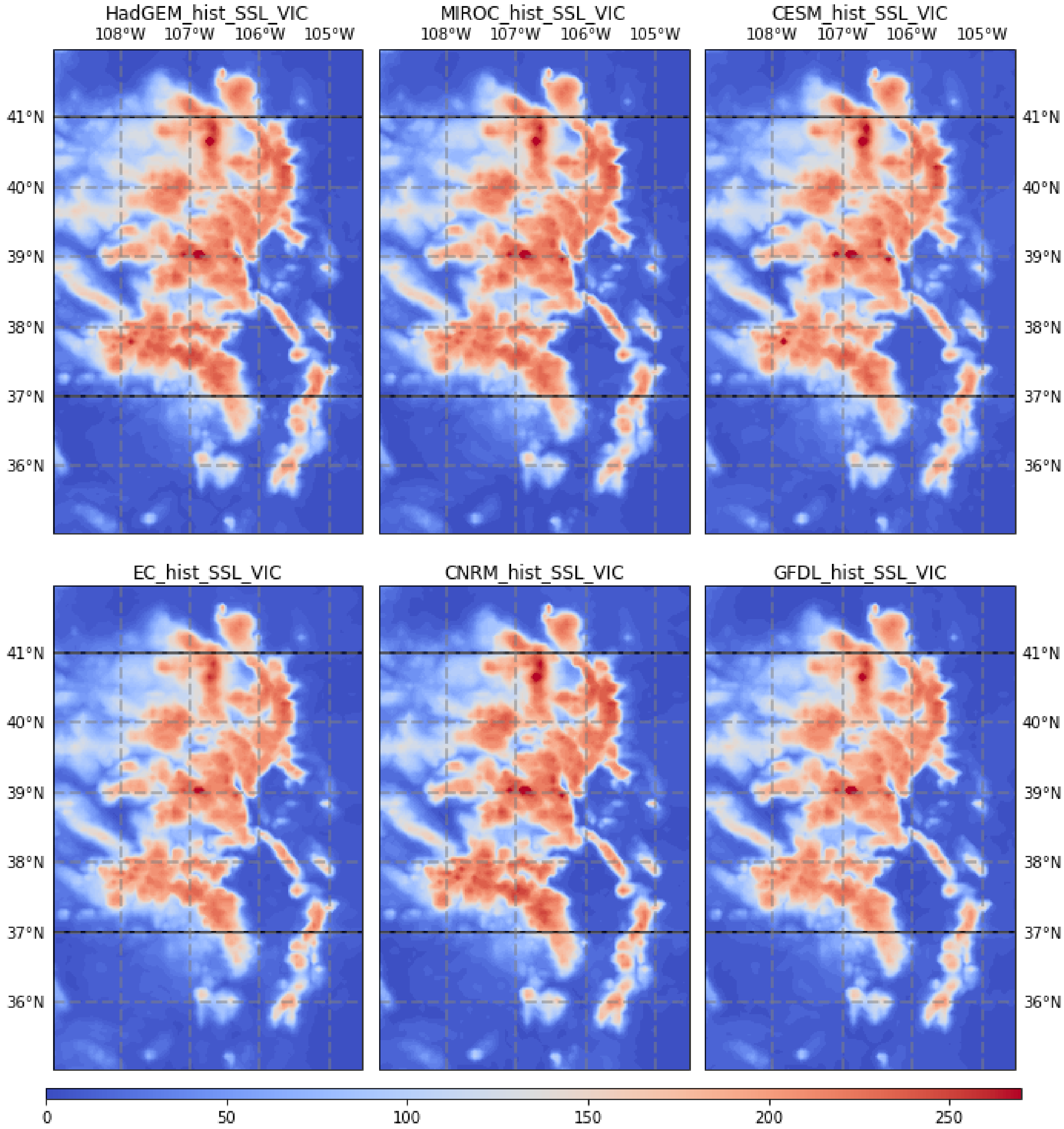

Figure S4: Historical SSL from LOCA-VIC simulations.

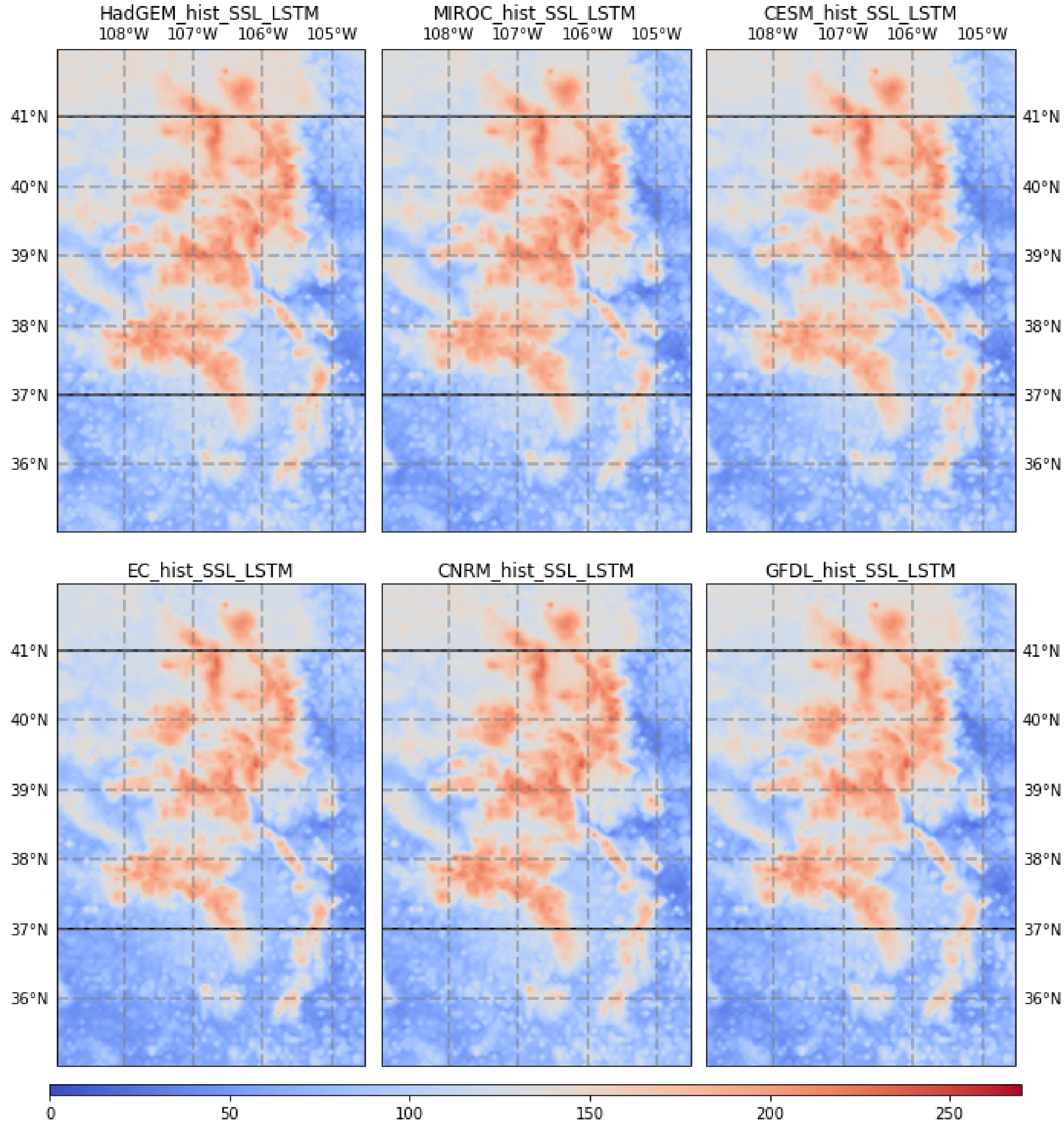

Figure S5: Historical SSL from LSTM simulations.

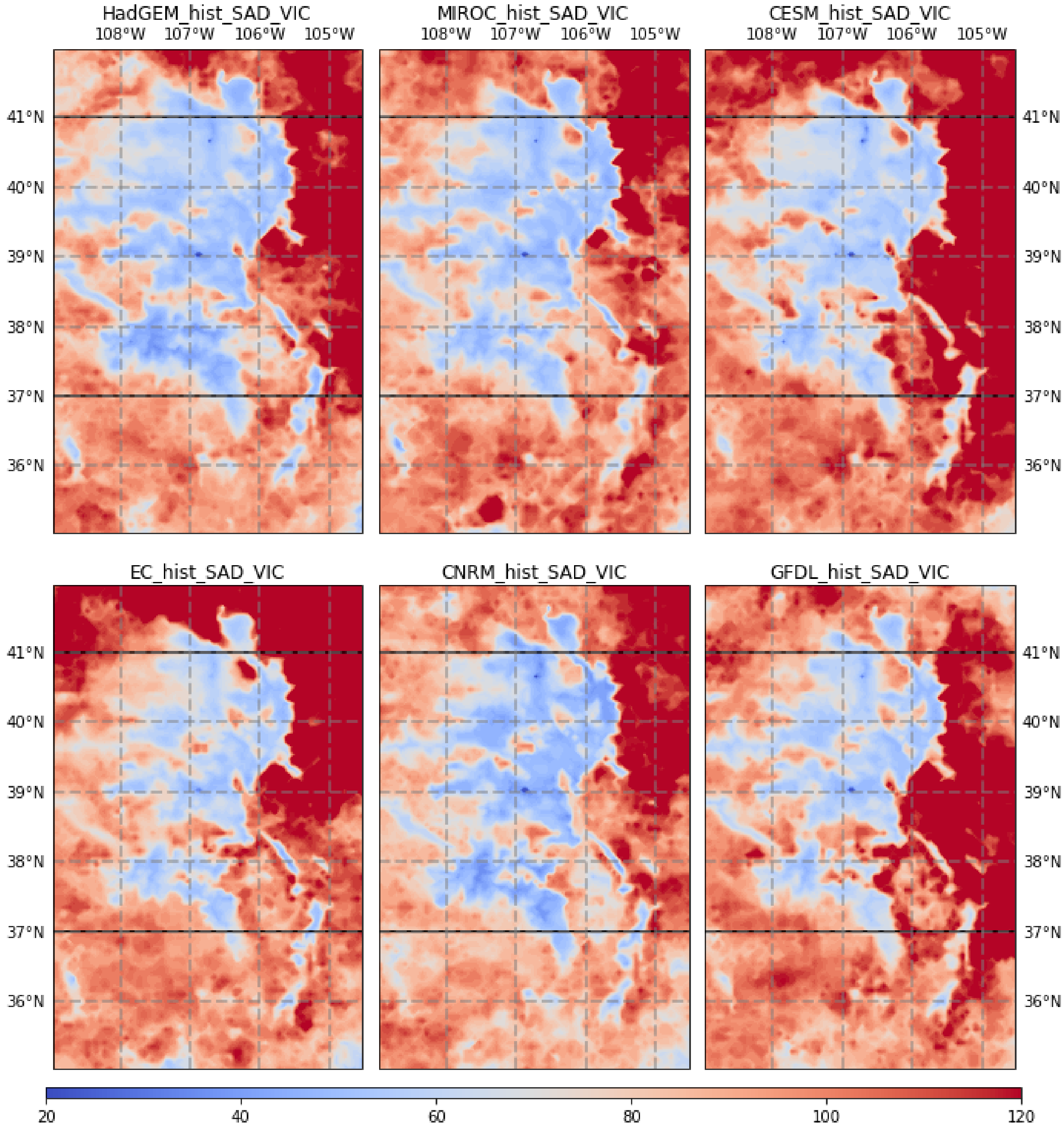

Figure S6: Historical SAD from LOCA-VIC simulations.

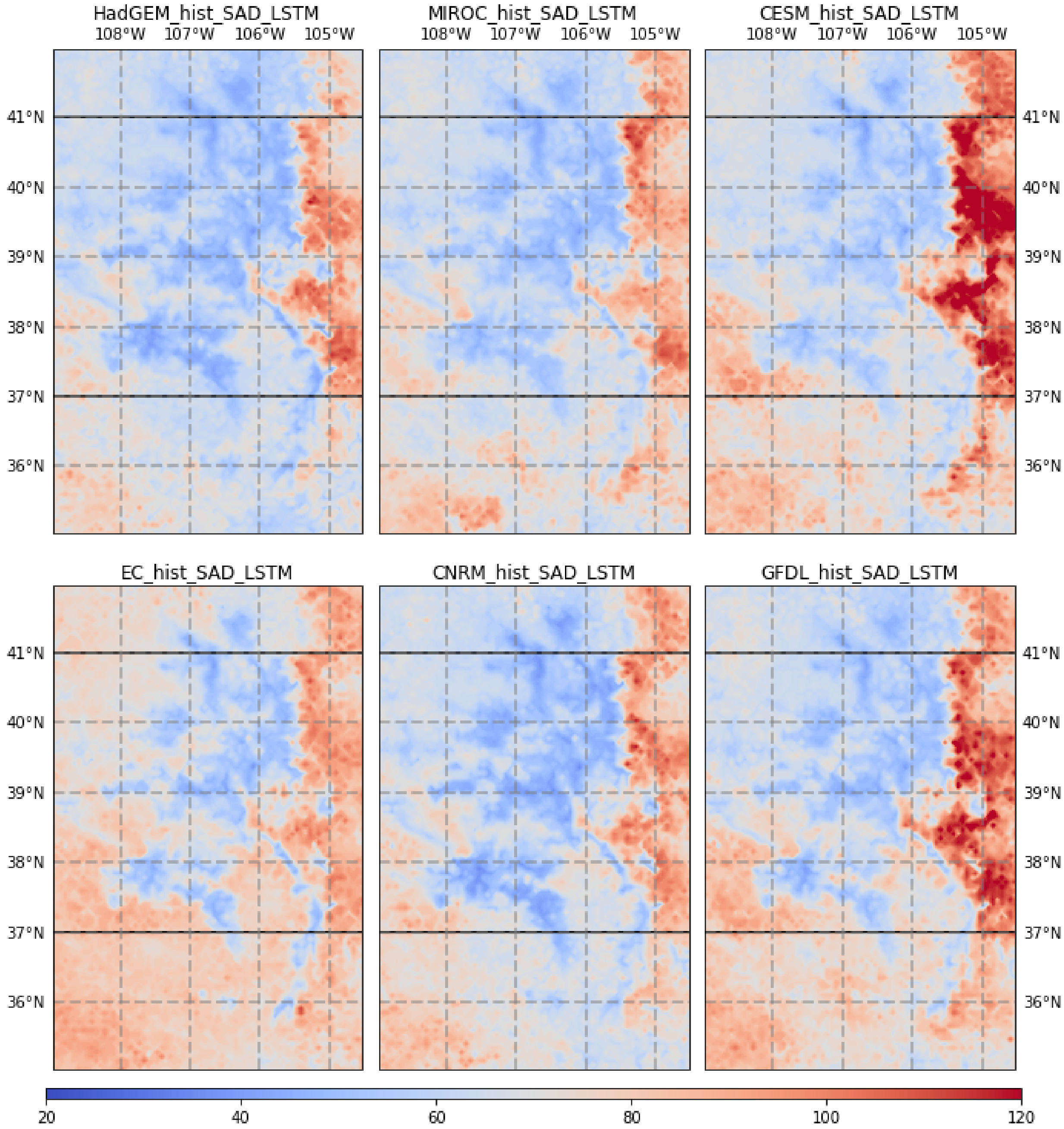

Figure S7: Historical SAD from LSTM simulations.

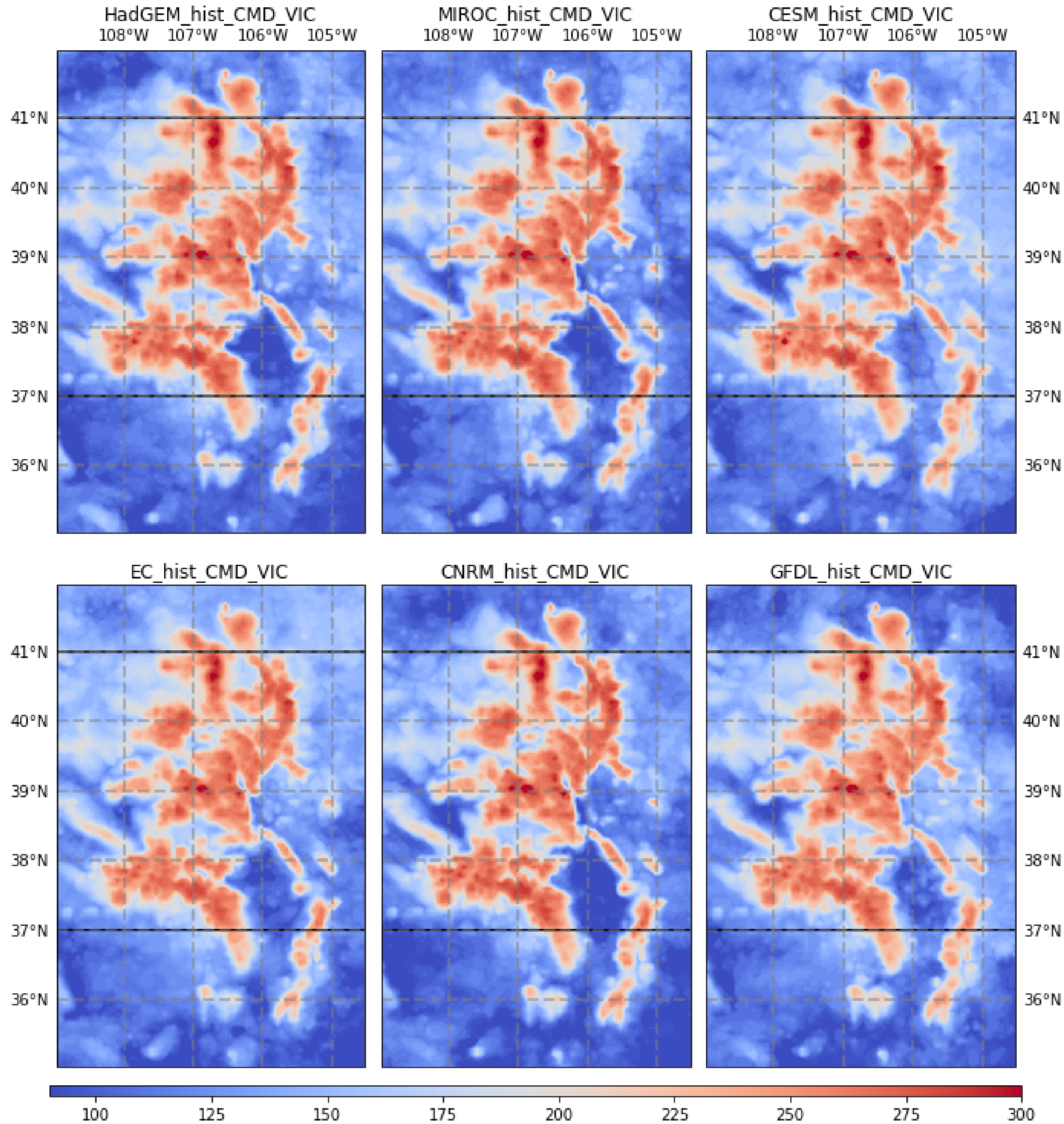

Figure S8: Historical CMD from LOCA-VIC simulations.

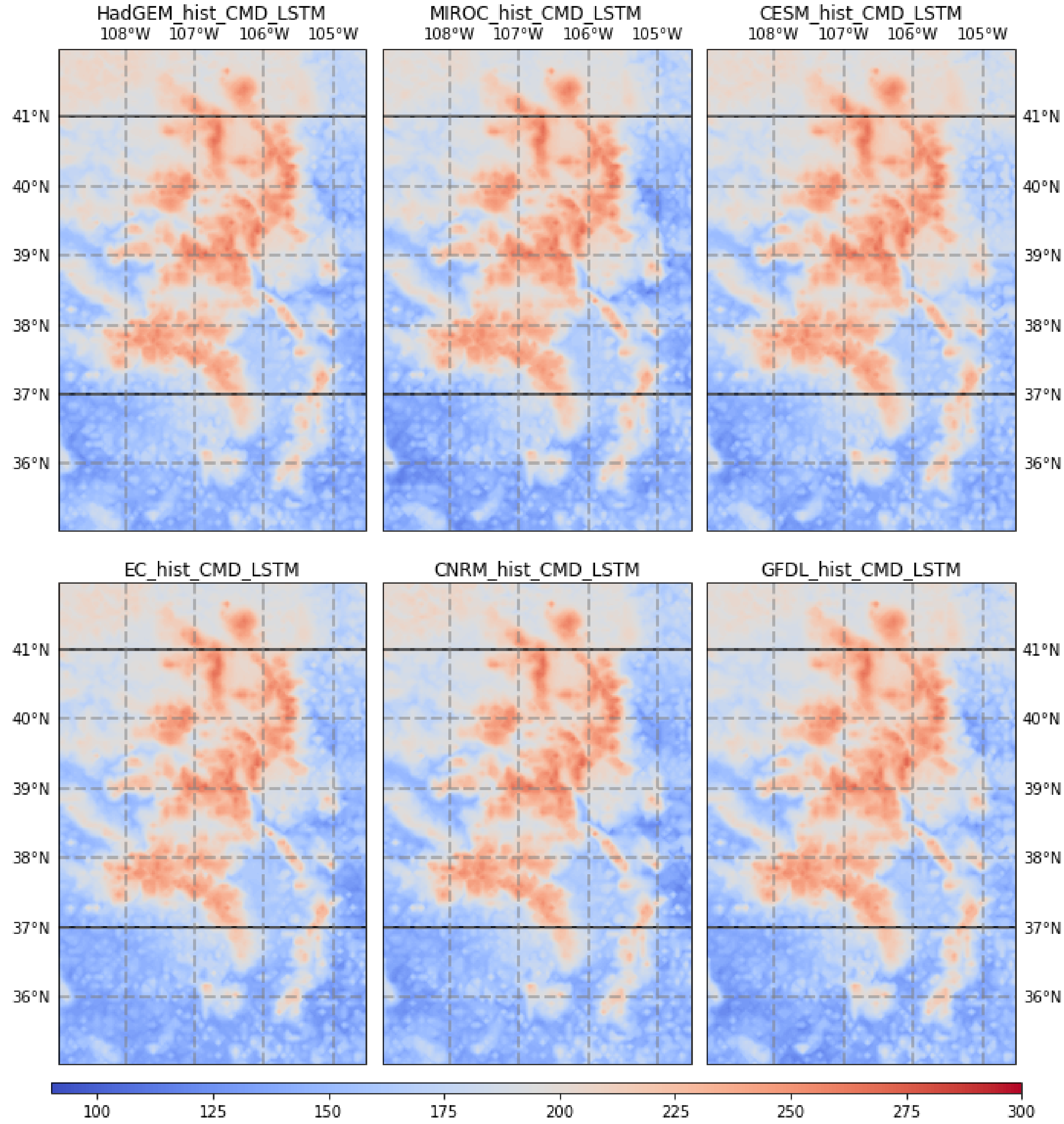

Figure S9: Historical CMD from LSTM simulations.

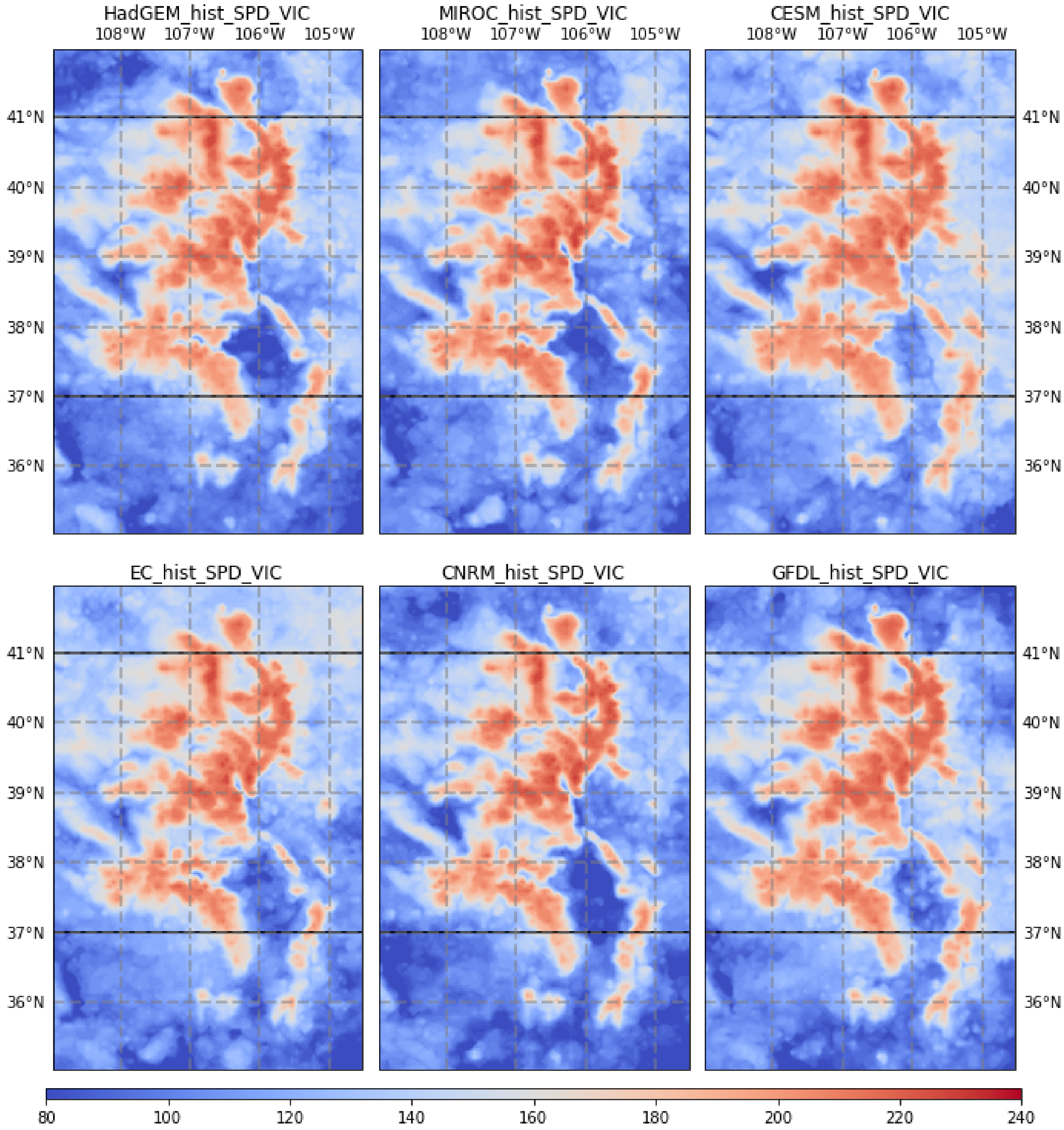

Figure S10: Historical SPD from LOCA-VIC simulations.

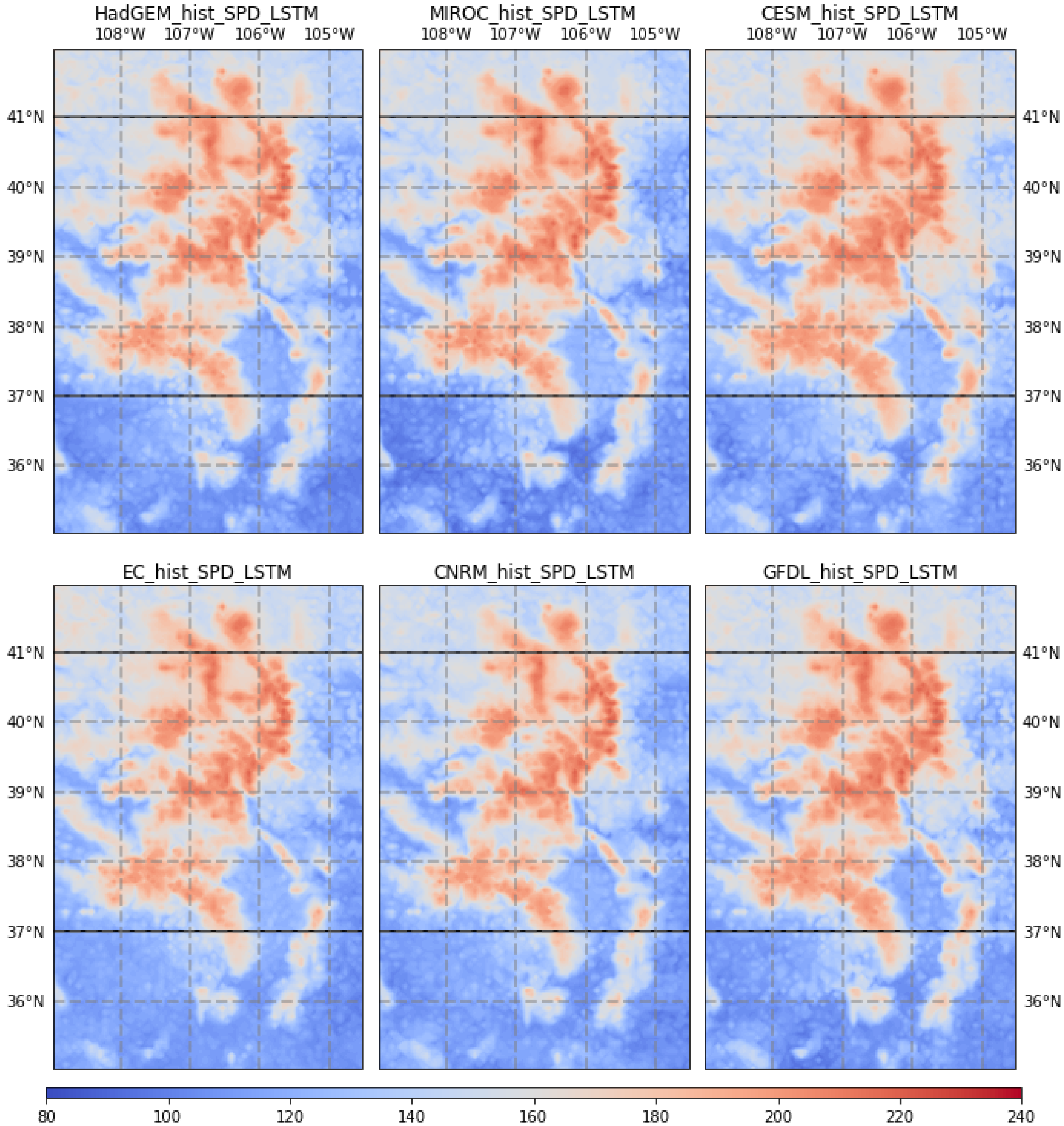

Figure S11: Historical SPD from LSTM simulations.

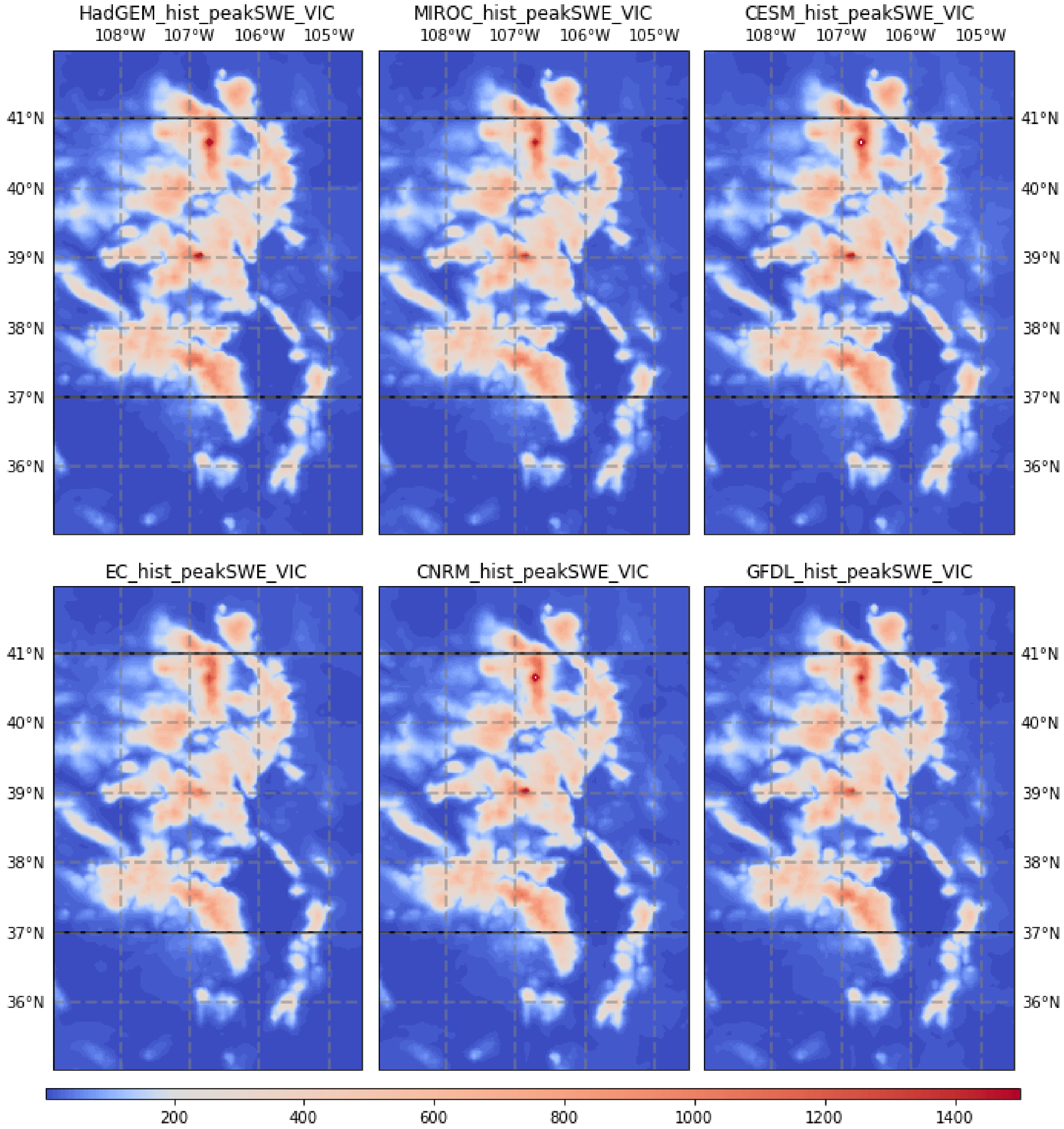

Figure S12: Historical snowpack peak from LOCA-VIC simulations.

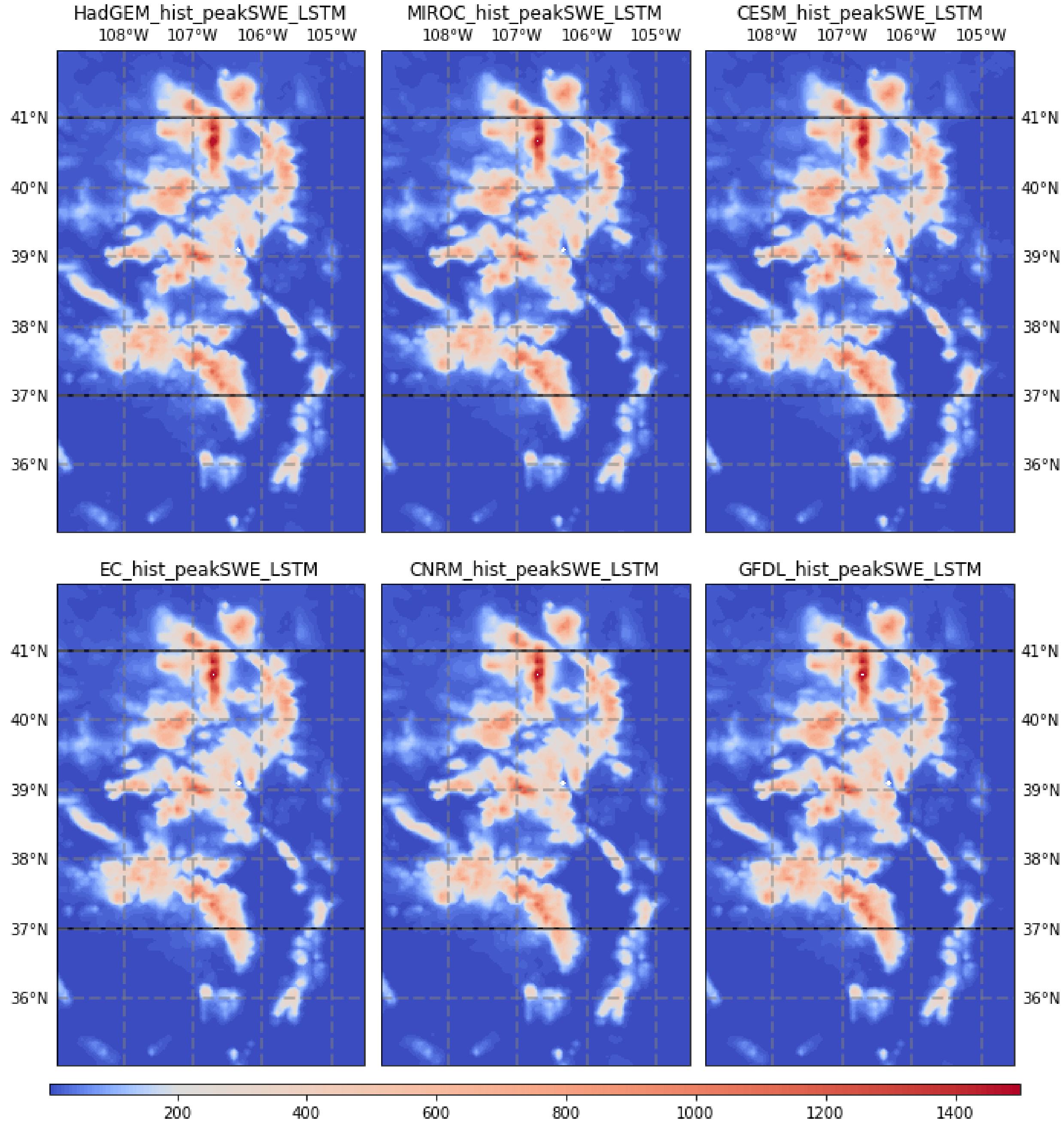

Figure S13: Historical snowpack peak from LSTM simulations.

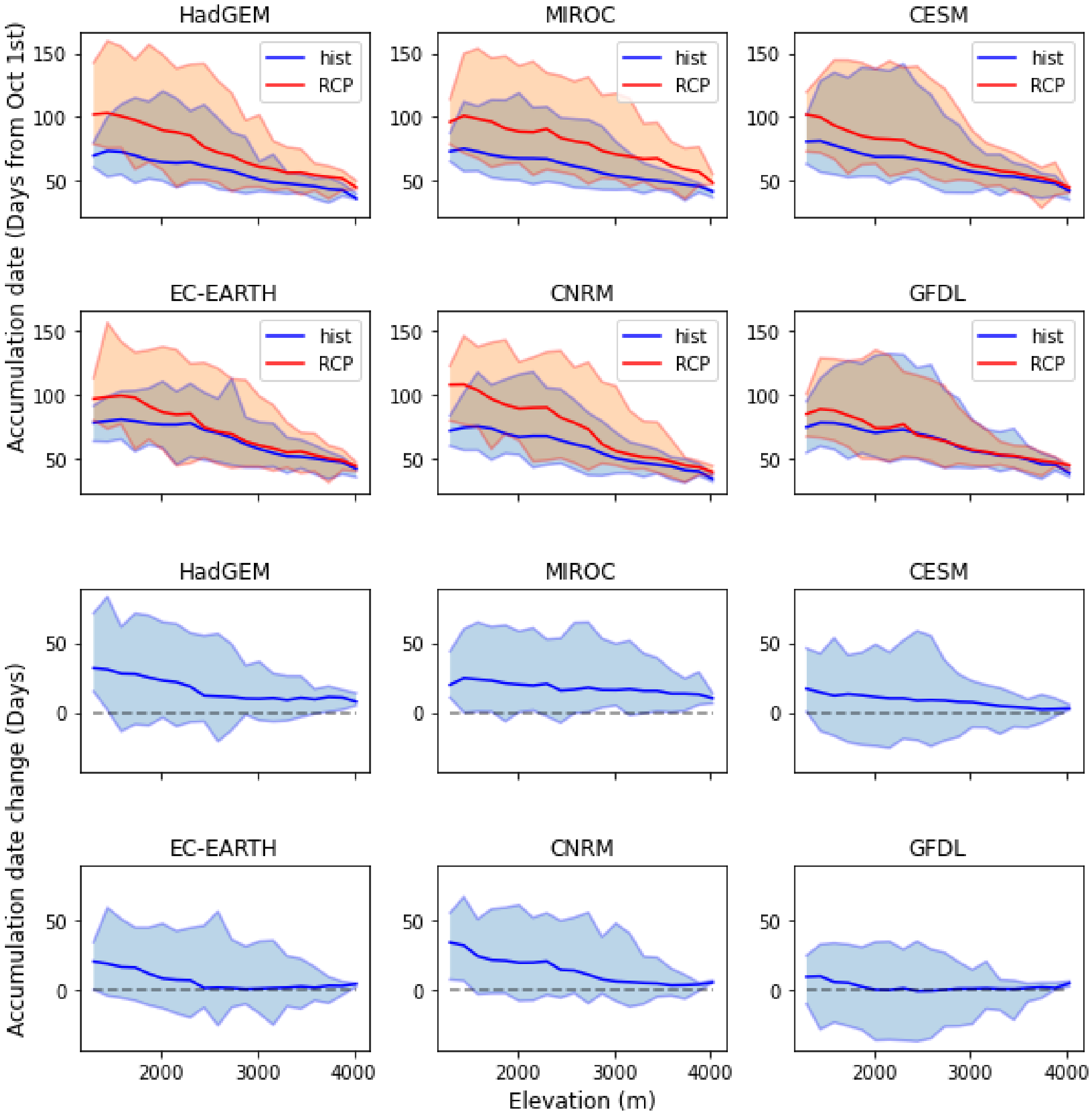

Figure S14: Elevation dependency of SAD. SAD change is calculated as the future SAD minus the historical SAD.

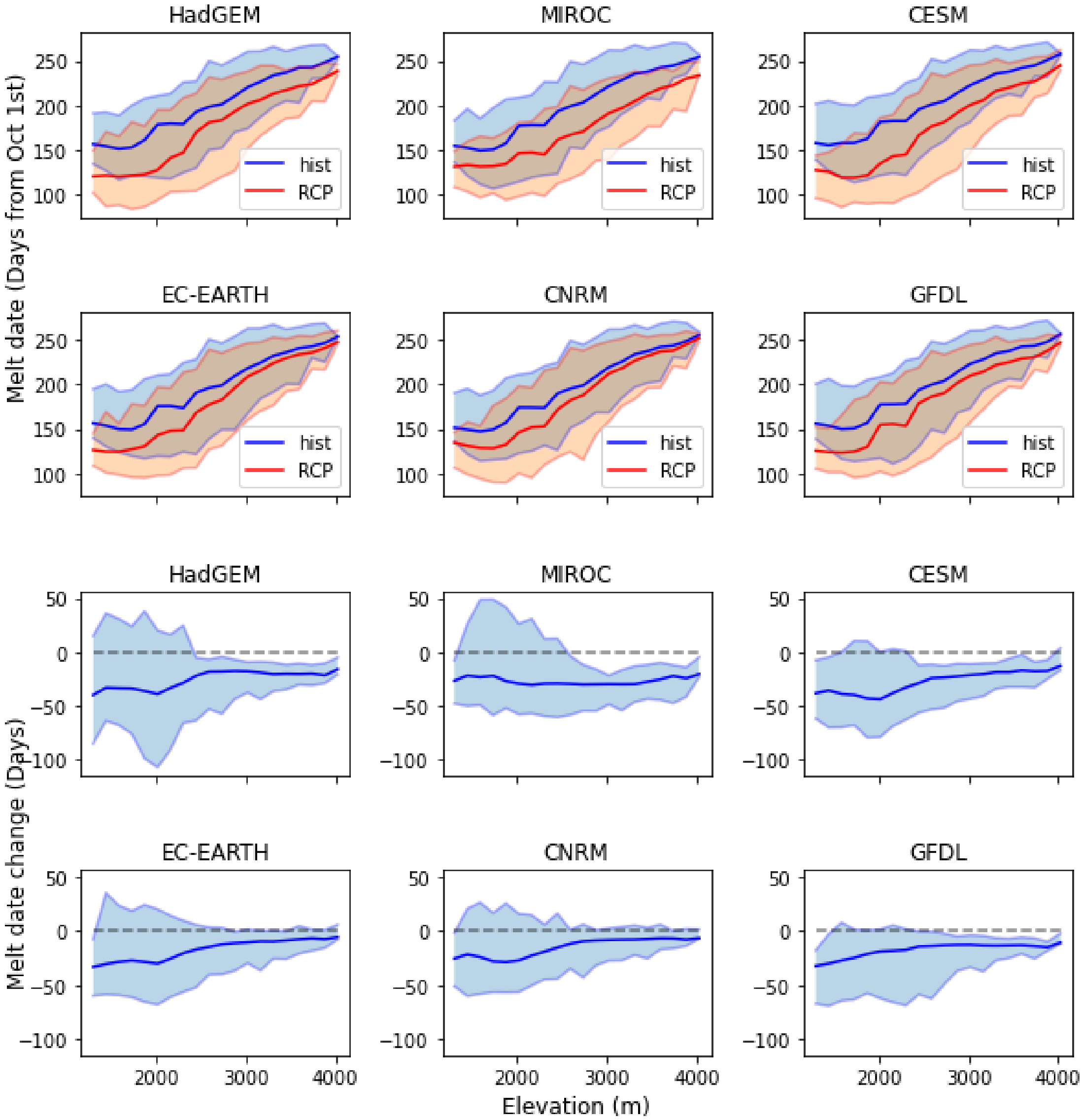

Figure S15: Elevation dependency of CMD. CMD change is calculated as the future CMD minus the historical CMD.

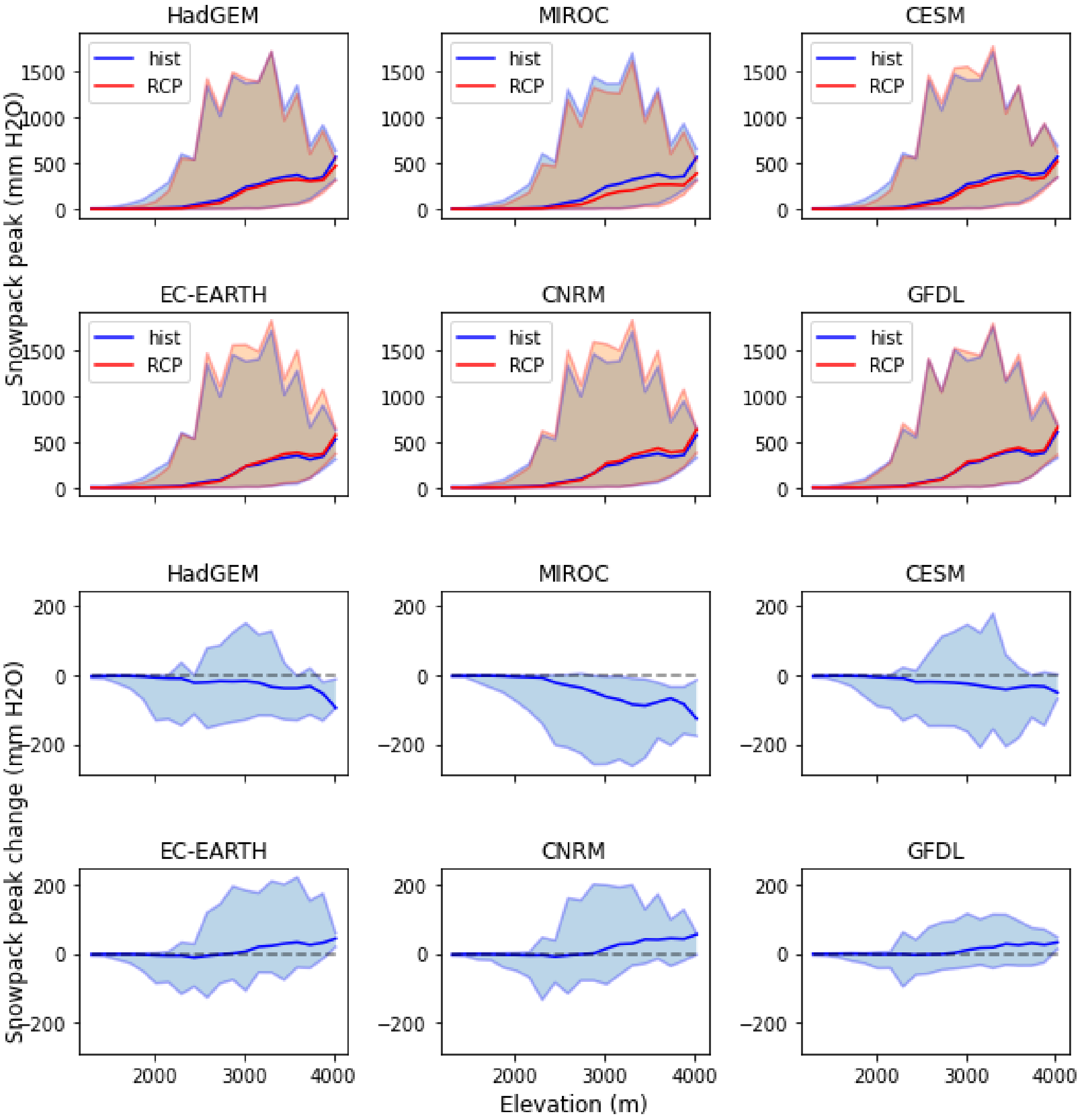

Figure S16: Elevation dependency of snowpack peak.

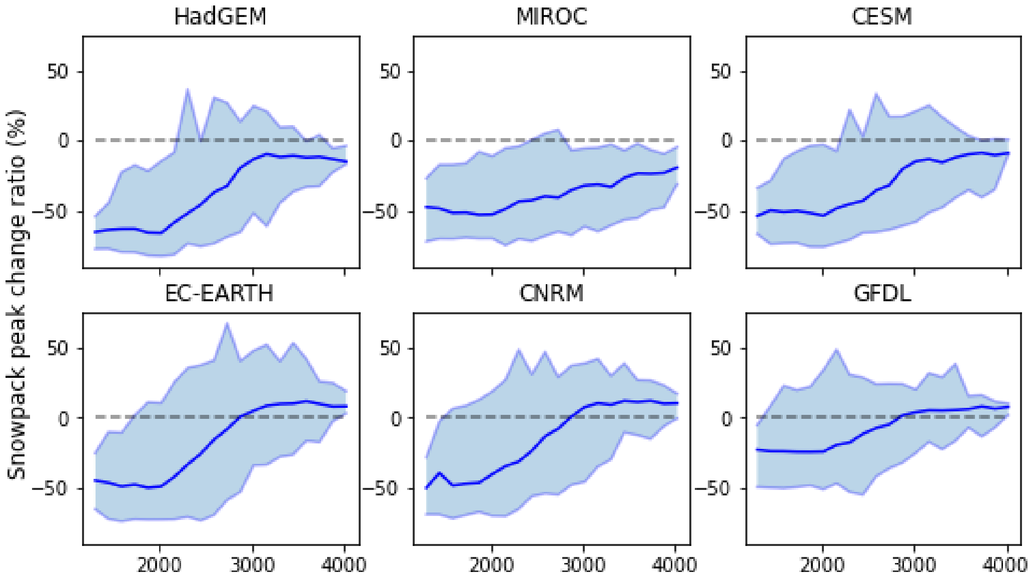

Figure S17: Elevation dependency of changes in snowpack peak ratio. Change ratio is calcualted as the difference between future SWE fraction and historic SWE fraction divided by historic SWE fraction.

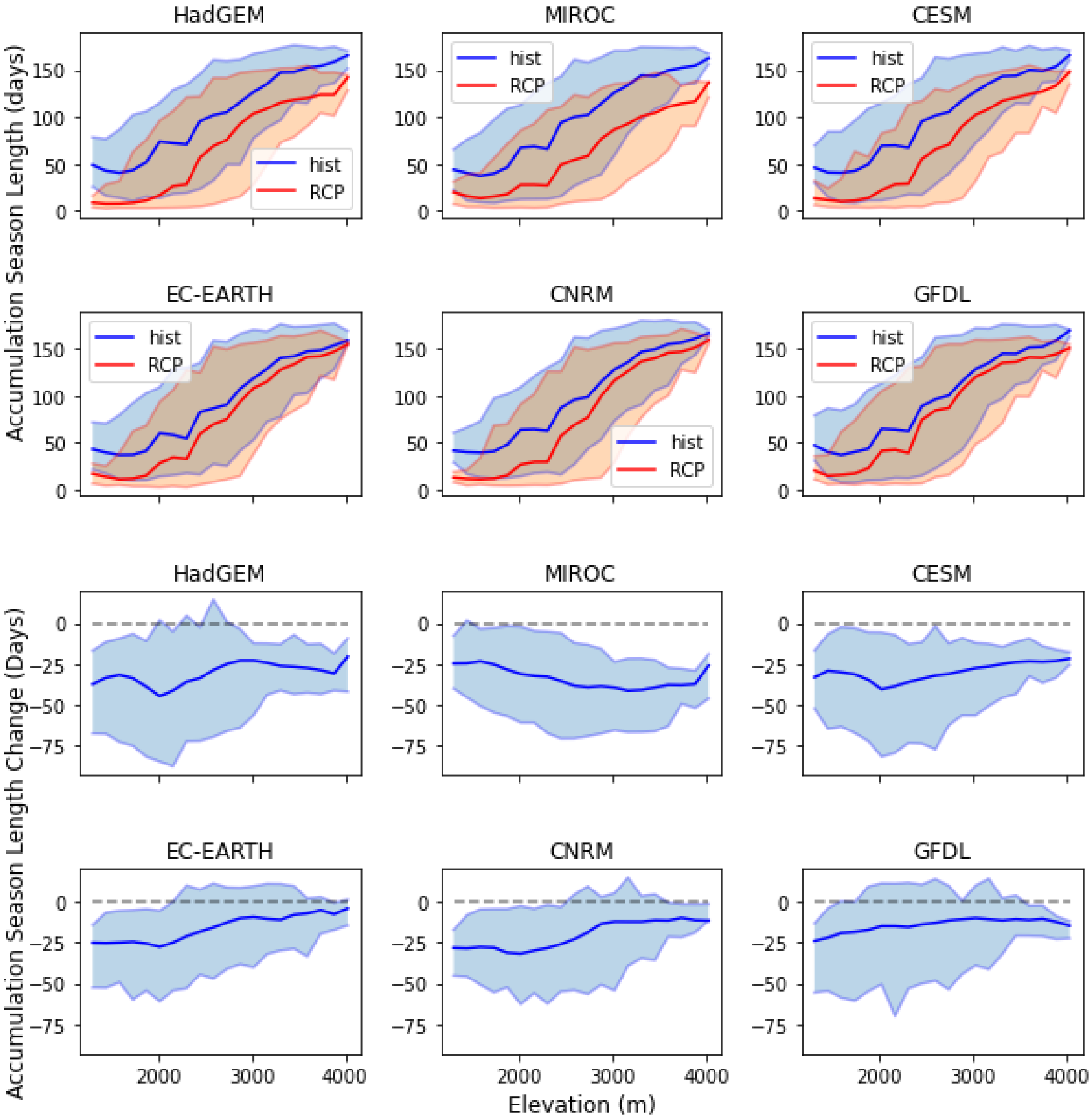

Figure S18: Elevation dependency of ASL. ASL change is calculated as the future ASL minus the historical ASL.

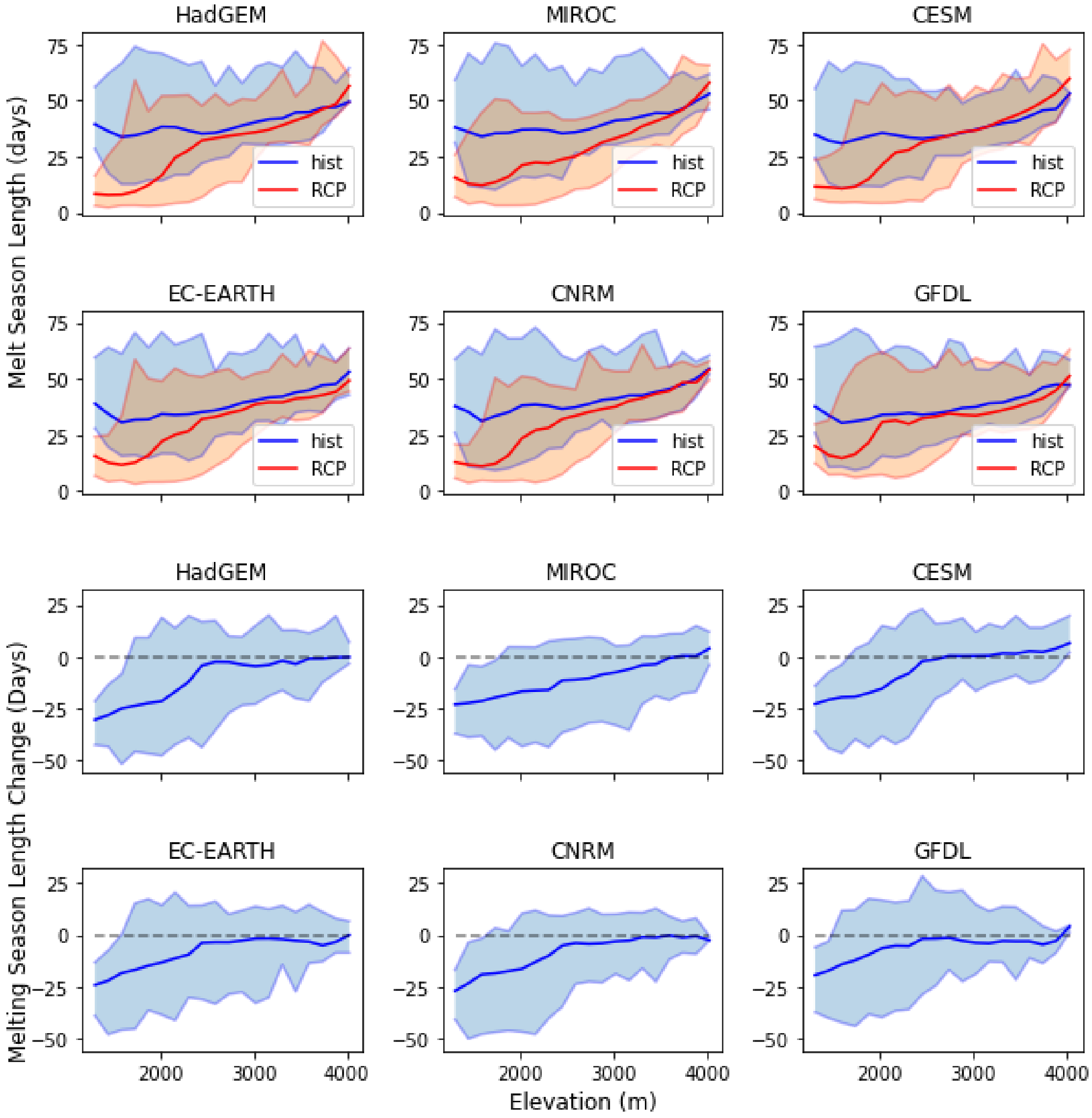

Figure S19: Elevation dependency of MSL. MSL change is calculated as the future MSL minus the historical MSL.

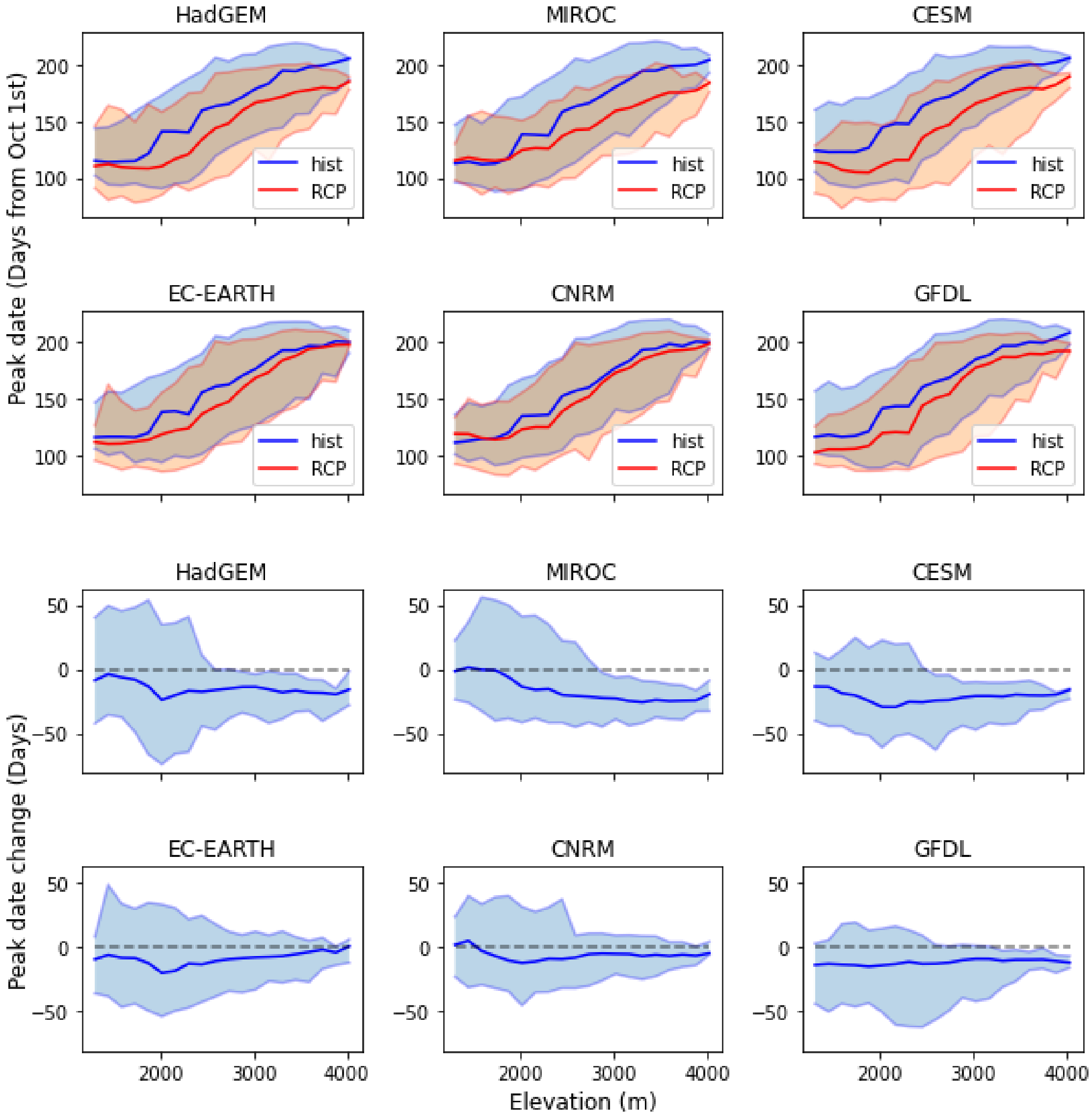

Figure S20: Elevation dependency of SPD. SPD change is calculated as the future SPD minus the historical SPD.

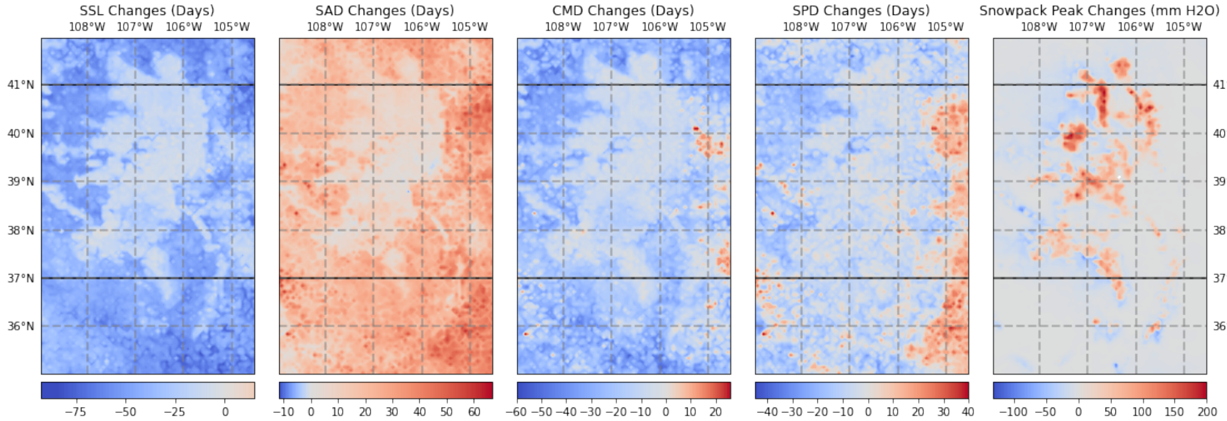

Figure S21: Projected changes of snowpack metrics from CNRM-LSTM.

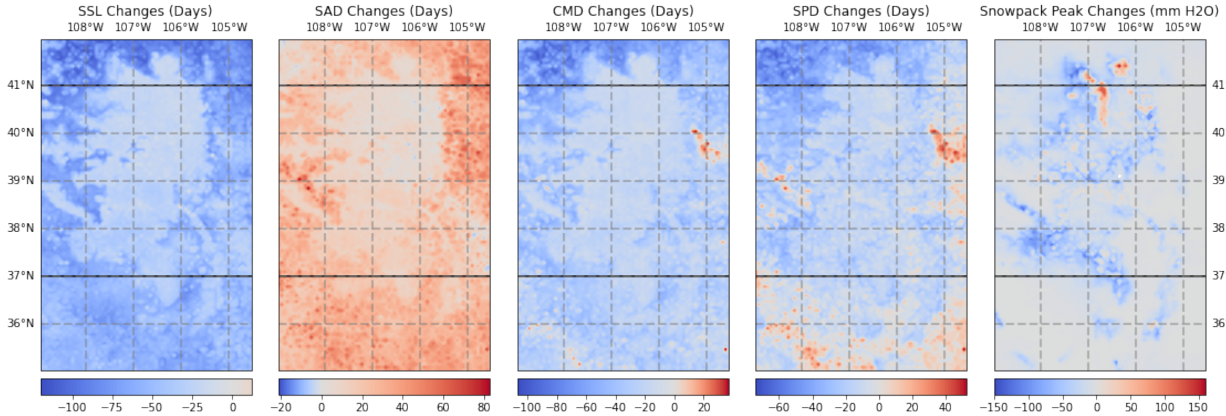

Figure S22: Projected changes of snowpack metrics from HadGEM-LSTM.

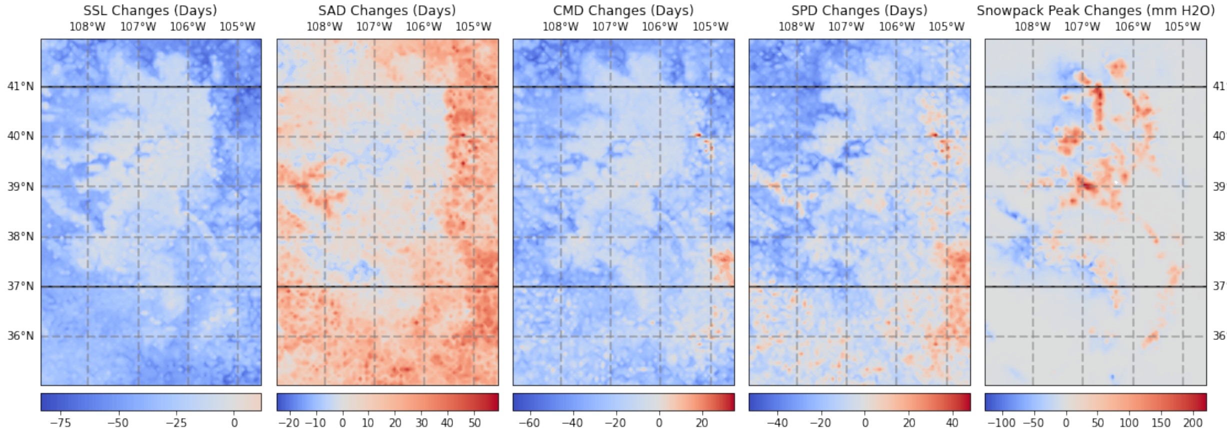

Figure S23: Projected changes of snowpack metrics from EC-LSTM.

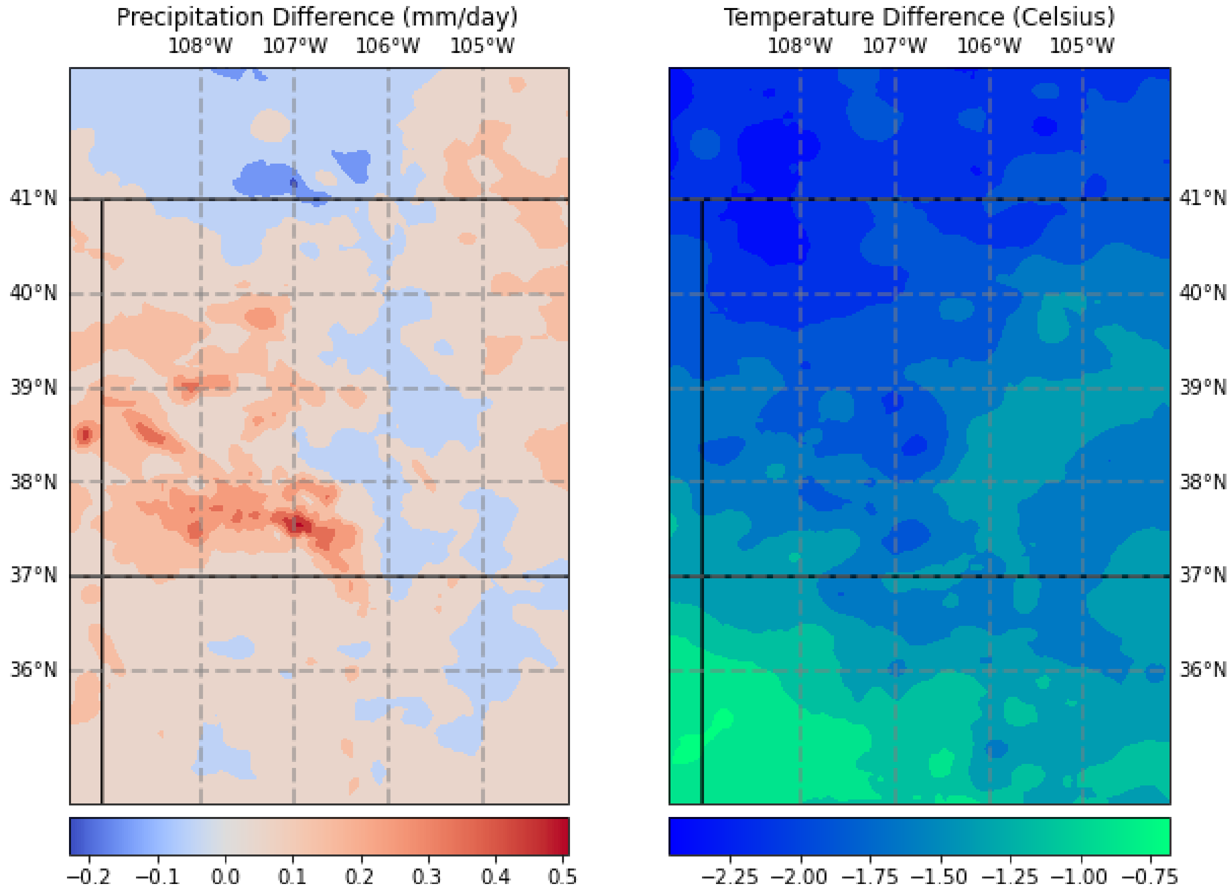

Figure S24: Differences of mean precipitation and temperature between GFDL-ESM2M and MIROC5. Difference is calculated as precipitation or temperature from GFDL-ESM2M minus MIROC5.

Chapter 4 North American Monsoon Precipitation and Extreme Precipitation Events

4.1 Background

Monsoons are continental-scale circulation systems that develop in response to seasonal changes in the thermal contrast between continents and adjacent oceanic regions (Vera et al., 2006). They are known for driving substantial regional precipitation, and are critical to the Earth's hydroclimate system. In this study we focus on the North American Monsoon (NAM) and examine the meteorological causes of both precipitation and extreme precipitation when the NAM is active.

However, actually defining the NAM region can be a challenge. Ramage (1971) used the reversal in the large-scale lower tropospheric circulation to identify the monsoon domain. This approach has been applied widely to define several monsoon indices, such as Webster-Yang monsoon index for the South Asian monsoon, the Australian monsoon index, the South Asian monsoon index and the dynamic Indian monsoon index (Goswami et al., 1999; Hung and Yanai, 2004; Wang and Fan, 1999; Webster and Yang, 1992). However, this circulation-based method is not suitable for the NAM region, since it does not exhibit the seasonal wind reversal that characterizes monsoons in other regions (de Carvalho and Jones, 2016). Precipitation has also been used to identify monsoonal regions. Liu et al. (2016) define global monsoon systems using the climatological precipitation difference between MJJAS (May-September) and NDJFM (November-March). If defined in terms of precipitation variability, the NAM region refers to the region roughly bounded by subtropical America through the southwestern US (Lee and Wang, 2014; Liu et al., 2016; Mohtadi et al., 2016; Wang et al., 2018). In the literature, the NAM also refers to a more localized area over

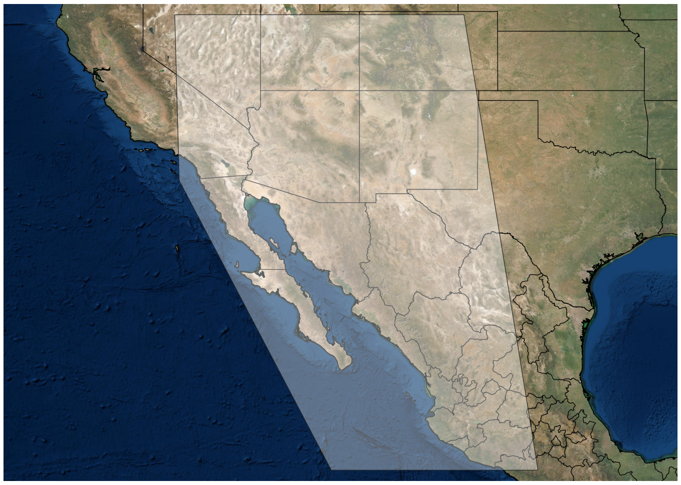

Figure 4.1: NAM regional domain from the North American Monsoon Experiment Forecast Forum.

the southwestern US and northwestern Mexico (Higgins et al., 2004). This particular region is much smaller and localized to the north of the region that emerges when defined via precipitation variability. The NAM Experiment (NAME) has defined a localized NAM domain as shown in Figure 4.1 (Higgins et al., 2006). In this study, the NAM region refers to this area encompassing southwestern US and northwestern Mexico.

Despite establishment of the NAM regional domain from the NAME, the regular trapezoid obtained from the latitude and longitude is infrequently used as an exact boundary. Instead, the term “NAM region” has been used to refer to a regular rectangular latitude-longitude box, or to specific states such as Arizona or New Mexico; this has especially been the case in climate change studies focused on long-term climatological precipitation signals (Cook and Seager, 2013; Douglas and Englehart, 2007; Finch and Johnson, 2010b; Varuolo-Clarke et al., 2019). Although these choices can simplify computations, such approximations are not appropriate for regional precipitation studies. Such structured regions cover areas with distinct precipitation mechanisms and drivers. This is especially true in the NAM area, where the complex terrain leads to precipitation being largely a product of orographic lifting (Boos and Pascale, 2021). As such, we argue that a delineation of the NAM region emphasizing localized precipitation features should be used for studies focused on NAM preciptiation.

Precipitation rate is frequently modeled using a gamma distribution (Martinez-Villalobos and

Neelin, 2019; Watterson and Dix, 2003). Extreme preciptiation events (EPEs), which occur when precipitation rate is in the long tail of this distribution, are of considerable importance for both scientific research, socioeconomic impacts, and water management. EPEs are generally defined as those periods when precipitation rate exceed a certain threshold, generally derived using one of two methods: a parametric approach or a non-parametric approach (Anagnostopoulou and Tolika, 2012). The parametric approach involves the probability density function of the tail distributions from the precipitation time series, and includes two popular methods: peaks-over-threshold (POT) and block maxima (Barlow et al., 2019). The POT method sets an initial threshold and fits the data with a generalized Pareto distribution (Acero et al., 2011), while block maxima focuses on the series of maximum values from a regular interval (such as maximum daily precipitation in each month), and fits the maximum data series with a generalized extreme value distribution (Alaya et al., 2020). The non-parametric approach doesn't make assumptions about the probability distribution of the data. It is often used with percentiles, such as the 95th percentile amount of rainy days (pq95) and the 99th percentile amount of rainy days (pq99) (Agel et al., 2018; Kunkel et al., 2012; Myhre et al., 2019). In this study, we adopt the non-parametric approach and define the threshold for EPEs from pq95.

To understand the meteorological causes of EPEs, Barlow et al. (2019) reviewed a set of potential meteorological systems for extreme precipitation over the North America, such as tropical cyclones, mesoscale convective systems, frontal systems and atmospheric rivers. Specifically for the NAM region, Kunkel et al. (2012) demonstrated the dominant role of frontal systems in summertime, and Sierks et al. (2020) revealed the occurrence of upper-level wave breaking with EPEs in the Lake Mead watershed. These studies provide candidate meteorological systems to comprehensively understand the drivers of NAM precipitation.

In this study, we first identify the NAM domain and its subregions from a gridded precipitation dataset to delineate regions based on local precipitation characteristics. The drivers of precipitation and EPEs are subsequently investigated using a linear decomposition method and a feature attribution method, respectively. This paper is organized as the following: Section 4.2 describes the precipitation and reanalysis datasets in this study. NAM domain identification and delineation are described in Section 4.3. Section 4.4 introduces the linear orthogonal decomposition method for decomposing precipitation time series and illustrates the LOD modes for mean precipitation. Section 4.5 introduces the candidate drivers of the NAM EPEs and corresponding detection methods or datasets. And lastly, the feature-based EPE analysis is connected with the LOD modes in Section 4.6.

4.2 Data

Precipitation data from the Climate Prediction Center (CPC) Global Unified Gauge-Based Analysis of Daily Precipitation (referenced to as the CPC dataset) is used in this study. It is based on gauge observations and provides daily precipitation analysis in a half degree resolution across the globe from January 1st 1979 to present (Xie et al., 2010). Leveraging previous research of the NAM, we select the area covering contiguous US (CONUS) and Mexico as our SOM clustering domain to identify the NAM region. Since the CPC dataset relies on gauge observations, the specific time period that defines a day varies across the globe. For CONUS and Mexico, they share the same time window: from 1200 to 1200 UTC. Meteorological conditions are derived from the ERA5 reanalysis dataset. This product provides hourly reanalysis atmospheric fields with a 30-km horizontal resolution (Hersbach et al., 2020). The record spans from 1950 to present, although we subset the period 1979 to 2018 to correspond with the precipitation data coverage. Additionally, when the hourly data is averaged to derive daily records, the time window is set to 12Z-12Z to keep accord with the CPC precipitation time interval.

4.3 Identification of NAM Subregions

4.3.1 Self Organizing Maps

Self organizing maps (SOMs) have been applied in previous studies for pattern recognition. For example, Agel et al. (2018) used SOMs with tropopause pressure anomalies to find the large-scale patterns associated with extreme precipitation. As for precipitation region identification, Swenson and Grotjahn (2019) used SOMs to classify different precipitation regimes over the CONUS. In our study, we use a method analogous to Swenson and Grotjahn (2019). Following Stidd (1953), we first take the cube root of precipitation to transform it from a highly skewed distribution to an approximately normal distribution. Then the long-term daily mean (LTDM) is calculated, excluding leap days. The LTDM is normalized to the range from 0 to 1 before training the SOMs following

$$\mathrm{LTDM}_{\mathrm{normalized}} = \frac{\mathrm{LTDM} - \min(\mathrm{LTDM})}{\max(\mathrm{LTDM}) - \min(\mathrm{LTDM})}. \tag{4.1}$$

This preprocessing informs us of the occurrence of extreme precipitation, rather than the actual precipitation amount.

The number of output nodes (i.e., the number of clusters) is prescribed before training SOMs. Since there is no prior knowledge of the correct number of clusters, to avoid arbitrariness and ensure robustness, an ensemble method is employed with the number of nodes ranging from 10 to 20. The final NAM region is then based on the intersection of all the ensemble results.

4.3.2 NAM Domain and Subregions

The SOM method doesn't ensure geographical continuity, so any singular grid point is manually added to the final region. Figure S1 depicts the SOM ensemble results for NAM domain identification. Although the cluster boundaries vary with the number of clusters, the general locations and patterns are consistent among all the SOM results. Compared with Figure 7 in Swenson and Grotjahn (2019), this region has a similar boundary and covers all of Arizona and part of California, Nevada, Utah, Colorado and New Mexico. The differences in the west and north boundaries from their results are likely due to sensitivity of the method to the addition of grid points outside of the CONUS. Although the overall NAM domain emerges naturally from this SOMs analysis, further delineation of precipitation subregions is still necessary given the heterogeneous geographical and topographical characteristics in this domain. The same SOM method is again applied to the identified NAM region, but only the summertime precipitation (June, July, August and September) is used as SOM inputs. Figure 4.2 depicts the 7 subregions identified from the SOM along with their LTDM precipitation signals. Subregions 1 through 7 (Sub1-Sub7), respectively, refer

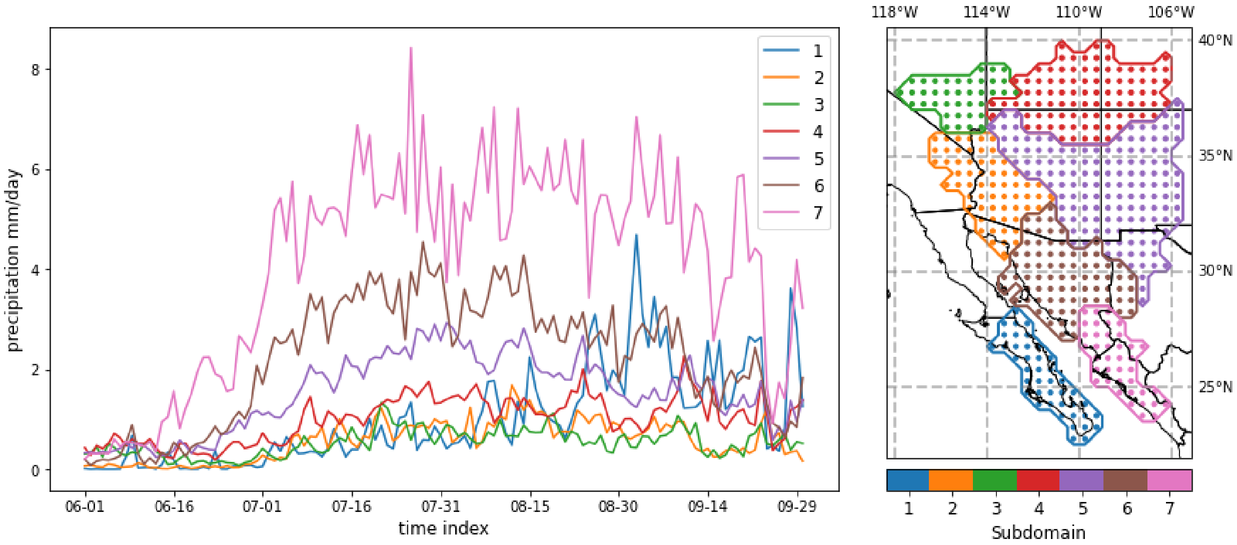

Figure 4.2: NAM sub regions and their long-term daily mean precipitation over summer season. The points denote the grid points from the CPC precipitation dataset.

to: (1) the southern half of the Baja California Peninsula; (2) Southeastern California, Northern Sonora and Eastern Arizona; (3) southwestern Utah and most of southern Nevada; (4) the Colorado Plateau and the 'Four Corners' region; (5) most of the Arizona desert, New Mexico and Northern Chihuahua; (6) most of Sonora; and (7) Southern Sonora and Northern Sinaloa. Comparing the LTDM precipitation, coastal areas such as Sub7, Sub6 and Sub1 are wetter regions, with higher overall precipitation rates, while the inland deserts are relatively drier (e.g., Sub2 and Sub3). The definition of monsoon onset date varies across previous studies. It is derived as the first day after June 1^{st} when precipitation rate exceeds 0.5 mm/day and lasts for 3 days in Higgins et al. (1997), while the threshold has changed to 1 mm/day and 5 consecutive days in Turrent and Cavazos (2009). The threshold difference is mainly due to the area of interest. Turrent and Cavazos (2009) examined the whole NAM area, whereas Higgins et al. (1997) focused on New Mexico and Arizona where the climatological precipitation is weaker. We adopt 1 mm/day and 5 days here, yielding median monsoon onset dates for Sub1-Sub7 of Aug 30^{th}, July 30^{th}, July 20^{th}, July 19^{th}, July 6^{th}, July 4^{th} and June 30^{th}, respectively. The onset dates are generally earlier for South subregions with Sub1 being a clear exception. The late onset date here is attributed to the impact of tropical cyclones (TCs), as demonstrated in the following sections.

4.4 Linear Orthogonal Decomposition

Our meteorological analysis of mean precipitation in this region focuses on the linear orthogonal models related to precipitation. This linear iterative method is helpful for decomposing the precipitation time series and correlating it to relevant meteorological variables.

4.4.1 Methodology

Linear orthogonal decomposition (Sukhdeo et al., 2022) proceeds as follows: In each iteration, a linear regression model is fitted between the time series of unpredicted precipitation and one candidate atmospheric variable. The residual from this linear model will be used as the new unpredicted component in the next iteration. The detailed process can be described as the following:

- Calculate the sample correlation between the precipitation to be predicted at this stage p^n and every atmospheric variable at each grid point. The value of the correlation field should range from -1 to 1.
- Identify the point of maximum absolute correlation x^n in field i^n. With the time series of p^n and x^n from i^n, we fit a linear regression model.
- With the fitted regression model, calculate the predicted component $\hat{p^n}$ and residual, which is defined as $p^n - \hat{p^n}$. The residual is used as p^{n+1} for the next iteration.

Here n represents the number of iterations, i stands for the input atmospheric variable and x for the most correlated grid point in i. After N iterations, we will get $\hat{p^N}$ and its corresponding residual p^{N+1}. The final prediction will be the summation of all the predicted components:

$$\hat{p} = \sum_{n=1}^{N} \hat{p^n} = p^1 - p^{N+1}. \tag{4.2}$$

For each predicted component $\hat{p^n}$, the correlation matrix of candidate atmospheric variables other than i^n with respect to the time series of i^n at grid point x^n is derived. This correlation map represents the large-scale pattern that is most correlated with the point x^n in i^n, which is equivalent to the associated predicted precipitation $\hat{p^n}$. Thus, these correlation maps can be viewed as the 2-dimensional projections of the time series of $\hat{p^n}$ on each candidate variable, and they represent the large-scale modes that are linearly correlated with precipitation.

R^2 score is used to quantify the predictability of the linear models and is given by

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \qquad (4.3) \\ &= 1 - \frac{\sum (p^1 - p^{N+1})^2}{\sum (p^1 - \bar{p^1})^2}. \qquad (4.4) \end{aligned}$$

Obviously the RSS, or the residuals will always decrease and R^2 score will increase with the number of iterations, however, the linear model may overfit the data and the corresponding linear modes are not generalizable. To avoid overfitting and assess the robustness of the linear decomposition modes, Sukhdeo et al. (2022) extracted and compared the linear modes from several independent datasets, and the modes are robust if different datasets yield similar patterns. In our study, instead of using multiple datasets, a 5-fold cross validation approach is used along with the linear decomposition. The cross validation approach splits the precipitation time series randomly into 5 folds along with its associated atmospheric data. Each time, 4 of the folds are used to fit the linear decomposition models and the remaining fold is used for testing. After 5 times, each fold has been used as the testing set and we would obtain the predictions for all the samples, which are used for the overall skill assessment. With different training data, the linear decomposition modes show some variation. A particular mode is considered to be robust only if different training folds generate similar patterns.

4.4.2 Large-Scale Modes from Linear Decomposition

The vertical integral of vapor transport (IVT), 500 hPa geopotential (Φ500), 850 hPa specific humidity (Q850), sea level pressure (SLP), 200 hPa potential vorticity (PV200), and total column water vapor (TCWV) are used as predictors in the linear orthogonal decomposition. IVT is a vector with two components: eastward and westward transport (IVT-E and IVT-N). To extract the influence from the GOC more intuitively, it reconstituted into vapor transport along the Gulf

(IVT-a) and perpendicular to the Gulf (IVT-b).

As described in section 4.4.1, 5-fold cross validation divides samples into 5 folds randomly, so each fold has different training samples and probably results in different linear modes. It could happen that the ith iteration of the xth fold yields a mode that is similar with the jth iteration from the yth fold where $x \neq y$ and $i \neq j$. To ensure the robustness of linear modes, the selected modes can be from different iterations to keep the similarity across all the folds. With the selected iterations of each fold, the corresponding predicted precipitation is used to evaluate the predictability quantified with R^2 score. Table 4.1 lists the R^2 scores for daily, 3-day and 5-day mean precipitation. Since the 5-day averaged precipitation is smoother, both the overall R^2 score and the R^2 for each mode are higher than daily and 3-day mean precipitation. The inland subregions such as Sub4 and Sub5 tend to have larger R^2 scores compared with the coastal areas like Sub1 and Sub7. This performance difference indicates the precipitation mechanisms are more complicated over the coastal sub regions and exhibit relatively low predictability, while the linear patterns are more representative over the in-land areas.

The extracted modes for 3-day mean precipitation are depicted in Figure 4.3 and Figures S2 to S14 as examples. Modes for daily and 5-day mean precipitation can be found in the supplement (S15-S41). Referring back to section 4.4, the subfigure titled with 'Mode' represents the variable i^n having the largest absolute correlation coefficient with the precipitation to be predicted in iteration n (p^n). The point of the largest correlation coefficient is labeled with a yellow plus sign, which is x_n, and the rest subfigures are correlation maps between fields other than i^n and the time series of i^n at point x_n. Comparing all subregions, the first modes or i^1 are mostly associated with high local water vapor content in either the Q850 or TCWV field. While Sub1 is the only exception, with IVT-N as the identified i^1, the moisture transport and water vapor fields are highly correlated. All subregions exhibit strong onshore moisture transport with positive correlation coefficients for IVT-a and IVT-N. This moisture transportation supports high values in the Q850 and TCWV fields. Along with the vapor transport, a strong sea level gradient emerges over the GOC with a low center to the southwest of Baja California and a ridge over the GOC coast. This strong pressure gradient has been identified to be favorable for GOC moisture surges (Bordoni and Stevens, 2006), which

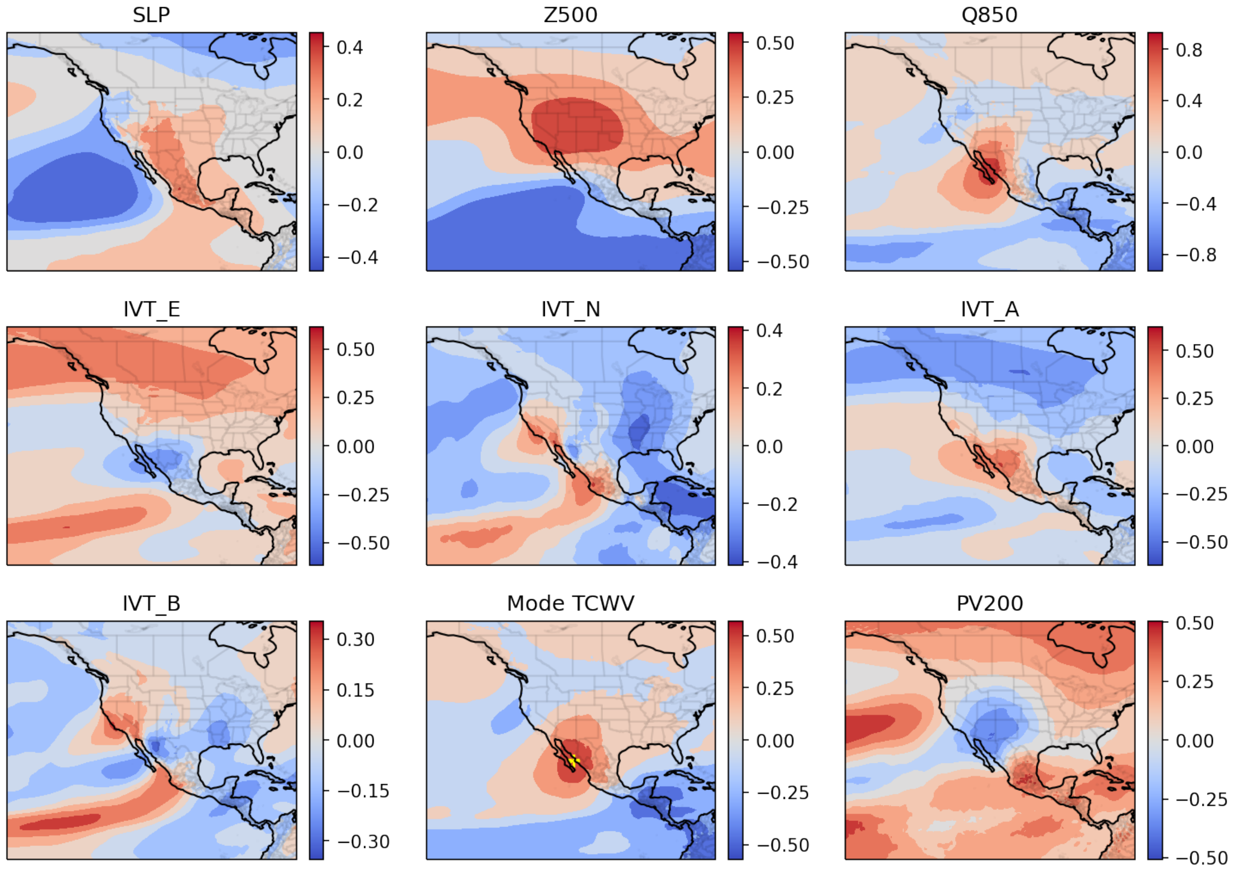

Figure 4.3: The first mode for 3-day averaged precipitation of subregion 7.

again enhances local water vapor content. Most fields share similar patterns across all subregions, while Z500 exhibits a unique pattern for Sub1 (again likely indicative of TC activity). The positive correlation coefficient of Z500 in Sub2 to Sub7 shows the general location of NAM ridge, which is a typical feature in the NAM season. Its location and shift have been proven to affect NAM precipitation (Adams and Comrie, 1997; Seastrand et al., 2015). For Sub1, the Z500 field reveals a local negative center, indicating the strong tropical cyclone impacts over this area (Breña-Naranjo et al., 2015; Farfán et al., 2014).

The second modes are associated with moisture transportation (IVT-N or IVT-b) for subregions besides Sub1, while Sub1 is associated with Q850. As with the first modes, the water vapor fields (Q850 or TCWV) are strongly correlated with the moisture transport, while the transportation direction accords with the local water vapor. Compared with the first mode that Z500 shows a smoother spatial gradient, there are more organized dipoles for Z500 and PV200 in the second modes which correspond with stronger spatial gradients. The PV200 anomalies are associated with the upper-level disturbances such as PV streamer in Martius et al. (2006) and Rossby wave breaking in Sierks et al. (2020). The Z500 anomalies are mid-troposphere features that favor the moisture transportation, providing moisture for precipitation. Again subregion 1 is an exception with the second mode showing the general location of the monsoon ridge.

Comparing the R^2 scores from different iterations, the most explained variance is associated with the first mode (32-58% for 3-day mean precipitation), with a smaller increase in R^2 score after adding the second (2-10%) and the third (1-2%) modes. The linear modes start to diverge among the 5 folds after 2 or 3 iterations. This sensitivity to data sampling suggests some differences among the upstream process drivers of precipitation by region. In essence, with the few years of data available in ERA5, only these 2-3 modes are distinguishable from climate variability.

Table 4.1: Linear decomposition R^2 scores of daily, 3-day and 5-day mean precipitation. Only the modes that are consistent across 5-fold validation are calculated.

daily	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7
Mode1	0.225	0.253	0.315	0.412	0.455	0.281	0.172
Mode2		0.050	0.044	0.116	0.067	0.072	0.058
Mode3		0.022	0.008	0.021	0.023		
3-day	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7
Mode1	0.350	0.400	0.420	0.528	0.576	0.445	0.321
Mode2	0.026	0.010	0.040	0.101	0.058	0.096	0.105
Mode3		0.027	0.004	0.021	0.015	0.009	
5-day	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7
Mode1	0.387	0.479	0.493	0.592	0.645	0.534	0.398
Mode2	0.031	0.041	0.042	0.083	0.043	0.081	0.115
Mode3	0.031	0.029			0.020	0.002	0.009

4.5 Synoptic and Mesoscale Features as Drivers for EPEs

4.5.1 EPE Definition

EPEs are herein defined as days when daily subregion-mean precipitation rate exceeds the 95th percentile of rainy days (precipitation rate larger than 1 mm/day). When there are consecutive days

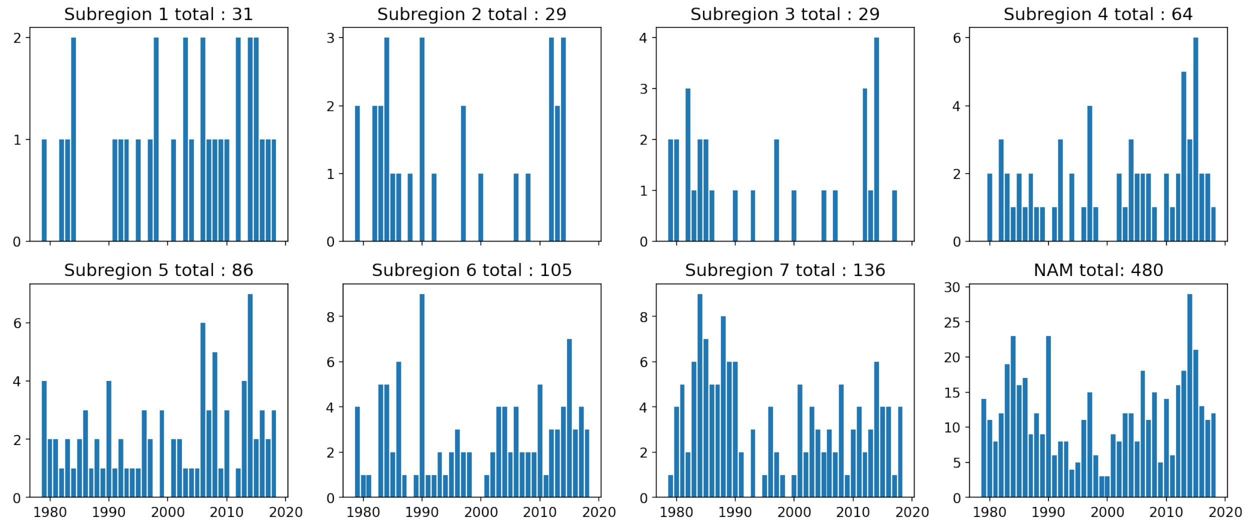

Figure 4.4: Number of extreme events in each sub region for each year and total number of extreme events.

exceeding this threshold, they are consolidated into a single event. Figure 4.4 shows the number of EPEs in each subregion from 1979 to 2018. Since the coastal regions have more rainy days, following our criteria, they also have more EPEs. A Mann-Kendall (MK) test is applied to each subregion to see if there is a historical trend in the number of EPEs, EPE precipitation amount and EPE precipitation rate each year from 1979 to 2018. This test has been shown to be effective in detecting monotonic trends in precipitation analysis (Wang et al., 2020a).

In most subregions, there are no significant trends under the 5% confidence level, however, EPE numbers and precipitation amount does exhibit a significant increase in Sub1 and Sub6. Sub1 also shows a rising trend in EPE precipitation rate, while Sub2 shows a declining trend. These changes are likely due to the influence of low-frequency climate variability.

4.5.2 Selected Features

For the purposes of identifying the process drivers of EPEs in the NAM region, we select and examine five synoptic features and one mesoscale feature: tropical cyclones (TC), Gulf of California moisture surges, upper troposphere troughs (UTT), frontal systems, mid-tropospheric lows and mesoscale convective systems (MCS). The following subsections introduce each feature and corresponding procedures to link these events with EPEs.

Tropical Cyclones

Tropical cyclones (TCs) play an essential role in the global hydroclimate system. They transport significant water vapor from the tropics and sub-tropics, and account for a large fraction of EPEs around the world (Zhao, 2022). In the NAM region, previous studies have revealed that TCs are largely responsible for precipitation over Baja California and Northern Mexico (Díaz et al., 2008; Englehart and Douglas, 2001). In this study, TC tracks from International Best Track Archive for Climate Stewardship (IBTrACS) are used. IBTrACS provides 3-hour intervals for TC locations and intensities around the world from 1842 to present (Knapp et al., 2010). We exclude tropical depressions (TDs) from this analysis, selecting only tropical storms (TSs), tropical cyclones (TCs), and hurricanes (HRs). An TC is linked to an EPE if its track is within a 5-degree radius of the given NAM region. This distance criterion is based on the general horizontal scale of TCs (Dominguez and Magaña, 2018; Jiang and Zipser, 2010; Kunkel et al., 2012).

Gulf of California Moisture Surges

As discussed in section 4.4.2, precipitation variability in the NAM region is strongly connected with northward surges of vapor transport along the Gulf of California (GOC) (Bordoni and Stevens, 2006). GOC moisture surges boost continental humidity, provide the necessary water vapor for precipitation, and decrease the stability of the environment. In Dominguez et al. (2016), a simulation using the Weather Research and Forecasting Model with water vapor tracer diagnostics (WRF-WVT) examined the origins of water vapor that contributes to precipitation during the NAM season. The sources were divided into four regions: two marine sources including Gulf of Mexico (GOM) and GOC, and two terrestrial sources including Sierra Madre and the NAM region, defined as regions in the east of Sierra Madre. Their 10-year simulation emphasized the importance of moisture from the GOC, which accounts for 16% of NAM precipitation in summertime, followed by the local NAM source with 15% contribution.

GOC moisture surges are identified using the vertical integral of northward and eastward vapor

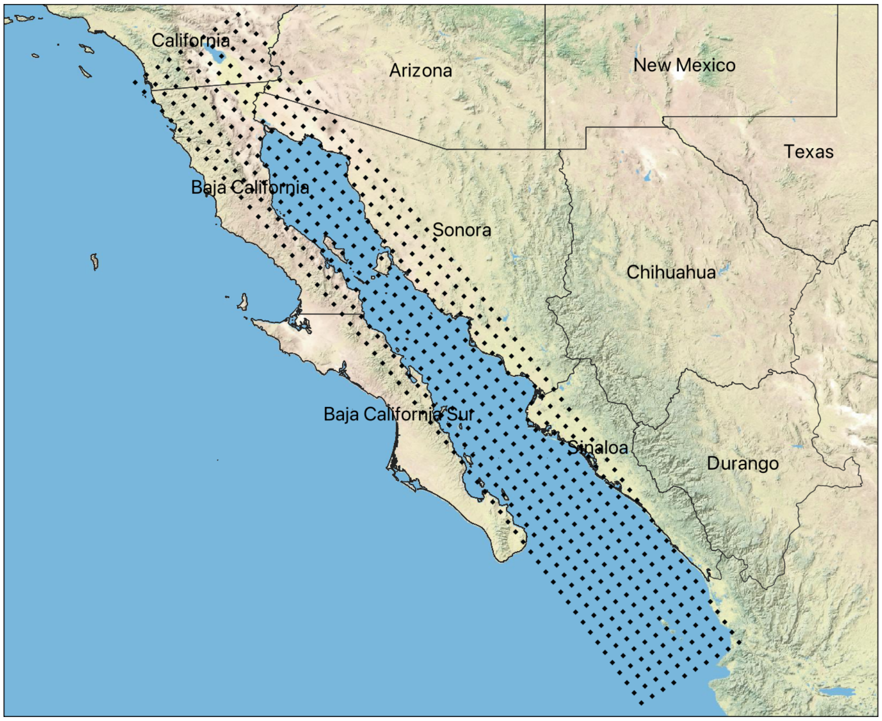

Figure 4.5: The GOC transect grid points.

flux (denoted as IVT-N and IVT-E) from ERA5 6-hourly reanalysis data. Figure 4.5 shows the GOC transect with grid points aligned along the gulf in a 25-km spatial resolution. The northward and eastward fluxes are reconstructed as fluxes parallel to and perpendicular to the GOC transect, and the grid points along the perpendicular axis are averaged to derive a one-dimensional flux profile along the Gulf. Surge candidates are defiend as fluxes that surpass the 95^{th} percentile of each grid point. The spatio-temporally consecutive candidate grid points are then characterized as a surge event, which must last at least 12 hours. The detection method is illustrated in Figure 4.6 with four surge events as an example. An EPE is classified as a surge event with a specific time window threshold, and this time window is determined for each subregion using a composite analysis.

Figure 4.7 shows the precipitation anomalies with respect to the surge occurrence. The x-axis denotes days after the onset of surges with negative values representing days before the surge and positive for days after the surge. Zero denotes the surge onset date. Most subregions show precipitation peaks after 2 or 3 days of the onset date while Sub7 shows double peaks with the first

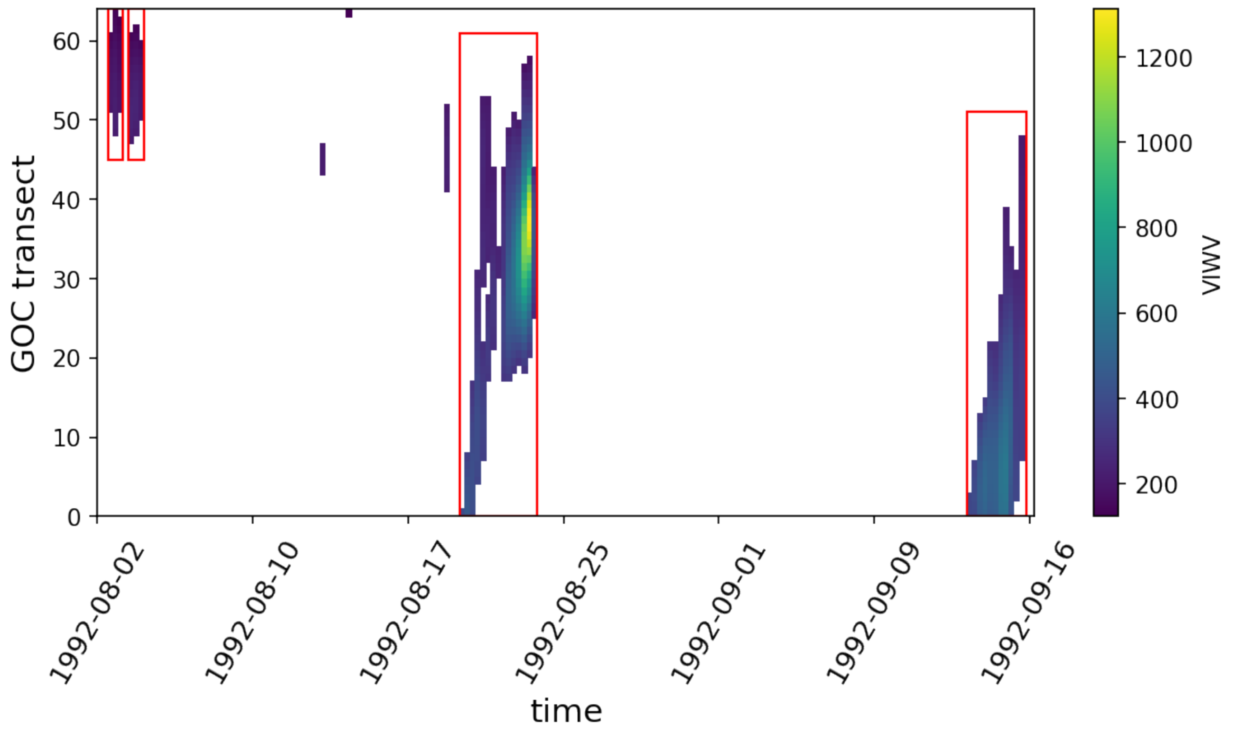

Figure 4.6: Examples from 1992 GOC surge detection result in Hovmoller diagram. Only the candidate surge grid points are shown. A surge is identified as a continuous band in the figure, and is denoted with a red box. Four surge events are identified in this figure. The specific candidate grid points are not included.

one on the onset date, which is probably due to its location at the southern end of the GOC. In addition, the precipitation anomaly is negative on the onset date in Sub3, Sub4 and Sub5. It is probability due to their far distance from the GOC Gulf. To associate EPEs with GOC surge, we check if there is a moisture surge within a specific time window before the EPEs. The window size is set to 1-day for Sub2, 2-day for Sub1, Sub3, Sub4, Sub5, 3-day for Sub6 and concurrent day for Sub7.

Upper Troposphere Troughs

Upper troposphere troughs (UTTs) are upper-level circulation patterns with a local low geopotential height and high potential vorticity around 200 hPa (Kelley and Mock, 1982). Among the various types of UTTs, Rossby wave breaking events (RWBEs) and inverted troughs (ITs) are two of the

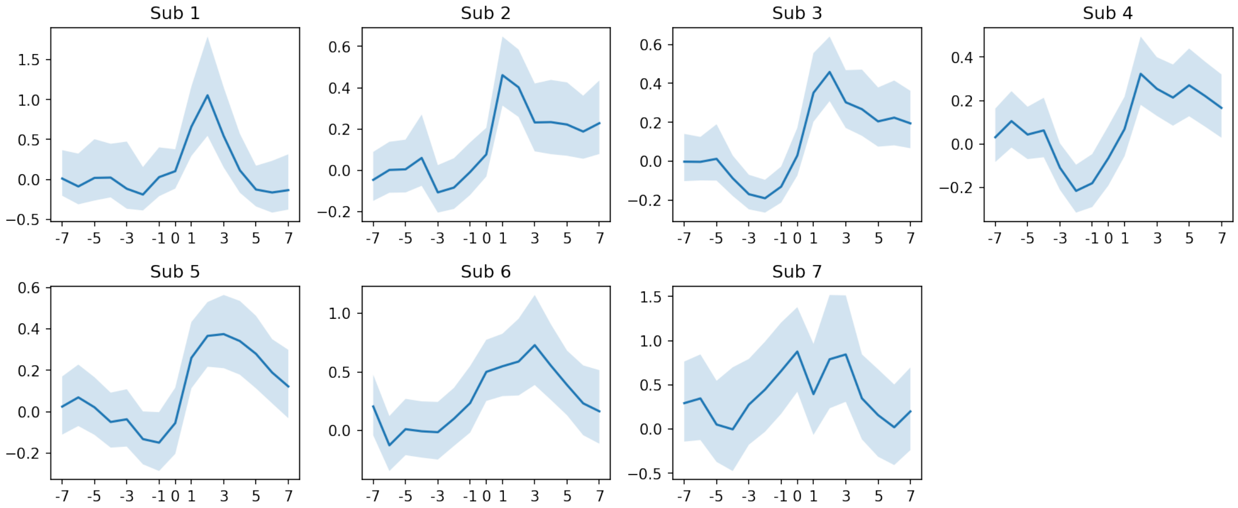

Figure 4.7: Precipitation anomaly composites of GOC Surges. Shading indicates the 95% confidence intervals generated by bootstrapping.

most common features employed in precipitation analysis. RWBEs are often characterized by a reversal in the latitudinal PV gradient near the tropopause (Zavadoff and Kirtman, 2019). When the length-width ratio of PV overturning is large, it is also referred to as PV streamer (Papin et al., 2020). The effects of RWBEs and ITs on precipitation in the Lake Mead Watershed were explored in Sierks et al. (2020). RWBEs events have also been linked to precipitation in Ryoo et al. (2013), who showed a strong correlation between PV200 and precipitation. Moore et al. (2019) links EPEs with RWBEs. According to their findings, the majority of EPEs in central and eastern United States are associated with concurrent PV streamers from RWBEs. Aside from RWBEs, another common upper-level disturbance that affects precipitation is IT. An IT is a trough with pressure increasing toward the poles, which is opposite to the most common mid-latitude troughs. For the NAM region, tropical upper-troposphere troughs are the most common IT type. TUTTs, unlike RWBEs, are more common in subtropical easterlies, albeit they are also connected to mid-latitude wave breaking events (Igel et al., 2021). In terms of its impact on precipitation, Finch and Johnson (2010a) utilized quasigeostrophic (QG) theory to study a TUTT event over the NAM region in July 2004, while Newman and Johnson (2012) used WRF to simulate the same event. Their results showed wind shear and convective available potential energy (CAPE) both increased during the TUTT event, particularly to the west of the TUTT. TUTT-induced convective enhancement was also identified in Bieda III et al. (2009), where it was shown that lightning event density increases when a TUTT is present. Interactions betweeen TUTTs and TCs and RWBEs were also investigated in Wang et al. (2020b). A comprehensive TUTT dataset was built based on the 200 hPa stream function in Igel et al. (2021), and their composite analysis showed an enhancement in precipitation to the southeast to the TUTT core.

The wide variety of upper level disturbances (RWBEs, PV streamers, TUTTs, ITs) all exhibit a local high in potential vorticity at tropopause, commonly approximated by 200 hPa level. In this study, UTT candidates are first identified as closed contours of $2 \times 10^{-6} m^2 s^{-1} K kg^{-1}$, or 2 PVU from the ERA5 6-hourly 200 hPa potential vorticity by TempestExtremes. The prospective candidates are then filtered out with any TCs at the near surface level to ensure that they are upper-level disturbances. UTTs are linked with NAM region EPEs if the UTT occurs within 10 degrees of the NAM subregion, a choice made based on precipitation anomaly composites.

For each tracked UTT, the precipitation within a 20 degree radius from its center is taken and the long-term daily mean precipitation is subtracted to derive anomaly precipitation. The 95% confidence interval is derived with a two-sided Student's t test. As depicted in Figure 4.8, the precipitation is depressed to the north and northeast of the UTT center, and enhanced to the south and southeast. Within 10 degrees, the enhancement reaches its peak and diminishes with distance. Thus, 10 degrees is set as the criterion for UTTs: if there is a concurrent UTT in the 10 degree radius from the subregion, the EPE will be assigned to this UTT.

Frontal Systems

Frontal systems, especially in the mid-latitudes, promote precipitation by inducing uplift. Catto et al. (2012) described the importance of frontal systems for precipitation around the world, arguing that they are responsible for 46 percent of overland precipitation in the Northern Hemisphere. According to Kunkel et al. (2012), 44 percent of EPEs in the southwestern US summer are attributed to fronts.

Despite of the existence of automated identification methods for frontal systems, available

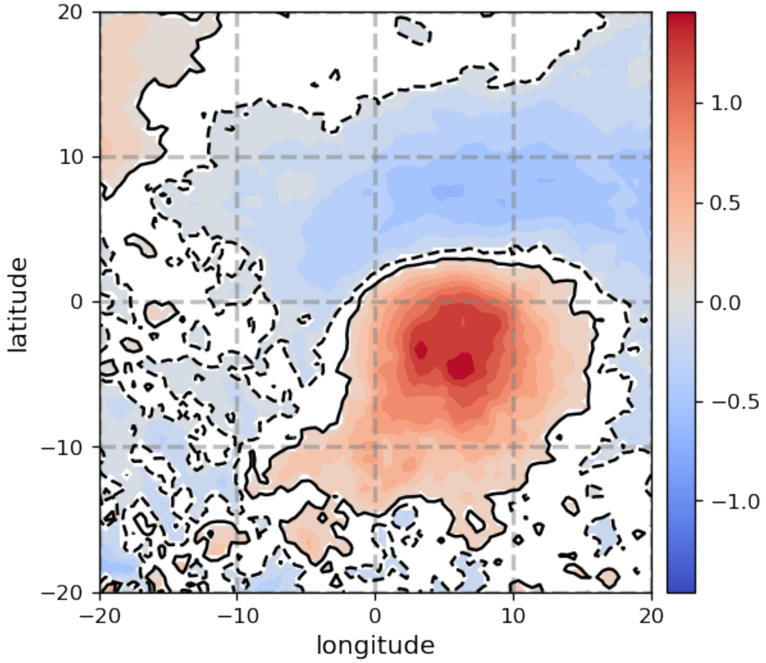

Figure 4.8: UTT-centered composites of precipitation anomalies with confidence level at 95%. Colors show the precipitation anomaly in mm/day. Solid and dash lines are for confidence interval contours.

schemes either require substantial computational power (Hewson, 1998), or are insufficiently validated over the NAM region (Biard and Kunkel, 2019; Parfitt et al., 2017). Instead of identifying fronts from reanalysis data, we instead use a manually labeled dataset from the National Weather Service (NWS) coded surface bulletins. From 2003, this NWS dataset provides the locations and types of frontal systems at 3-hour intervals, which are determined by National Weather Service meteorologist (Biard, 2019). To link EPEs with frontal systems, we used the method from Catto et al. (2012): If a concurrent front is 5 degrees or less away from the EPE area, this precipitation event is associated to front.

Mid-troposphere Lows

Often the moisture transport is driven by mid-tropospheric (i.e., 500hPa) disturbances that do not strongly manifest at the surface level or in the upper atmosphere. Wibig (1999) used 500 hPa geopotential height to identify circulation patterns related to winter precipitation over the Euro-

Atlantic sector. The atmospheric circulation patterns related with the EPEs over Greece emerged by analyzing the clustering results of 500 hPa geopotential height meteorological fields in Houssos et al. (2008). In this study, we detect anomalous lows at the 500 hPa level and examine their importance as a driver of EPEs. The composite mean of 500 hPa geopotential anomaly during EPEs is shown in Figure 4.9. The low centers are generally located to the west of the inland subregions, while the centers are not deep for coastal subregions, though they are all significant with a 95% confidence interval. Based on this analysis, the distance criterion is set to 5-degree and -1000 m^2/s^2 as the threshold for Φ500 anomaly: where a concurrent Φ500 anomaly low stronger than -1000 m^2/s^2 is less than 5 degrees away from the subregion, the EPE is classified as mid-troposphere low.

Mesoscale Convective Systems

Mesoscale convective systems (MCS) drive are significant drivers of global precipitation (Zhao, 2022). Specific to the NAM region, Finch and Johnson (2010b) and Mejia et al. (2016) used observation records to show that MCS activity increases over the summer in the NAM region. While MCSs are difficult to resolve in modern reanalysis data, a variety of observational products possess sufficiently high resolution to enable MCS detection. Feng et al. (2021) tracked MCS globally based on infrared brightness temperature and precipitation from satellite datasets. They selected MCS candidates based on brightness temperature and precipitation. In this study we analyzed a subset of this tracking data covering the NAM region. However, since this dataset only begins in 2001, we only employ this dataset in conjunction with the frontal systems dataset which begins in 2003. A MCS event is deemed to be associated with an EPE only if there are labeled MCS grid points inside the precipitating area.

4.5.3 EPE Feature Drivers

Since the frontal system record starts from 2003, only TCs, UTTs, GOC surges and mid-troposphere lows are considered for EPEs before 2003, and fronts and MCS are included for events after 2003. Figure 4.10 and 4.11 shows the precipitation amount fraction with different drivers for EPEs before

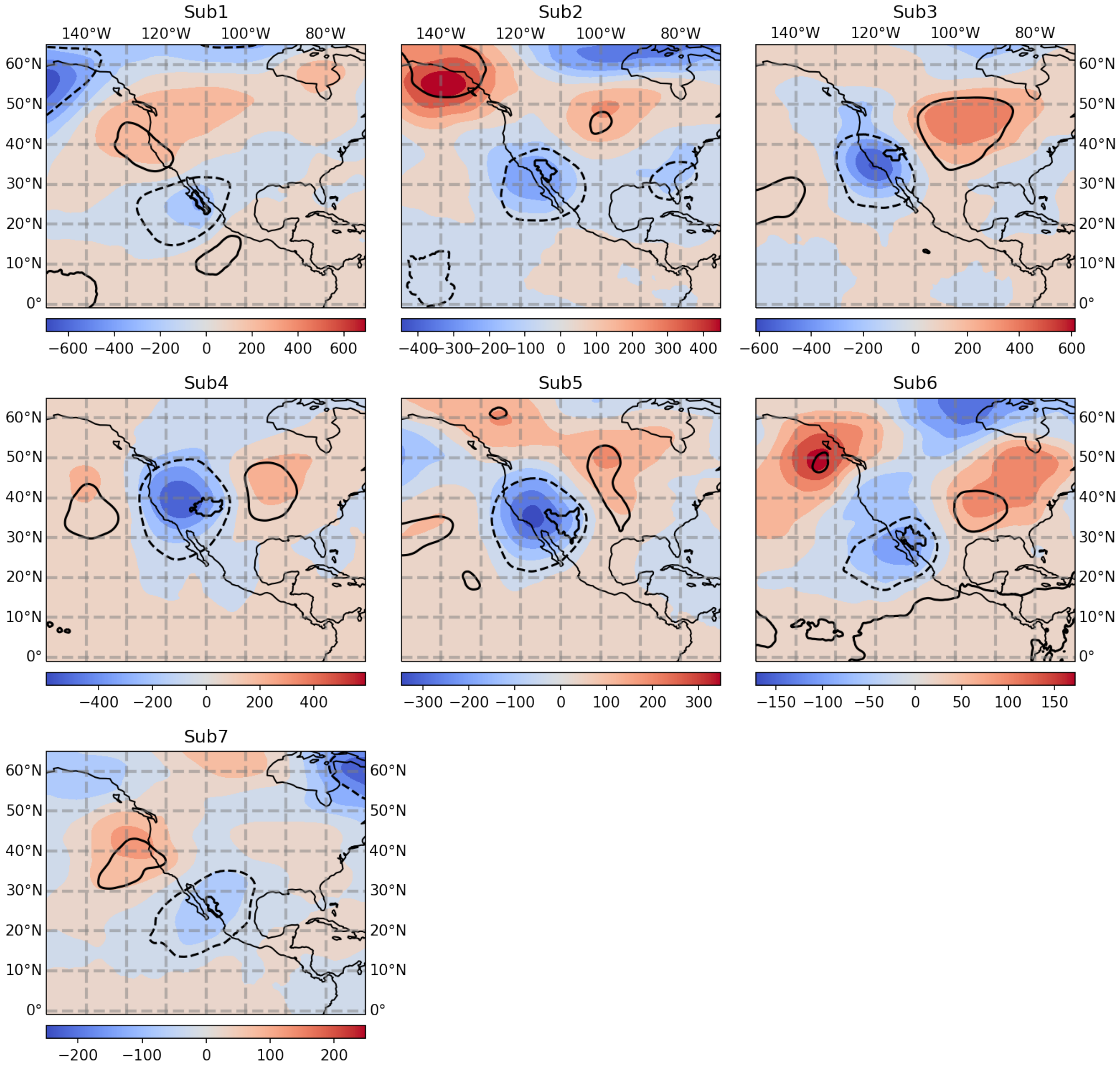

Figure 4.9: EPE 500 hPa geopotential anomaly composites. Black contours denote the 95% confidence interval (solid line for positive anomalies and dashed line for negative anomalies).

and after 2003. The fraction of EPE numbers associated with the candidate drivers are depicted in Figure 4.12 and 4.13. The events that are not linked with any candidate drivers are denoted as ‘Unclassified’ (abbreviated as ‘Unclass’). Although there are several ’Unclassified’ events before 2003, the inclusion of frontal systems and MCSs produces only one remaining ‘Unclassified’ event.

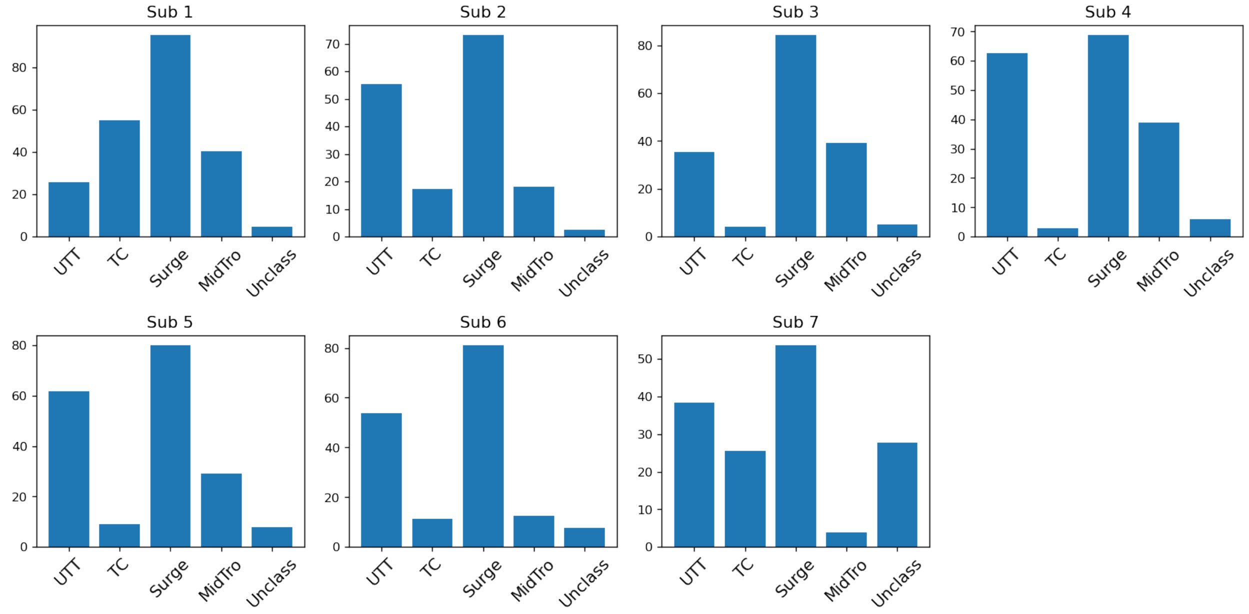

Figure 4.10: EPE precipitation amount (%) associated with different feature drivers before 2003. Since a given EPE could be associated with more than one feature, the percentages do not add up to 100%.

For most subregions, GOC surges and fronts are the two leading drivers, account for both more relevant events and precipitation amount. TCs have greater impacts on Sub1 and Sub6, and MCSs dominate Sub5 and Sub7. Mid-troposphere lows are more frequent drivers of EPEs over inland subregions (Sub3, Sub4 and Sub5) than coastal areas, which is consistent with Figure 4.9 where the geopotential low is more pronounced in these subregions.

It should be noted that the feature classification in Figures 4.10 and 4.11 is not exclusive (i.e., a UTT event can also be linked with other drivers like GOC surges or MCS). Combined events (i.e., two features simultaneously) are further investigated with EPEs after 2003, since all but one of the EPEs can be assigned to at least one candidate driver. The results are illustrated in Figure 4.14. In general, most of the EPEs are caused by two to three drivers. However, there are fewer

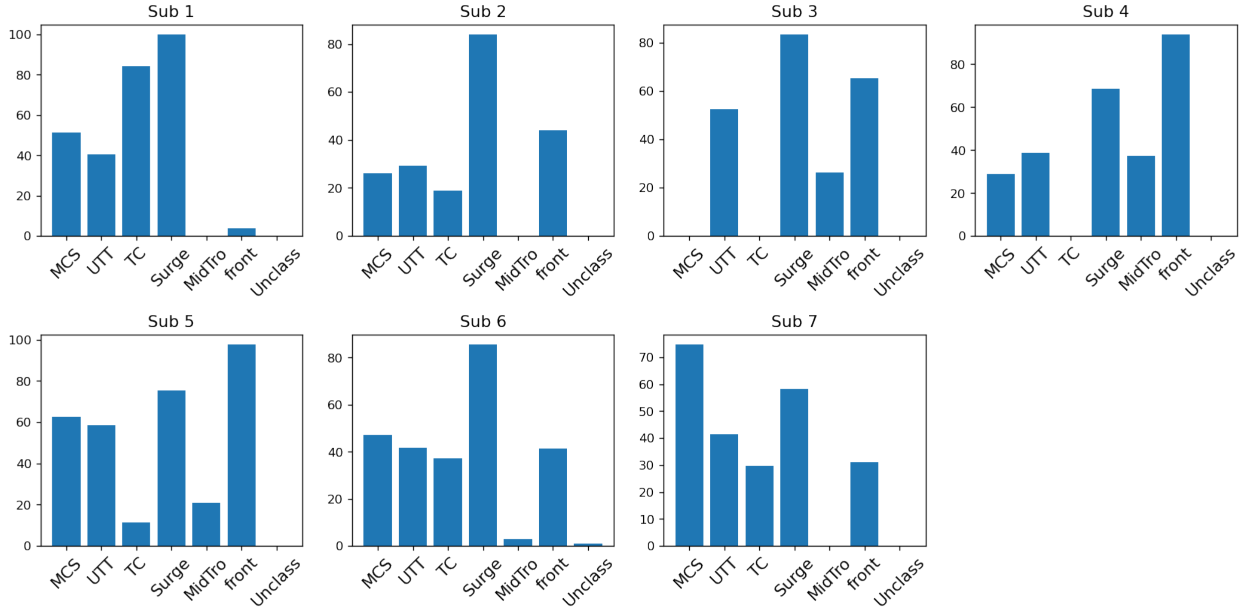

Figure 4.11: EPE precipitation amount (%) associated with feature drivers after 2003. Since a given EPE could be associated with more than one feature, the percentages do not add up to 100%.

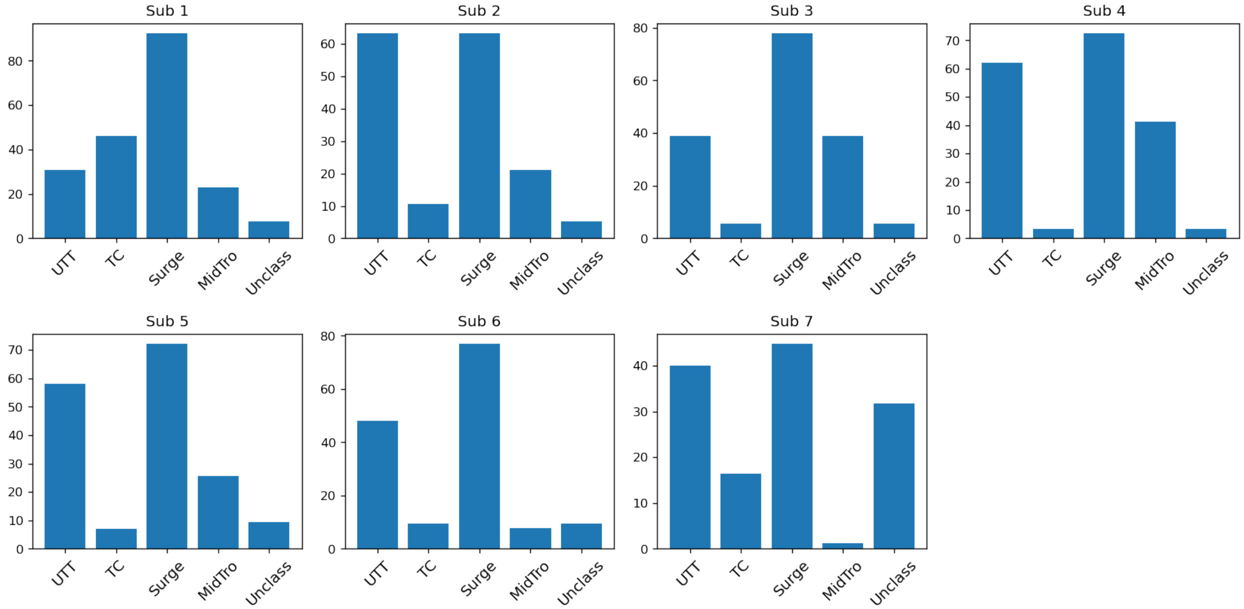

Figure 4.12: EPE occurance percentage associated with different drivers before 2003.

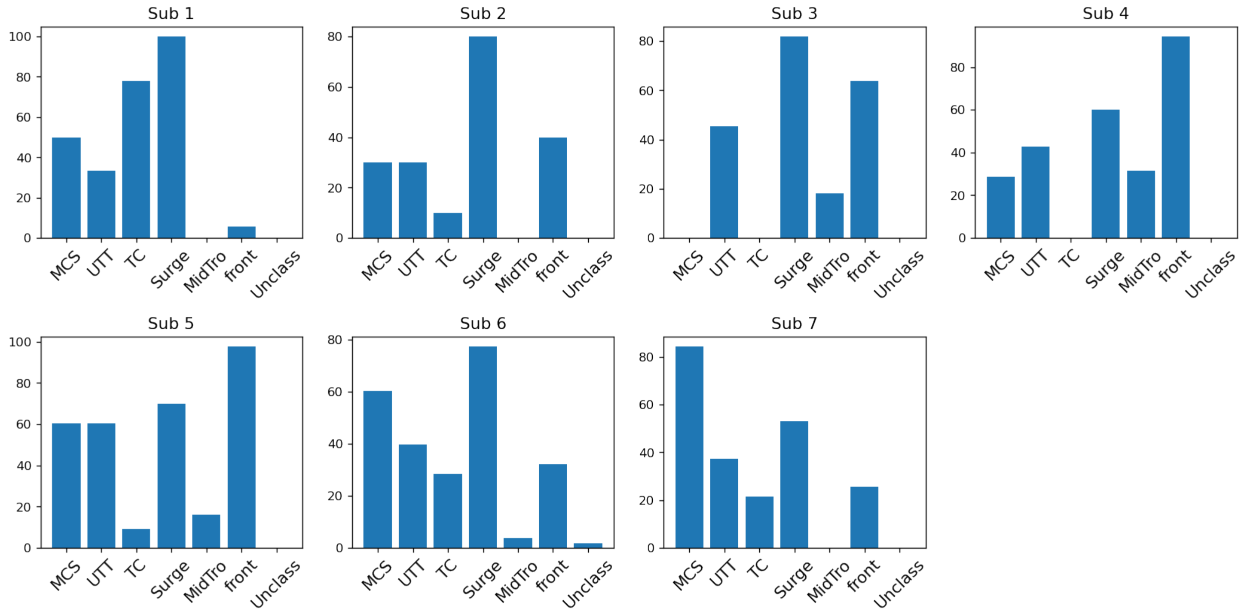

Figure 4.13: EPE occurance percentage associated with different drivers after 2003.

categories in Sub1, Sub2 and Sub3, while the interactions are more complex in Sub6 and Sub7. Perhaps what stands out the most are those events induced solely by a single driver. Particularly for Sub7, MCSs are the dominant driver of EPEs, with the MCS-alone precipitation exceeding 10%. This result indicates the importance of MCSs in this area for EPE attributions, and explains why this region suffered from a large percentage of 'Unclassified' events before 2003. Fronts are another feature unavailable in our analysis before 2003, and one that is particularly important over inland subregions (Sub2, Sub3 and Sub4), where the front-only precipitation exceeds 5%. Contrary to MCSs and fronts, TCs are an important feature for EPEs, yet they never occur by themselves; generally TC-related EPEs usually occur with GOC surges.

As showed previously, key EPE metrics (both number of EPEs and EPE precipitation amount) have increased in Sub1 and Sub6, while EPE precipitation rate has trended down in Sub2. Since we have now classified EPEs by feature type, the trends for each EPE category in these three subregions are further examined with the same MK test. Since 6 categories are being tested in the same time, a Bonferroni correction is applied to adjust the confidence level from 0.05 to 0.05/6. For the number of EPEs in each year, only the trend of TC-related EPEs is significant in Sub6,

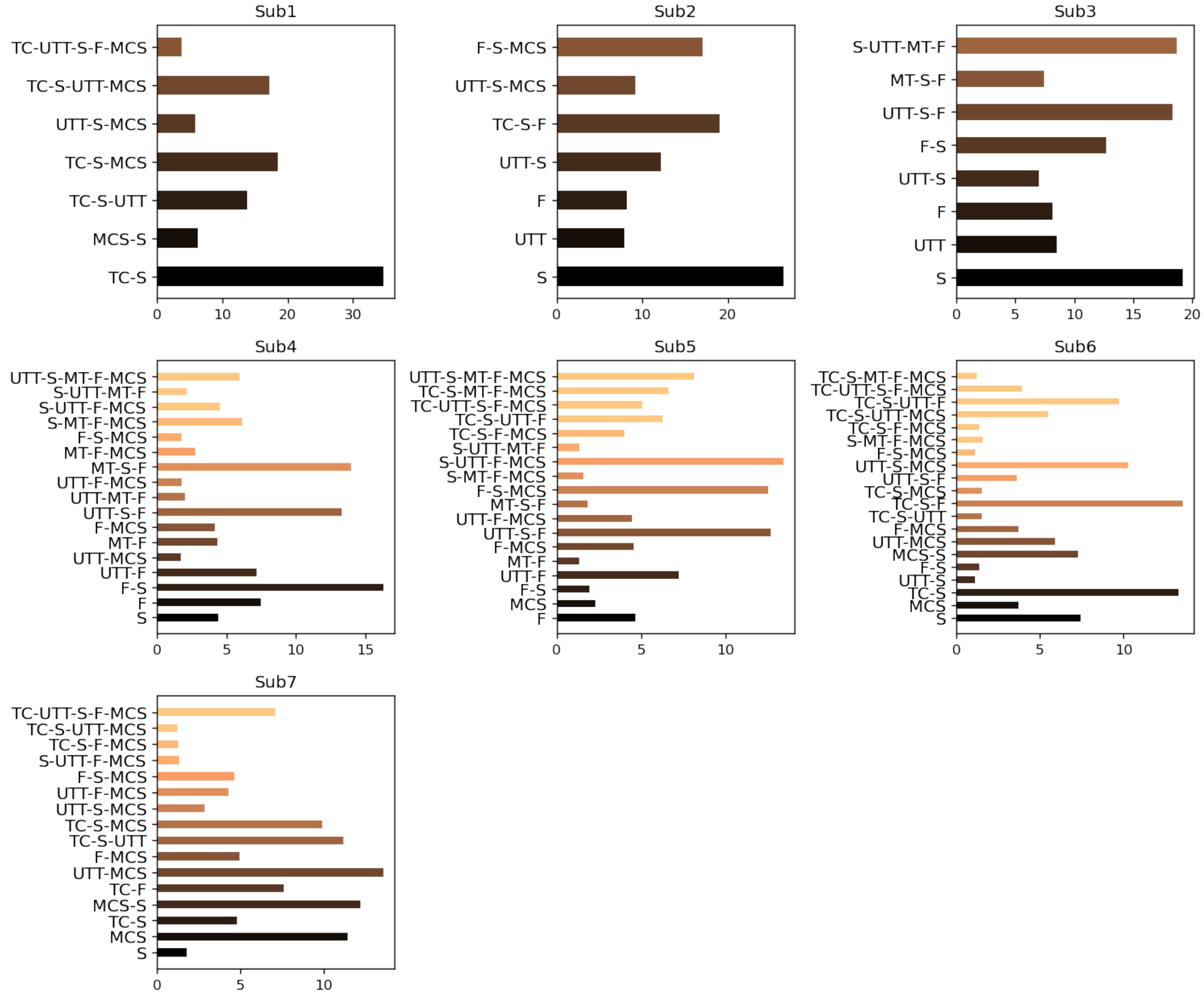

Figure 4.14: Combination of different drivers for EPE precipitation after 2003. Bar length represents the fraction of EPE precipitation amount. 'UTT' represents upper-troposphere troughs, 'F' for fronts and 'S' stands for GOC surges.

while there are no significant trends for other categories or regions. Although an upward trend of number of EPEs is found in Sub1, none of the EPE categories have increased significantly, likely due to the strict p-value from Bonferroni adjustment (Perneger, 1998). The trend for precipitation amount is only significant for TC-EPEs in Sub1 and Sub6, and there are no significant trends for the remaining categories. Only Sub6 exhibits an increasing trend for TC-EPE precipitation rate. This result suggests that the significant trends of EPE numbers and total precipitation in Sub6 are explained by an increase in TC-related EPEs and their associated precipitation rates. The increasing trend in TC-EPE precipitation rates is indicative of more intense TCs. This upward trend in TC-EPE numbers could be the result of more frequently TCs, which may be affected by low-frequency variability (Pazos and Mendoza, 2013), or global warming, (i.e., the observed increase over Baja California (Murakami et al., 2020) and increase trend in eastern North Pacific (Klotzbach et al., 2022)). But it is worth noting that although the increasing trend is significant in Sub6, the change rates are small with the Theil-Sen slopes being 0 and OLS slopes less than 0.01. A further careful analysis is required to better relate these trends with potential upstream drivers.

4.5.4 Meteorological Conditions for EPEs

The meteorological field composites for EPEs in each subregion are constructed to reveal the general circulation conditions. All the subregions show local high water contents in TCWV and Q850 fields, and they are mostly associated with strong moisture transport along the GOC channel. As we discussed in section 4.5.2, when GOC surge is onset, Sub4 shows negative precipitation anomalies because its far distance from the GOC. This is also observed in the concurrent composites where IVT-A shows negative anomalies for Sub4. Φ500 low center is always onset and all the subregions show upward lifting with negative Ω500 anomalies. Besides synoptic-scale uplifting, the positive CAPE anomalies indicate a convective uplift environment. Both the moisture and vertical ascent create a favorable environment for extreme precipitation. In spite of the common patterns of moisture and uplift motions, the upper-level disturbance exhibits different behaviors across the subregions: Sub1, Sub6 and Sub7 (coastal areas) show local anomalous low in PV200, while the strong gradient of PV200 with positive to the west and negative to the east is significant in Sub2, Sub3, Sub4 and Sub5 (inland areas). This difference indicates that UTTs (high PV200 contours) are more influential over Sub3, Sub4 and Sub5, which is consistent with the higher UTT-EPE precipitation fraction over inland areas in Figure 4.10 and 4.11. There are also magnitude differences across the subregions. Take TCWV and Z500 as examples, composite magnitudes are relatively larger for inland areas like Sub3 and Sub4 compared with Sub6 and Sub7 (Figure S42 and S43). This is probably due to the fact that MCSs are more dominant in Sub6 and Sub7, as shown in Figure 4.11. Its small horizontal and temporal scales make it hard to be resolved in daily reanalysis datasets and thus would result in lower anomaly magnitudes.

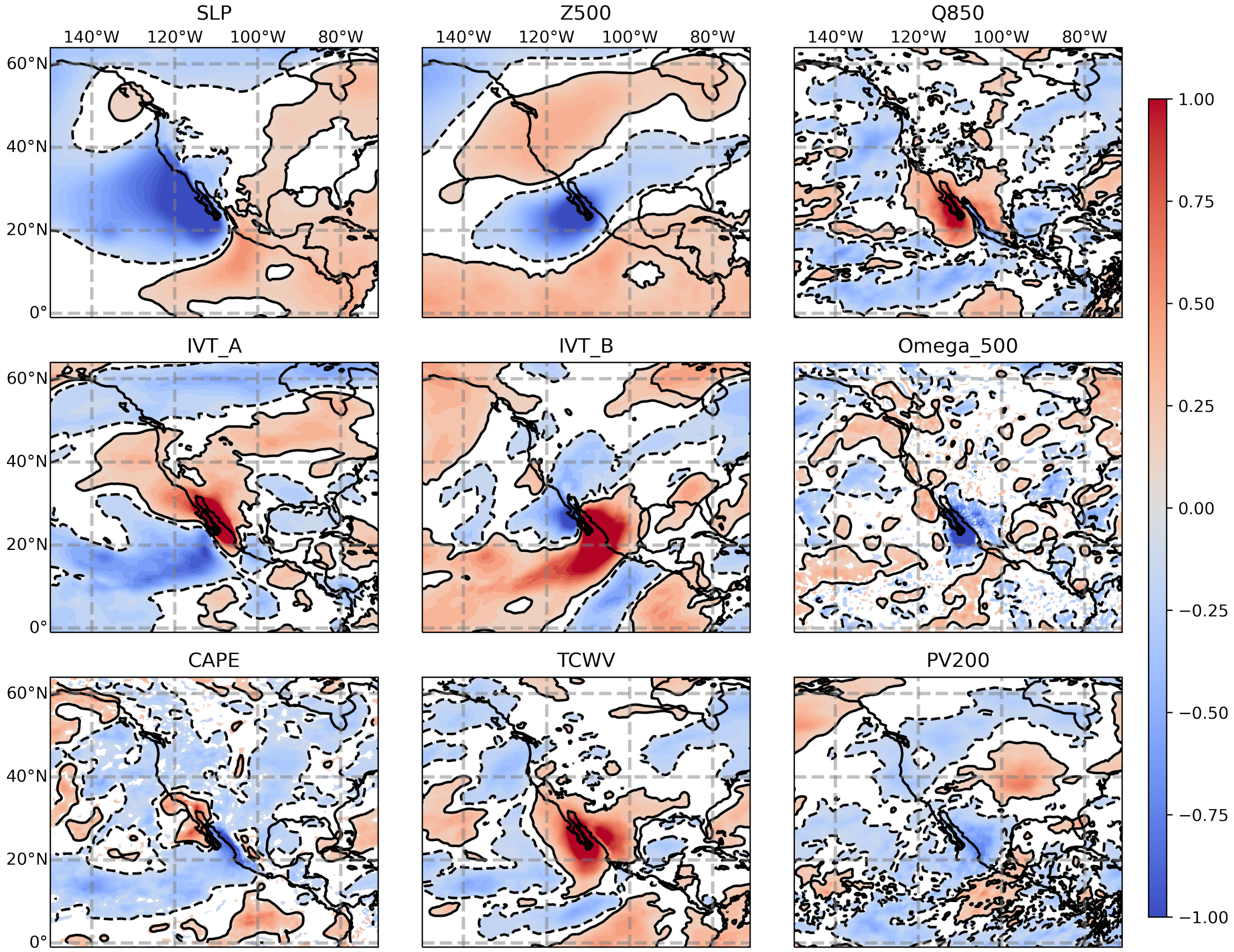

Figure 4.15: Composites of standard anomalies over EPEs in Sub1.

UTTs and mid-tropospheric lows

We now turn our attention to UTTs and mid-troposphere lows since they share some common features in PV200 and Φ500. Their composites both show anomalous high PV200 and Φ500 lows. Despite the similarities, no obvious closed contours are found for mid-troposphere lows, whereas significant closed contours are present in both PV200 and Φ500 for UTT composites. This is likely related with their horizontal scales: mid-troposphere lows are parts of planetary Rossby waves with longer wavelengths, while UTT features are shorter waves that break from the long waves (RWBs), or tropical disturbances with average wavelengths around 3000km (TUTTs) (Chen and Chou, 1994; Kelley and Mock, 1982). In addition to its difference with mid-troposphere lows, UTTs exhibit distinct movements across subregions. For every UTT-related EPEs, the movement direction is determined by UTT's initial and ending longitudes. Table 4.2 lists the number of westward and eastward moving UTTs for each subregion. The westward movement agrees with Igel et al. (2021) that most TUTTs are advected by the background easterlies. As for RWBs, they are often attached to mid-latitude high PV and propagate eastward (Zavadoff and Kirtman, 2019). Take Sub6 as an example, Figure 4.16 shows the composites of PV200 and U200 for eastward-UTTs and westward-UTTs. PV200 shows positive anomalies that span to extratropical regions for eastward-moving UTTs, and the high PV200 disturbances are located in the westerlies. These behaviors align with the RWB features in Zavadoff and Kirtman (2019). As contrast, the positive PV200 anomalies are relatively smaller in sizes for westward-moving UTTs, and they are located in tropical easterlies. This follows Igel et al. (2021) that most TUTTs are advected by the background easterlies. Moreover, the boundary of westerly and easterly moves further north during westward-UTT events. This transition favors the TUTT advections from tropics to the NAM region. Thus, although we use UTT as a category for all upper-level disturbances, they can be classified into tropical and subtropical features based on their locations and propagation directions.

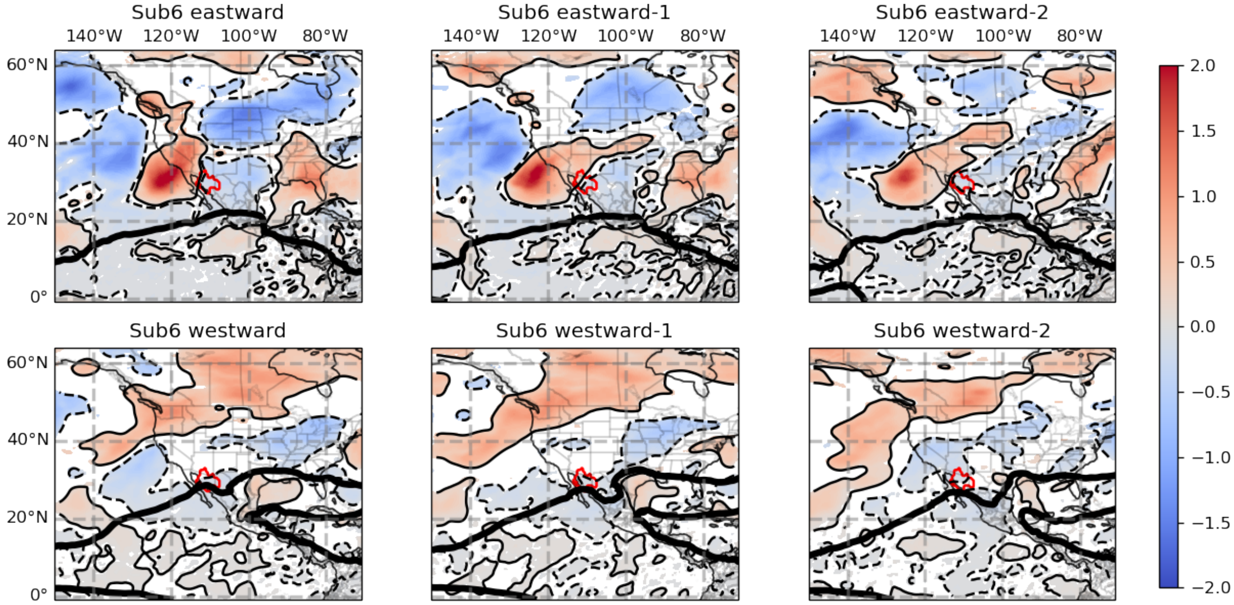

Figure 4.16: PV200 standardized anomalies and U200 contours for eastward and westward UTTs in Sub6. Red shape denotes the location of Sub6. Solid black contour represents the boundary line of 200 hPa easterlies and westerlies. Shading depicts PV200 standardized anomalies with 90% confidence interval. The left column is for concurrent days. Middle and right columns are for one day prior and two days prior the onset date, respectively.

Table 4.2: Number of UTT events by propagation direction in each sub-region.

	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7
UTT events	10	15	12	33	51	46	53
Westward	6	9	2	8	21	36	35
Eastward	4	6	10	25	30	10	18

Fronts and mid-tropospheric lows

Fronts and mid-tropospheric lows are more pronouced in in-land subregions (Sub4 and Sub5). They are connected since fronts tend to occur near a mid-tropospheric low where temperature is anomalous low. Further examination shows mid-tropospheric lows are always associated with fronts for EPEs after 2003, as shown in Figure 4.14. It probably indicates our tracking of mid-level lows is an approximation of frontal systems. Although both fronts and Mid-lows can introduce uplift motions, their composites show difference in magnitude and spatial extent. Figure 4.17 depicts the composites of front EPEs with and without mid-tropospheric lows in Sub4 as an example. Larger anomaly magnitudes are observed for fronts with mid-tropospheric lows. This is related to our geopotential magnitude criterion for mid-tropospheric lows. With -1000 m^2/s^2 as the threshold, an anomalous low temperature is expected. This cold air tends to facilitate cold front formations. In Sub4, for instance, among the 26 front-EPE days without mid-tropospheric lows, stationary fronts predominate with 20 days, while there are only 6 days with cold fronts. To the contrary, there are 8 cold front days and 5 stationary fronts in the 13 front-EPE days associated with mid-tropospheric lows. Similar patterns indicating mid-tropospheric lows would enhance the cold front fraction have also been found in Sub5 and Sub6 where fronts are common.

The unclassified EPE of 2003

In addition to the composites for each EPE category, meteorological conditions for the unclassified event after 2003 is examined. Local high water content is shown in TCWV and Q950 fields. PV200 and CAPE indeed show positive anomalies near the precipitation area, and the EPEs are likely related to these disturbances given that there are no clear disturbances found in IVT, SLP and Z500 fields. However, the upper-level disturbance is below 2PVU, which leads to missed UTT features based on our tracking criteria. Associated with relatively weak upper-level anomalies, the precipitation rate of this unclassified events (10.87mm/day for Sub6) is close to the 95th percentile thresholds (10.65 mm/day for Sub6).

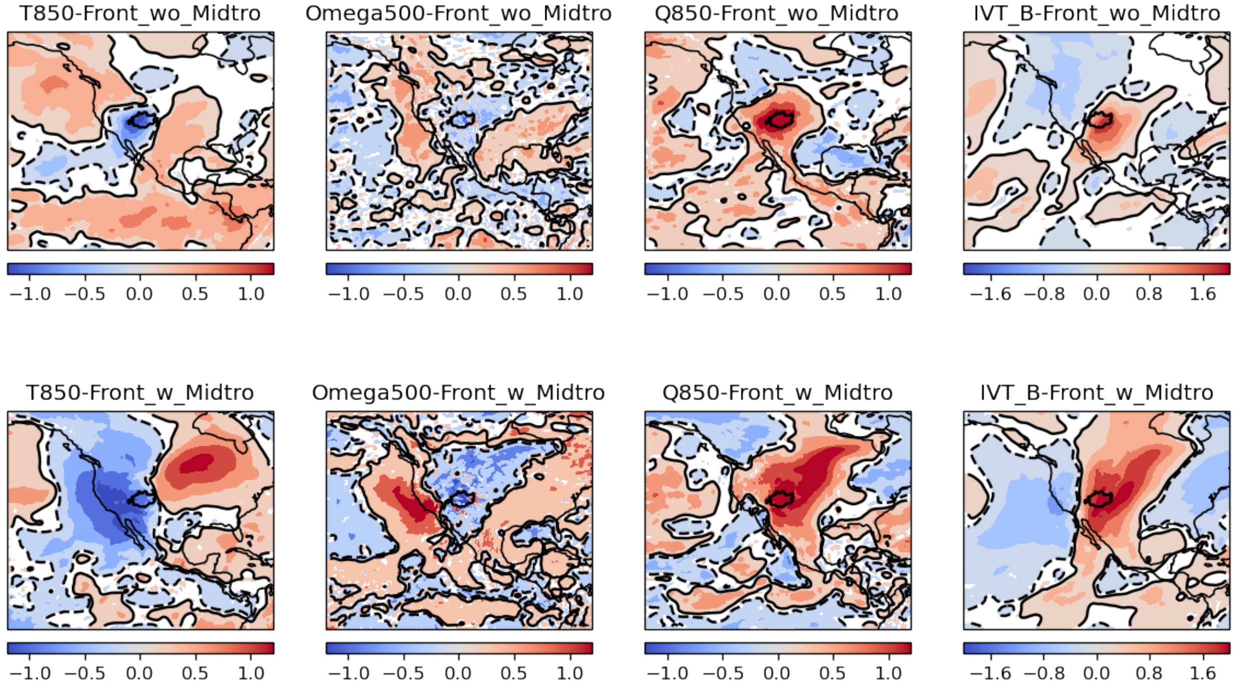

Figure 4.17: Front EPE composites in Sub4. Upper row shows fronts without mid-trpospheric lows and bottow for fronts with mid-tropospheric lows. Black contours show the 90% confidence interval.

4.5.5 Precipitation Distributions Associated with Atmospheric Features

Although the NAM EPEs can be assigned to various atmospheric drivers, the presence of a particular atmospheric driver is not a sufficient condition for occurrence of an EPE. To assess precipitation response in the presence of a particular atmospheric feature, we composite the precipitation rate with respect to different drivers and compare the probability of EPEs. Following the definition of rainy days, only those precipitation rates larger than 1 mm/day are analyzed. As the precipitation rate generally follows a gamma distribution (Martinez-Villalobos and Neelin, 2019; Watterson and Dix, 2003), for precipitation rates larger than 1 mm/day, a truncated Gamma distribution (TGD) is applied with cutoff threshold as 1 mm/day. The generalized Pareto distribution (GPD) is also employed here since it is commonly employed for assessing the tail of various distributions (Dargahi-Noubary, 1989). Both the GD and GPD have three parameters: shape, location, and scale. The TGD probability density function (PDF) is derived from the GD PDF normalized with its cumulative distribution function (CDF). When fitting the data, the location parameter is fixed to 0 mm/day for the GD and 1 mm/day for GPD. Shape and scale are optimized using their maximum likelihood estimate.

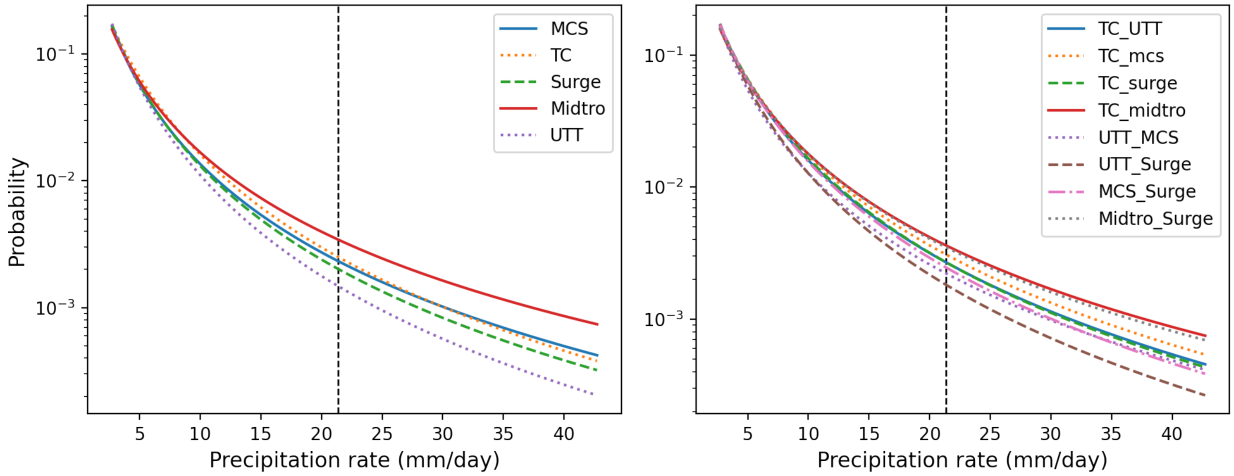

Figure 4.18: Precipitation rate distribution with respect to atmospheric drivers for Sub1. The dashed vertical line denotes the 95th percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

Figure 4.18 and Figure 4.19 shows the precipitation rate PDF function with single and double

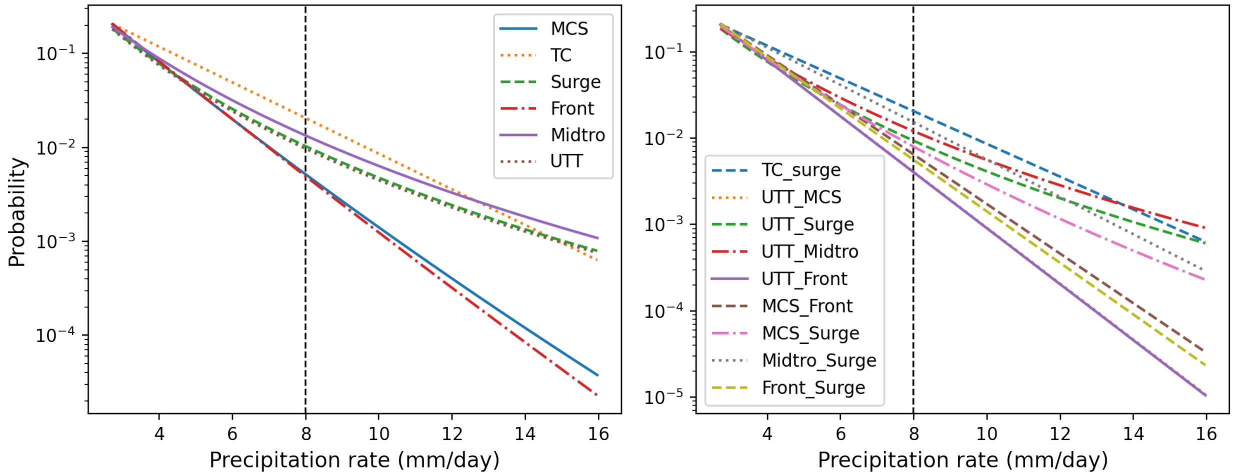

Figure 4.19: Precipitation rate distribution with respect to atmospheric drivers for Sub2. The dashed vertical line denotes the 95th percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

Table 4.3: Probabilities of extreme precipitation rates associated with candidate drivers. Probability values are shown in percentages (i.e., $\times 10^{-2}$). Values in parenthesis after subregion index are the 95^{th} percentile of precipitation rates in mm/day. Values in parenthesis after the probabilities are number of samples. Highest probabilities for single driver and two drivers are shown in bold. Only the drivers with sample size larger than 10 are compared.

	Sub1 (21.36)	Sub2 (7.98)	Sub3 (7.63)	Sub4 (6.30)	Sub5 (6.47)	Sub6 (10.65)	Sub7 (17.25)
TC	3.0 (227)	**4.7 (23)**	(1)	(2)	**6.7 (29)**	**4.5 (89)**	**3.8 (178)**
UTT	1.7 (255)	2.9 (336)	3.2 (335)	2.2 (786)	1.9 (1315)	0.9 (1114)	0.3 (1175)
Surge	2.6 (529)	3.1 (496)	4.0 (465)	3.1 (814)	2.4 (1287)	1.0 (1519)	0.7 (1050)
MCS	3.3 (135)	0.8 (99)	**4.6 (22)**	3.4 (98)	2.6 (481)	0.7 (623)	0.5 (959)
Front	5.1 (9)	0.7 (117)	2.3 (200)	4.0 (548)	2.3 (799)	2.2 (379)	0.4 (262)
Midtro	**5.9 (27)**	4.1 (42)	2.6 (122)	**5.0 (234)**	5.7 (196)	3.8 (62)	2.7 (24)
TC-UTT	3.6 (56)	(3)	(0)	(0)	4.9 (10)	4.9 (21)	5.2 (50)
TC-Surge	3.4 (208)	**4.7 (23)**	(1)	(2)	**7.4 (28)**	6.1 (84)	3.9 (160)
TC-MCS	4.2 (42)	(3)	(0)	(0)	(3)	10.3 (15)	2.2 (50)
TC-Front	(2)	2.0 (7)	(0)	(0)	5.7 (15)	**25.2 (15)**	**6.9 (14)**
TC-Midtro	**5.9 (24)**	11.3 (5)	(0)	(1)	17.4 (8)	1.4 (10)	7.1 (9)
UTT-Surge	2.1 (171)	2.6 (204)	3.2 (201)	2.0 (394)	2.2 (705)	1.0 (697)	0.6 (400)
UTT-MCS	3.3 (39)	0.5 (46)	**6.9 (14)**	2.6 (56)	2.6 (279)	0.7 (307)	0.4 (398)
UTT-Front	(3)	0.5 (45)	3.4 (80)	2.5 (287)	2.3 (423)	2.5 (172)	0.1 (95)
UTT-Midtro	4.0 (7)	3.6 (17)	2.4 (60)	3.0 (105)	4.0 (84)	5.6 (22)	1.9 (7)
MCS-Surge	3.1 (102)	1.7 (65)	3.4 (14)	4.8 (48)	2.9 (270)	0.9 (393)	1.0 (351)
MCS-Front	4.3 (4)	1.0 (38)	1.6 (12)	3.5 (83)	2.7 (391)	0.6 (223)	0.7 (194)
MCS-Midtro	(2)	(1)	2.5 (5)	**8.9 (15)**	5.6 (21)	2.8 (6)	(3)
Midtro-Front	(0)	2.5 (6)	5.0 (27)	7.9 (70)	4.7 (43)	2.7 (11)	(1)
Midtro-Surge	5.4 (25)	3.1 (27)	2.4 (61)	6.7 (87)	6.8 (87)	4.0 (38)	3.5 (18)
Front_Surge	2.6 (6)	0.8 (77)	3.0 (127)	5.2 (287)	2.6 (437)	2.8 (226)	0.7 (89)

atmospheric drivers in Sub1 and Sub2. Figures for other subregions are available in the supplements (S44 to S48). Overall the decaying tendency of precipitation rate does not change when composited on various categories. The PDF curve changes among subregions for different drivers, and the driver associated with the highest precipitation rate varies as well (i.e., MCSs are more likely to bring heavy precipitation in Sub1 while its precipitation probability is relatively lower in Sub2). The PDF plots illustrate the distribution of precipitation rates, but they don't show which driver is more likely to generate extreme precipitation, because EPE thresholds vary per subregion. In addition, the probability of the EPE threshold being met may be lower for some drivers, though it is higher for precipitation rates greater than that threshold, such as the MCS curve in Sub3. To compare the probability of EPE occurrence more intuitively, the area under the PDFs above the regional 95^{th} percentile of precipitation (given by $1 - \mathrm{CDF}(95^{\mathrm{th}}\mathrm{percentile})$), are calculated and listed in Table 4.3. The upper and lower panels show single and double drivers, respectively. Here we mainly focus on the drivers with more than ten samples, as the distribution parameters estimated from too few samples may be unreliable. The single driver with the highest extreme precipitation probability differs among the subregions, and it is not the most dominant driver shown in Figure 4.11. Especially for Sub5, TCs are the driver with the highest probability of extreme precipitation rates, whereas both the number and precipitation amount of TC-related EPEs are the lowest as shown in Figure 4.10 and 4.11. This result actually reflects the inland location far from the coastlines. Coastal areas like Sub1, Sub2, Sub6 and Sub7 with more TCs typically have higher thresholds for extreme precipitation rates. In this case, the strength of the TC and its distance from the area will most likely determine if it would trigger extreme precipitation. Sub5, on the other hand, is an interior desert region where TCs are relatively rare. Its overall precipitation rates and 95^{th} percentiles are also lower. Once a TC makes landfall nearby, it would have a higher chance to bringing with it extreme precipitation. As a result, a higher EPE category frequency does not necessarily indicate a higher chance of extreme precipitation rate associated with the driver.

Compared with single drivers, the probability for an EPE to occur when two drivers are present is not necessarily higher due to the addition of another driver (i.e., the cumulative probability of UTT-Surge in Sub1 is less than UTT), implying that the multi-driver interactions are not additive.

When the second driver is included, the extreme precipitation probability may increase, decrease or remain unchanged, depending on the subregion and associated drivers. In the remainder of this section we investigate some more interesting combinations of features.

TC-Surge interactions

Given their close association, TC and TC-Surge PDF curves are similar in Sub1 and Sub7 (figures in the supplement). In addition, the number of TC-Surge-related precipitation days is about equal to the number of TC-related days, suggesting that TCs are generally associated with GOC surges. Given the similar locations towards the south end of GOC, the behaviors of TCs and TC-Surges are nearly identical in Sub1 and Sub7.

TC-UTT interactions

The PDF curves for TC, UTT and TC-UTT precipitation are further compared in Sub7 since TC and UTTs are both frequent here. The TC-UTT precipitation curve is close to the TC curve, while the UTT curve is far below these two curves in the high precipitation rate regimes. This result is likely due to the larger distance criterion for UTTs. TCs are more frequent to the west of Sub7 while easterly UTTs are prevailing as shown in Table 4.2. Since, in a compound event, UTT centers are usually far from the TC centers, the TCs are unlikely to be affected by UTTs. When we decrease the distance criterion to 5 degrees, however, the TC-UTT curve indeed shows lower probabilities for high precipitation rates. A further examination of the composites shows UTTs hinder the eastward moisture transportation by TCs, which decreases the local water content in Sub7. This is in accord with previous research showing that UTTs can decrease the TC activities (Wang et al., 2020b; Zhang et al., 2016, 2017).

Mid-troposphere lows and fronts

Mid-troposphere lows and fronts are selected as major drivers of EPEs for Sub4 and Sub5 areas, since they are frequent in these inland areas. Interactions between these features are different between Sub1 and Sub5. The combination of fronts and mid-troposphere lows yield mediocre precipitation probabilities with front and mid-troposphere lows as lower and upper boundaries in Sub5. To the contrary, Front-Midtro shows higher probabilities for extreme precipitation rates with lower probabilities for both fronts and mid-troposphere lows. Despite the discrepancy, the composites reveal that front would increase the local water content while mid-troposphere lows are associated with ascending motions. Thus, the distinct interactions are probably due to the trade-offs between the water vapor and lifting motions, which may be caused by the different orientations of fronts and mid-troposphere lows in these two regions.

4.6 Relating Features with LOD Modes

Although the LOD modes focus on mean precipitation whereas synoptic and mesoscale features are utilized to understand extreme precipitation, they can be linked together by the average standardized anomalies of LOD modes under the condition of EPE features. The time series of the LOD modes, representing the relative strength of each LOD modes, can thus reveal which large-scale meteorological conditions are most strongly connected with the selected features. The first two LOD modes from the 3-day mean precipitation for each subregion are selected here, because not enough robust LOD modes are obtained using 1-day precipitation, and 5-day mean precipitation is too smooth. The standardized anomaly of the LOD coefficients are calculated as the anomaly (actual values minus the monsoon season mean) divided by its standard deviation. When composited on different features, larger standardized anomalies suggest higher co-occurrence probabilities of the LOD mode and the selected feature, and vice versa. Table 4.4 lists the composite standardized anomalies of Mode1 and Mode2 with respect to candidate features for all the subregions.

As we discussed in section 4.4.2, the first modes are viewed as the long-term seasonal background for precipitation while the second modes represent short-term phenomena with sharper spatial gradients such as weather systems. The standardized anomalies of Mode2 for TCs are substantially larger than Mode1 for most subregions, indicating TCs' shorter life cycles, with the exceptions of Sub1 and Sub5. The magnitudes of Mode1 and Mode2 for TCs are close in Sub5, and TCs are rather rare here because of its inland location. The relatively lower standardized anomaly for Mode2 is probably due to its small sample size. In Sub1, Mode1 shows considerable larger

Table 4.4: Composite LOD mode anomalies for selected features. Percentage in parenthesis denotes the explained variance of precipitation by each mode, which is equivalent to the R^2 scores in Table 4.1. The left panel shows mean standardized LOD coefficients associated with all precipitation days, while only EPEs are listed in the right panel.

	All Precipitation Events						Extreme Precipitation Events					
	TC	UTT	Surge	Midtro-Low	Front	MCS	TC	UTT	Surge	Midtro-Low	Front	MCS
Sub1												
Mode1 (35.0%)	2.097	0.624	1.187	3.001	1.344	1.323	4.142	3.571	3.953	4.764	3.052	4.883
Mode2 (2.6%)	0.387	0.330	0.446	-0.149	0.452	0.650	0.326	0.659	0.503	-0.957	2.013	1.311
Sub2												
Mode1 (40.0%)	1.195	1.290	1.348	0.387	1.242	1.316	1.462	1.940	1.907	0.953	2.275	1.955
Mode2 (5.0%)	3.816	0.419	0.828	1.551	0.799	0.632	2.735	0.957	1.989	1.758	2.005	2.734
Sub3												
Mode1 (42.0%)	1.752	1.308	1.535	0.569	1.347	1.028	1.752	1.916	2.043	1.394	1.795	NA
Mode2 (4.0%)	7.929	0.526	0.736	1.466	0.575	1.004	7.929	1.676	2.321	2.901	2.654	NA
Sub4												
Mode1 (52.8%)	1.242	1.028	1.073	0.456	1.012	1.064	1.218	1.683	1.610	1.102	1.700	1.602
Mode2 (10.1%)	3.094	0.457	0.478	1.629	0.412	0.504	4.647	1.832	2.289	2.692	2.316	1.992
Sub5												
Mode1 (57.6%)	0.653	0.747	0.764	0.456	0.777	0.819	1.387	1.284	1.228	1.134	1.490	1.522
Mode2 (5.8%)	0.594	0.284	0.195	1.297	0.169	0.135	2.189	1.447	1.347	1.929	1.240	1.035
Sub6												
Mode1 (44.5%)	0.774	0.694	0.790	-0.106	0.612	0.750	1.010	0.921	0.964	0.456	1.221	1.190
Mode2 (9.6%)	2.840	0.144	0.339	1.823	0.262	0.182	3.466	1.431	1.739	1.971	2.935	1.506
Sub7												
Mode1 (32.1%)	0.883	0.440	0.685	0.497	0.570	0.570	1.028	0.821	0.997	0.695	0.903	1.064
Mode2 (10.5%)	1.914	0.124	0.466	1.967	0.126	0.144	2.449	0.881	1.495	2.359	1.863	1.005

standardized anomaly for TCs. As seen in Figure S2, Mode1 in Sub1 is related with the local lows in SLP and Z500, which correspond to TC characteristics. This result again underscores the critical roles of TCs in Sub1. Despite the limited life span for a single TC, prevalent TCs can provide a seasonal background for precipitation here. The dominance of TCs in Sub1 also justifies the late monsoon onset date here. As we used 1mm/day and 5 days as thresholds for monsoon onset dates, although TCs bring abundant precipitation, they may not last for 5 consecutive days given their short lifetime and result in a postponed onset date.

In all subregions, Mode1 anomalies associated with GOC surges are always larger than Mode2, indicative of their dominant role in affecting NAM precipitation. Midtropospheric lows, on the other hand, occur more often in conjunction with the second mode (except for Sub1 where they are less frequent). This suggests that midtroposphere lows have stronger spatial gradients than the monsoon ridge, and it agrees with the LOD results that local Z500 lows tend to occur with the second mode while monsoon ridges are present in the first mode. In the second modes for Sub2 through Sub7, there are positive PV200 coefficients, which are similar to UTTs that are detected as positive PV200 contours. Nonetheless, Mode1 obtains higher anomalies associated with UTTs than Mode2. It is probably due to a potential mismatch between the UTT centers and PV200 highs, as can be seen from Figure 4.17 that UTTs can occur in the tropical easterlies while Mode2 shows positive PV200 in mid-latitudes (Figure S9-S14).

As for EPEs, the anomalies are higher than mean precipitation as expected. In addition, since EPEs have shorter time periods and stronger precipitation rates, they are generally more correlated with the second mode (deeper gradients) except for Sub1. This result again agrees with our previous analysis that the second modes are more representative for short-term phenomena. The exception for Sub1 is probably due to the fact of prevailing TC activities, which are responsible for both mean precipitation and EPEs. There are indeed several exceptions that Mode1 has a larger anomaly than Mode2 for EPEs (i.e., UTTs in Sub2 and Sub3, and MCS in Sub5 and Sub7). Since MCSs can not be fully resolved in current reanalysis datasets, it is hard to analyze their impacts with ERA5 dataset. UTTs in Sub2 and Sub3 are more correlated with Mode1 for mean precipitation, and this relationship doesn't change when composited on EPEs. It indicates that EPEs share the same precipitation mechanisms with mean precipitation. The occurrence of EPEs are largely due to the increase of disturbances represented by Mode1. On the other hand, fronts are more connected with Mode2 for EPEs in Sub3 and Sub4, whereas Mode1 anomalies for mean precipitation associated with fronts are larger. This LOD mode transition suggests different precipitation processes for mean precipitation and EPEs. EPEs are brought on by both the increase in Mode1 and the shift to Mode2.

4.7 Conclusions

This work investigates the meteorology drivers for both mean precipitation and EPEs in the NAM region from 1979 to 2018. We first determine the NAM domain and its subregions from the CPC precipitation dataset, rather than using individual states or lat-lon based areas. Since the SOM-based identification method emphasizes the extreme precipitation characteristics and doesn't rely on topographical features or state borders, it is better suitable to regional precipitation studies. Robust linear modes associated with mean precipitation are derived using the cross-validated linear decomposition method. Compared with other clustering and composite-based methods, the linear models can be used to assess the precipitation predictability with their R^2 scores. In general, the predictability is higher for in-land subregions, suggesting more complex mechanisms for precipitation in coastal areas. Furthermore, the spatial scales of LOD modes can be reflected in their hierarchy order. The first modes tend to have lower spatial gradients. The local high water contents are selected in the first modes for most of the subregions, along with the monsoon ridge and GOC moisture transport shown as the positive coefficients in Φ500 and IVT fields. The negative coefficients of Φ500 and SLP imply TCs have a greater influence on Sub1. The second modes correspond to weather patterns, which are represented by closed contours in Φ500 and PV200 fields, and these systems generally have sharper gradients than the first modes. The extracted modes tend to have higher spatial gradients with more iterations, making them more sensitive to specific precipitation events. The cross-validation prevents overfitting by keeping only the generalizable modes from the iterations.

Candidate meteorological features to link with EPEs include TC, UTT, GOC moisture surge, front, mid-troposphere low and MCS. Almost all the EPEs fall into at least one of these categories. For the unclassified EPEs after 2003, the PV200 anomalies are quite weak, and its precipitation rate is close to the EPE threshold as consequence. This suggests a potential quantitative link between precipitation and meteorology conditions. Different subregions have various dominant drivers, and most EPEs are associated with more than one driver. Given the larger EPE precipitation fraction associated solely to them, GOC surge, MCS and front are substantially important. This finding highlights the importance of developing MCS and front datasets for the NAM region prior to 2003. The attribution of EPEs does not indicate these drivers are sufficient conditions. Further investigation reveals that the probabilities of EPEs given the presence of these drivers are less than 0.1. Additionally, the driver with the highest extreme precipitation probability for each subregion is not the most dominant driver shown in EPE associations. This is the indication of the discrepancy between the frequency of EPE drivers and the probability of EPEs under the condition of drivers.

EPE composites are constructed to reveal the general conditions for extreme precipitation.

High local water vapor content (Q850, TCWV) and upward lifting (Ω500, CAPE) are significant, which create a favorable environment for precipitation. GOC moisture transportation is shown with positive IVT-A anomalies for most subregions, while the negative IVT-A anomalies in Sub4 suggests its long distance away from the GOC and relatively longer time-window for GOC surges. Composites for EPEs with different drivers are examined further. Specifically, for UTT-EPEs, the directions of the upper-level disturbances are analyzed. There are more westerly disturbances for northern subregions (Sub3) whereas easterlies are more common for southern subregions (Sub6 and Sub7). The differences in propagation directions can be observed in the PV200 composites. PV200 highs are in the westerlies for eastward UTT cases and vice versa. Thus, although we use UTT to represent all upper level disturbances, the propagation differences could potentially separate them into easterly tropical features (i.e., TUTTs, ITs) and westerly extratropical features (i.e., RWB, PV streamers).

The LOD modes for mean precipitation are linked with EPE-related features by the standard anomalies for each mode. Anomaly magnitude can reflect the mode occurrence along with the features. Larger anomalies represent a closer link between the feature and the mode. Most features are more related with the first mode for mean precipitation with TCs and Midtro-Lows as two exceptions. Midtro-Lows yield various results across the subregions. TCs are more correlated with Mode2 for most subregions except for Sub1 where TCs are prevailing. Comparing the mean precipitation with EPEs, most features are more correlated with Mode2, which agrees with our LOD analysis that the second mode is more related with shorter-time phenomena. Additionally, the same prevalent mode for mean precipitation and EPEs suggests a similar precipitation mechanism and the EPEs are caused by the strengthened disturbances, while the transition of modes from mean precipitation to EPEs suggests different precipitation drivers for EPEs.

We are primarily interested in the co-occurrence of atmospheric drivers with EPEs, which does not indicate causality. In terms of future research, a causal inference analysis could be conducted. As for the linear decomposition method, all the modes are extracted based on the linear correlation between meteorological variables and precipitation. With the rise of machine learning and deep learning applications, it is possible to use a nonlinear machine learning model to replace the linear regression model (Wehner et al., 2021). It can incorporate any possible nonlinear relationships, and the difference in the R^2 score compared with linear models can be used to quantify the nonlinearity. Finally, give the modest PV200 anomalies for unclassified EPEs with lower precipitation rates, we see chances to incorporate quantitative analysis between atmospheric drivers and precipitation rates.

4.8 Supplements

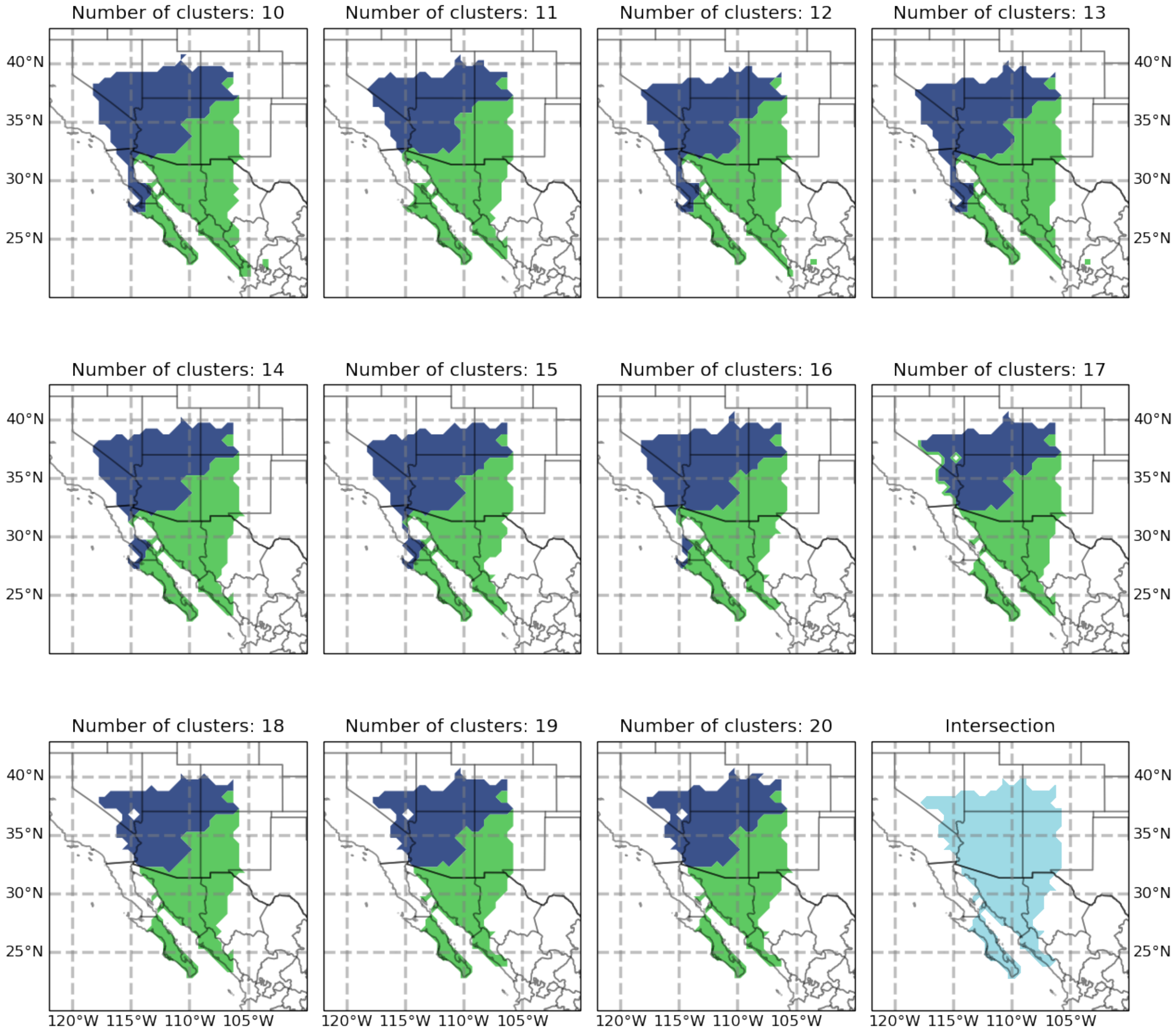

Figure S1: Ensemble SOM cluster results for the NAM domain.

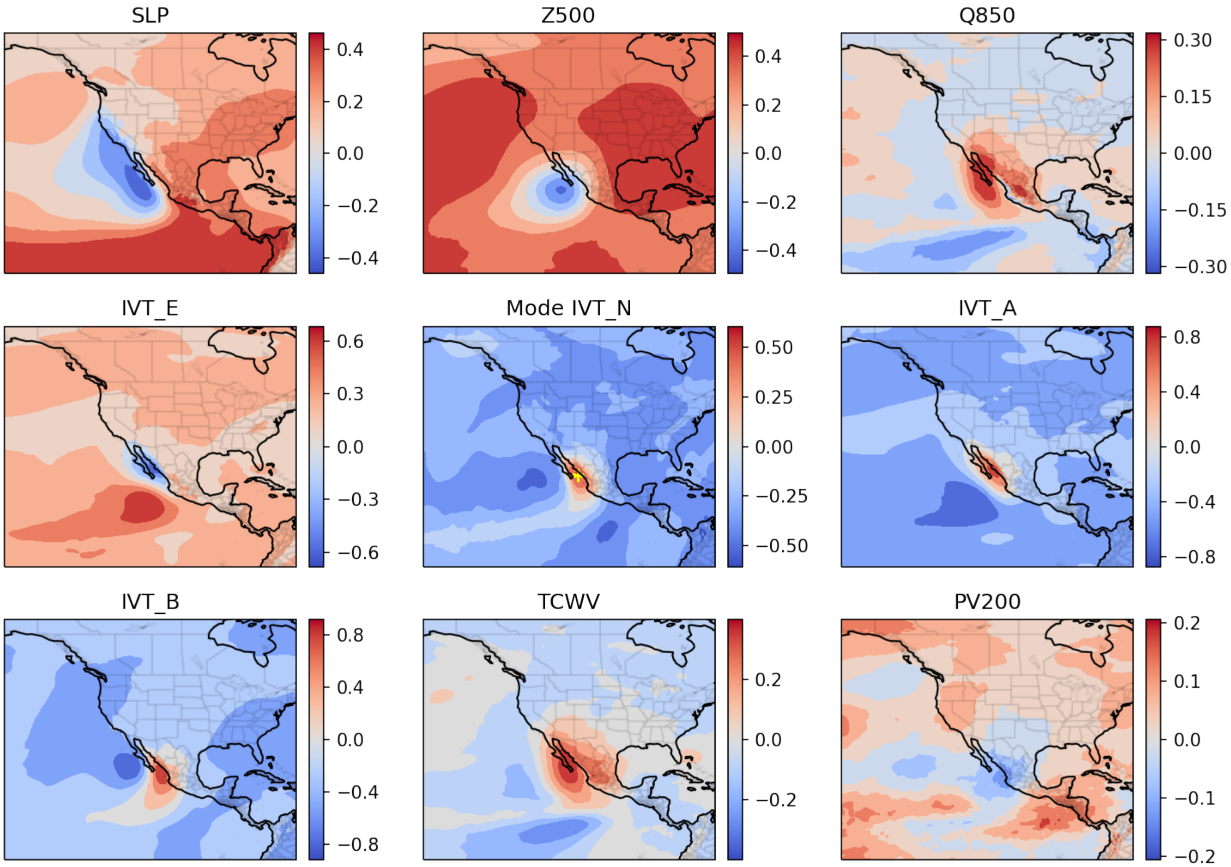

Figure S2: The first mode for 3-day averaged precipitation of subregion 1.

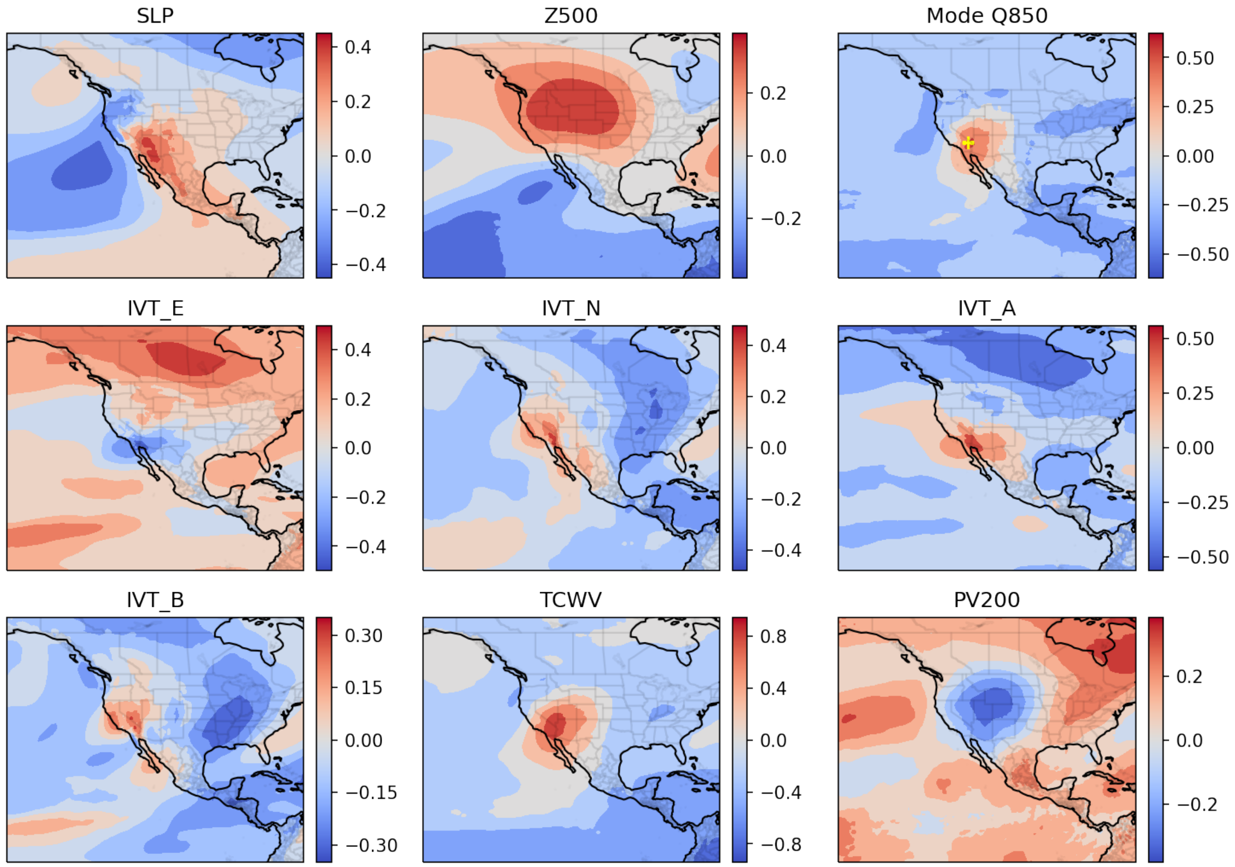

Figure S3: The first mode for 3-day averaged precipitation of subregion 2.

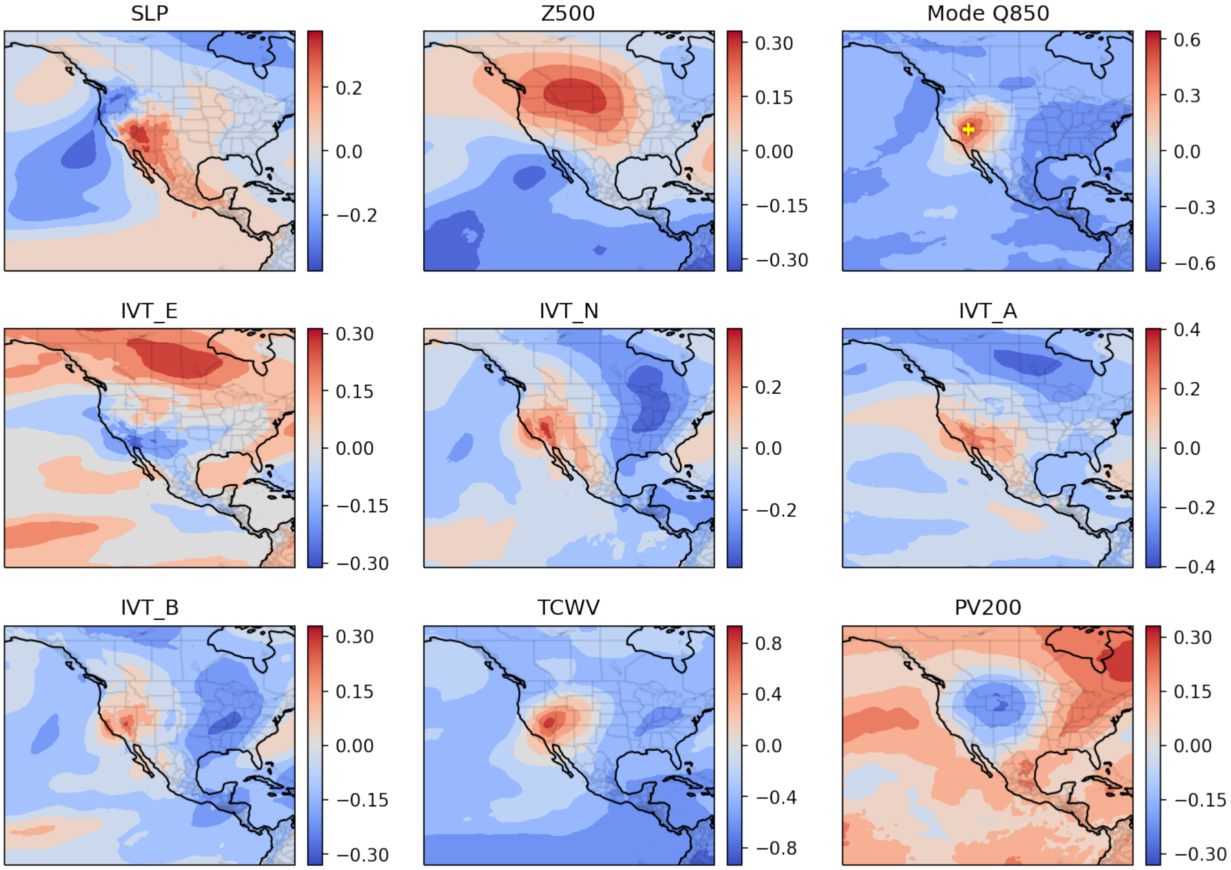

Figure S4: The first mode for 3-day averaged precipitation of subregion 3.

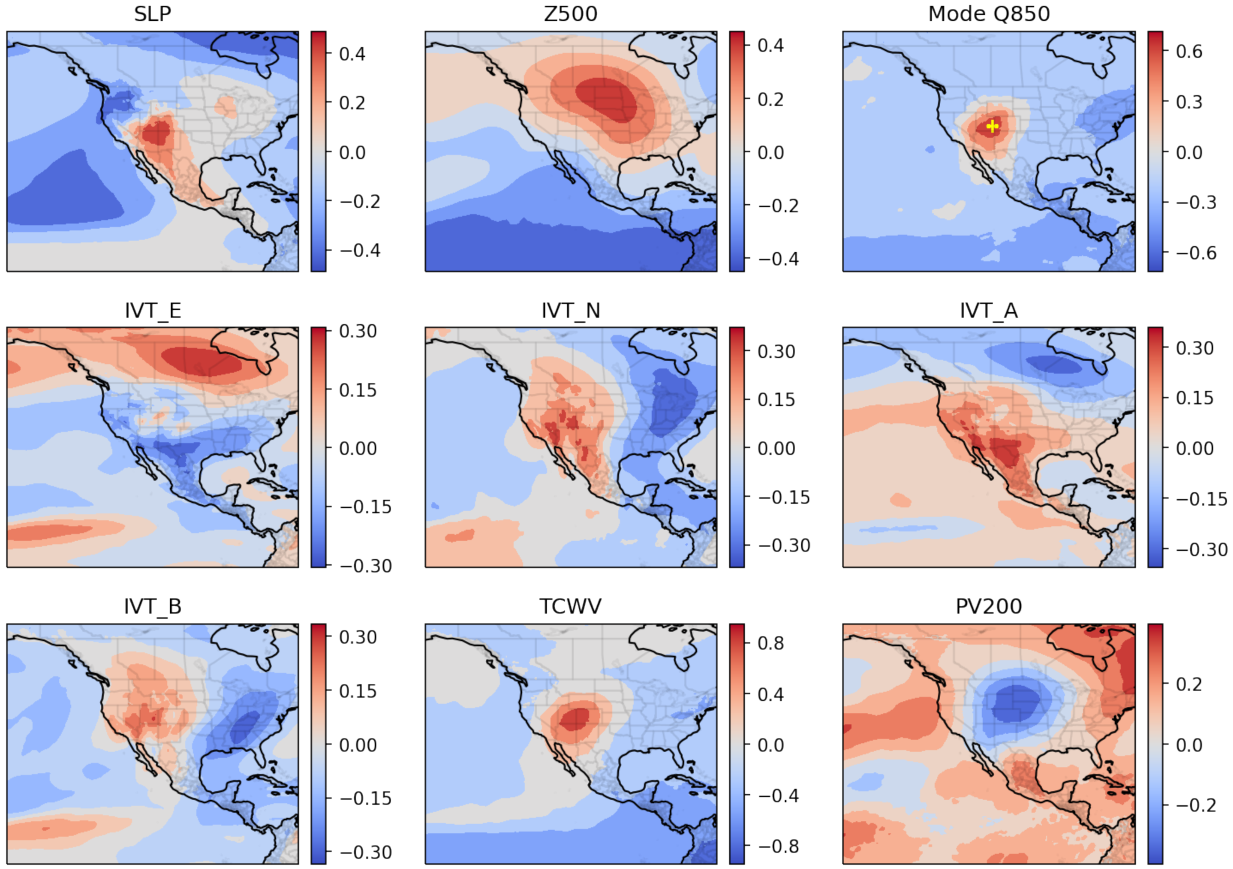

Figure S5: The first mode for 3-day averaged precipitation of subregion 4.

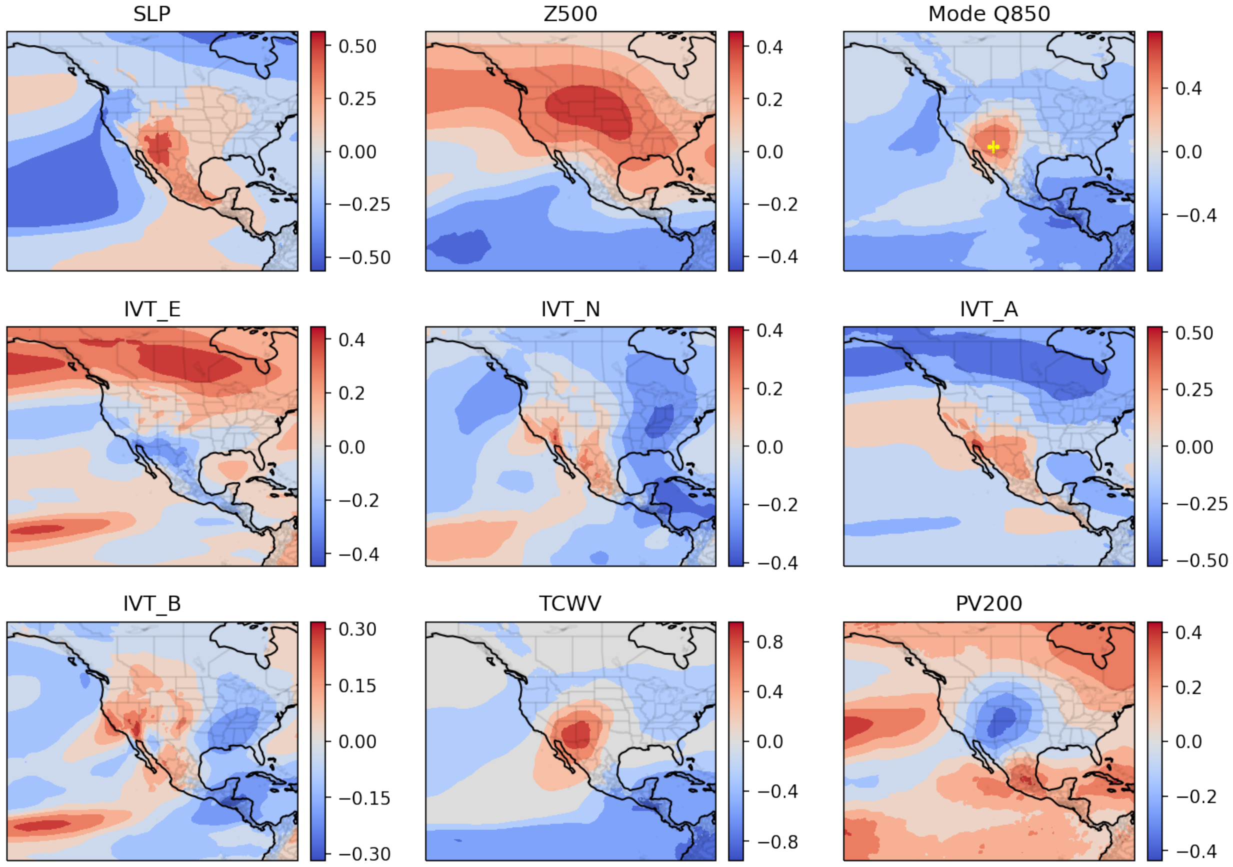

Figure S6: The first mode for 3-day averaged precipitation of subregion 5.

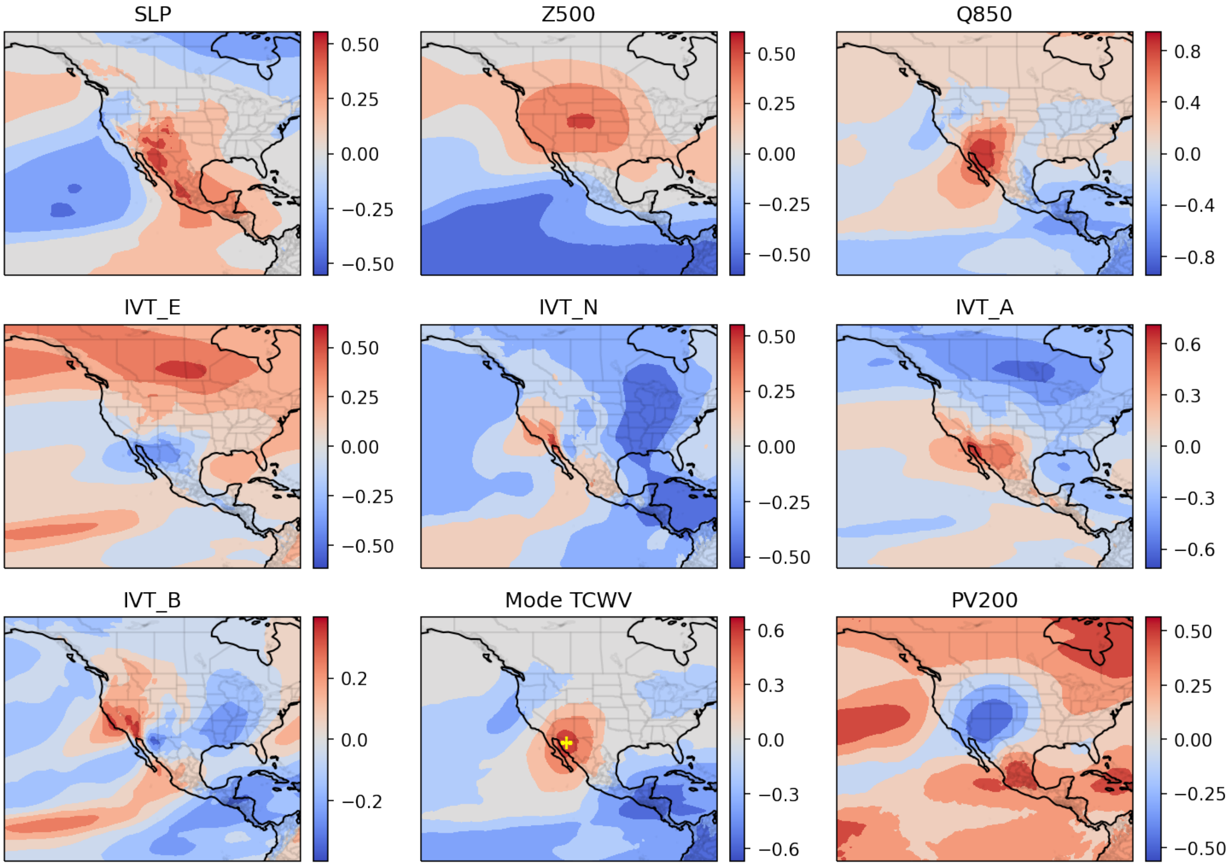

Figure S7: The first mode for 3-day averaged precipitation of subregion 6.

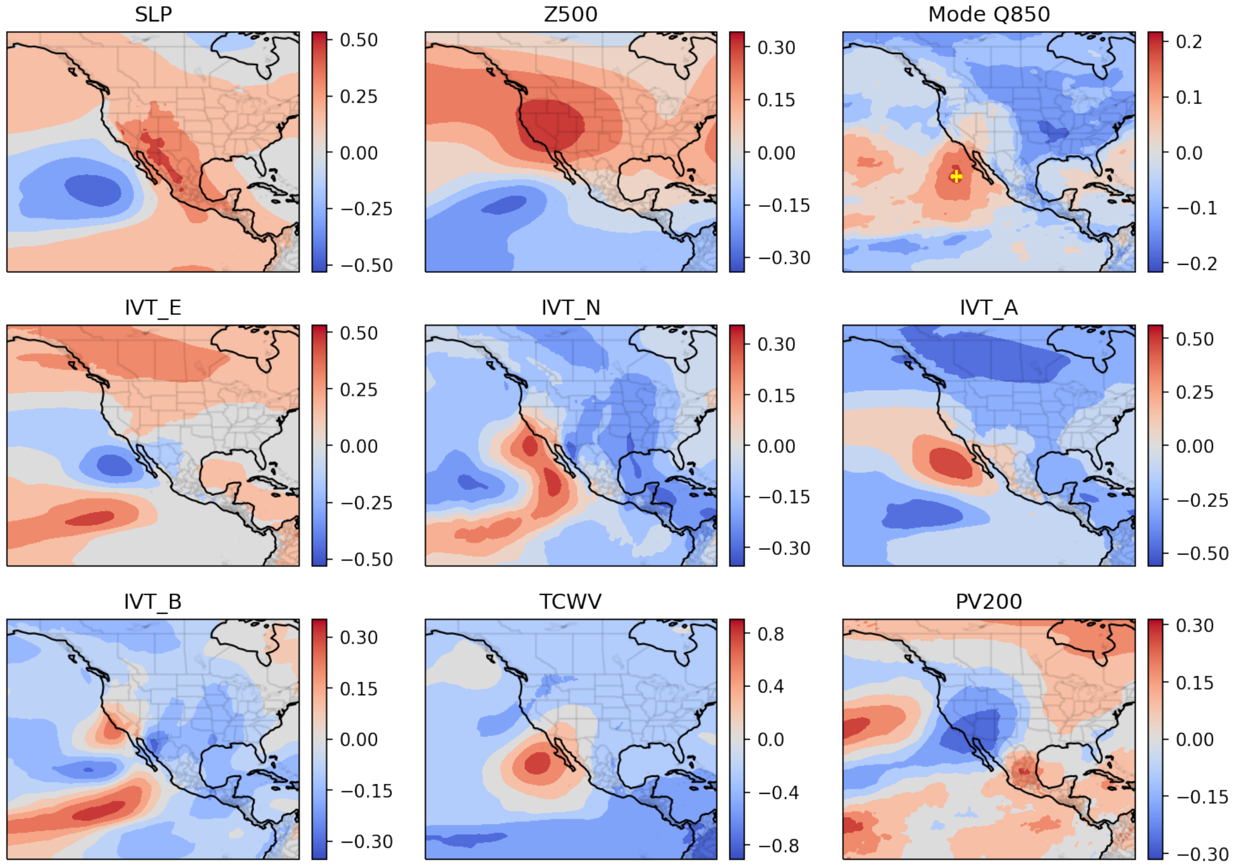

Figure S8: The second mode for 3-day averaged precipitation of subregion 1.

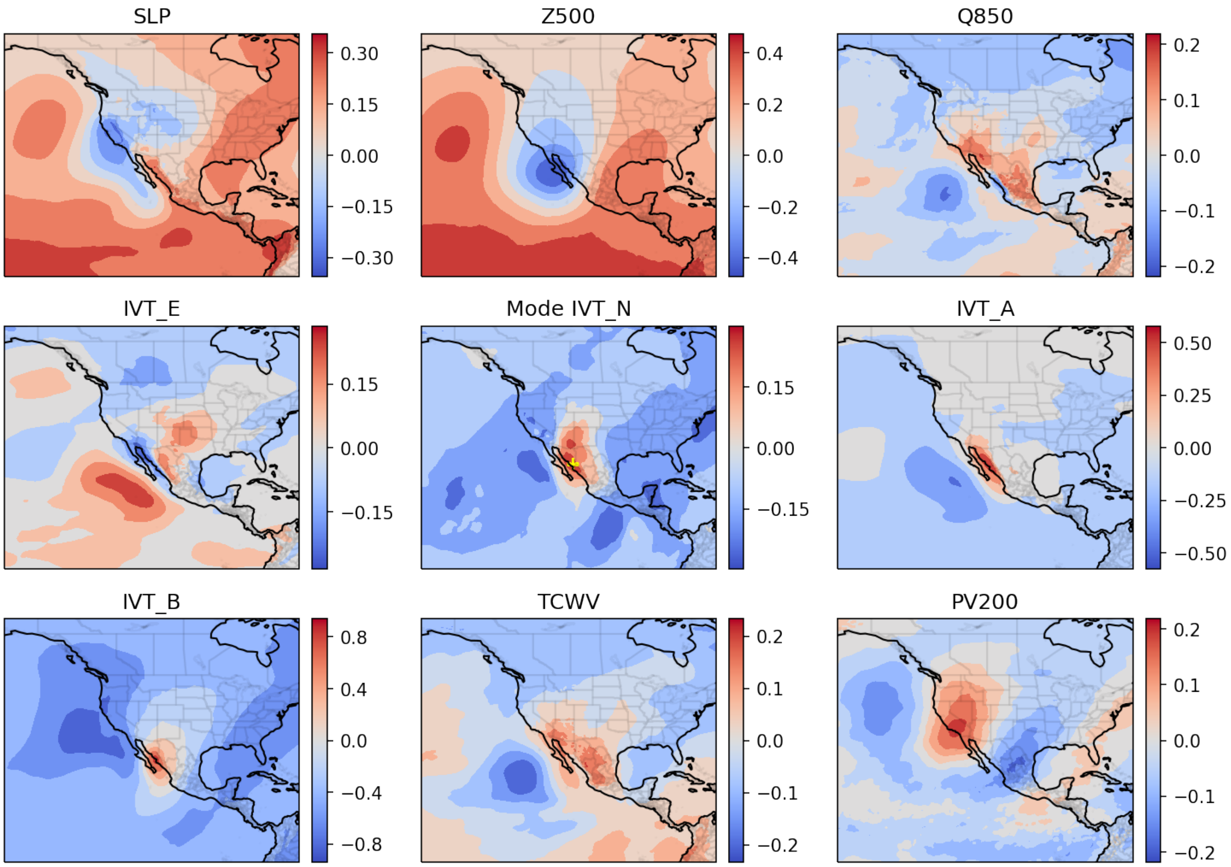

Figure S9: The second mode for 3-day averaged precipitation of subregion 2.

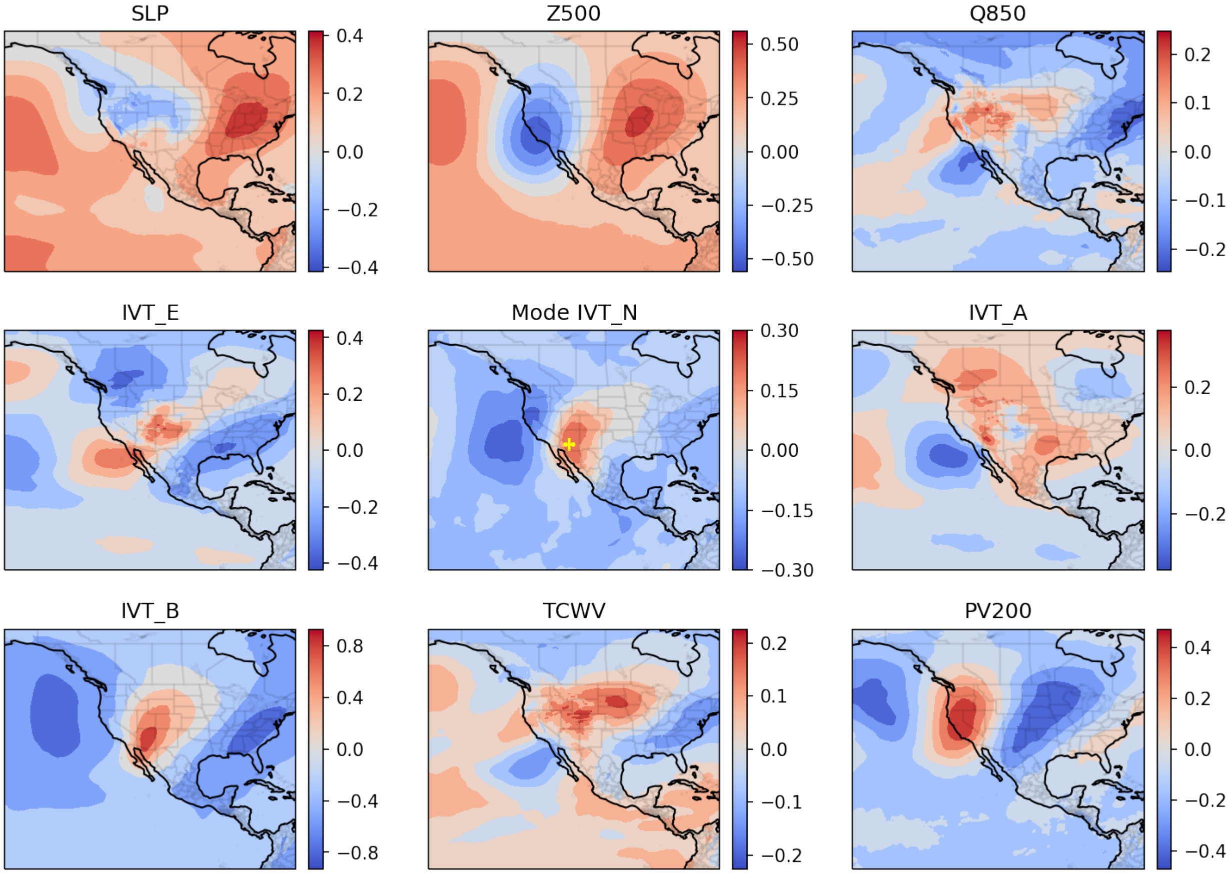

Figure S10: The second mode for 3-day averaged precipitation of subregion 3.

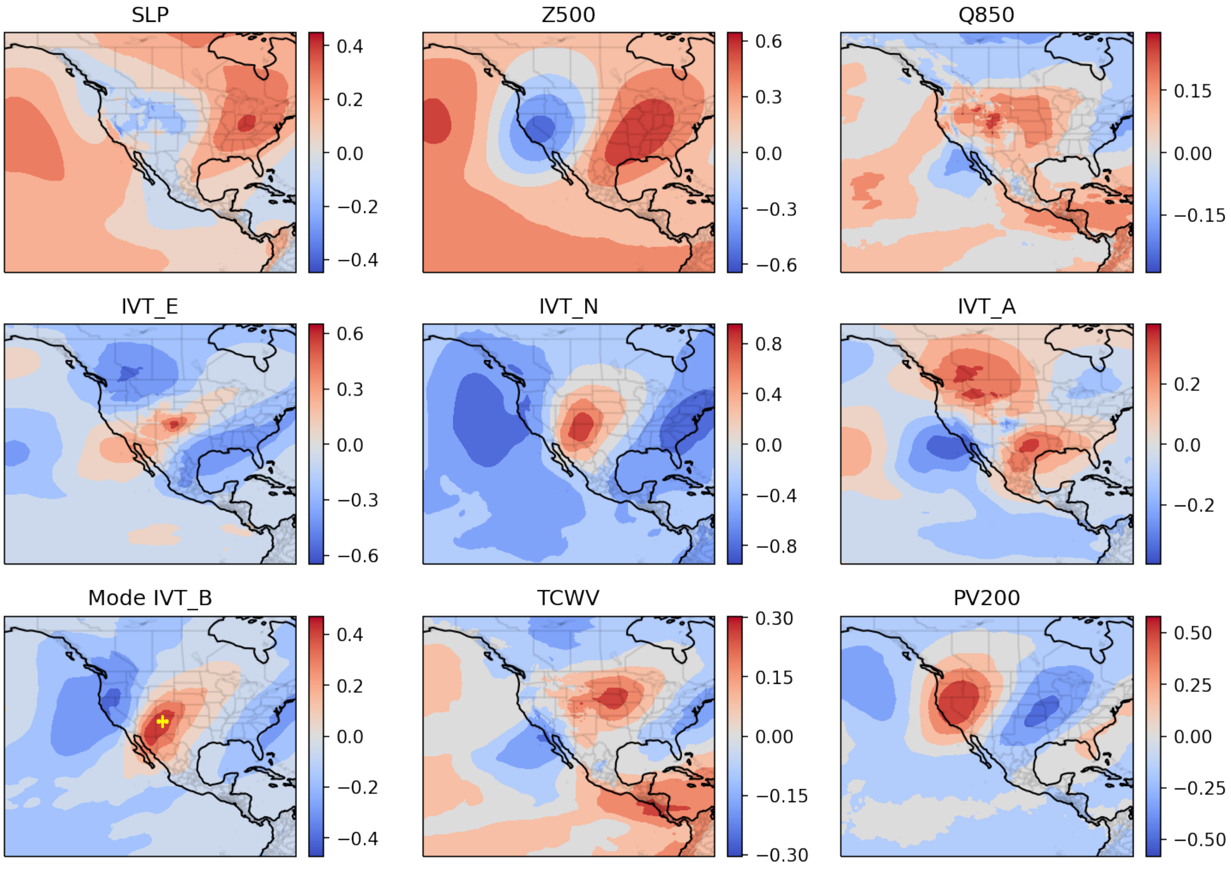

Figure S11: The second mode for 3-day averaged precipitation of subregion 4.

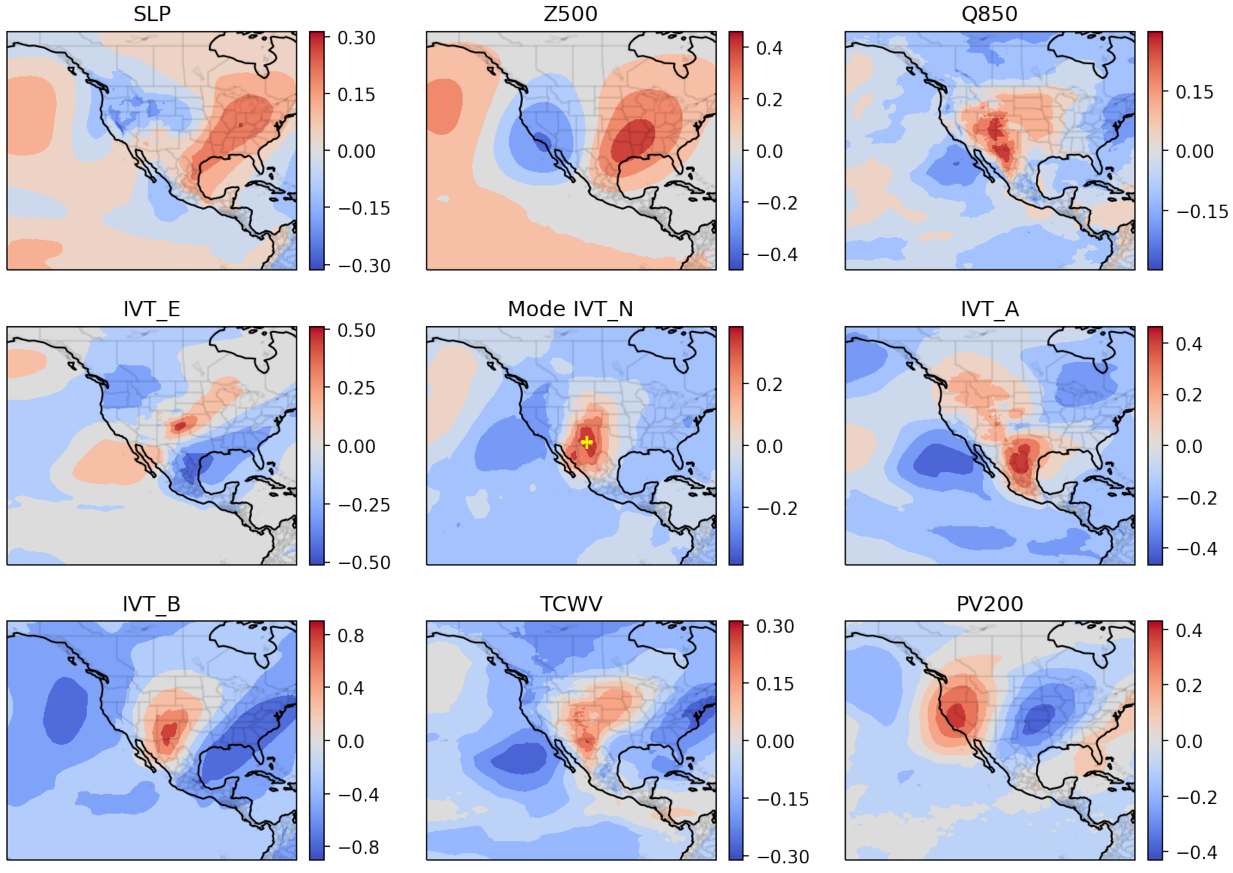

Figure S12: The second mode for 3-day averaged precipitation of subregion 5.

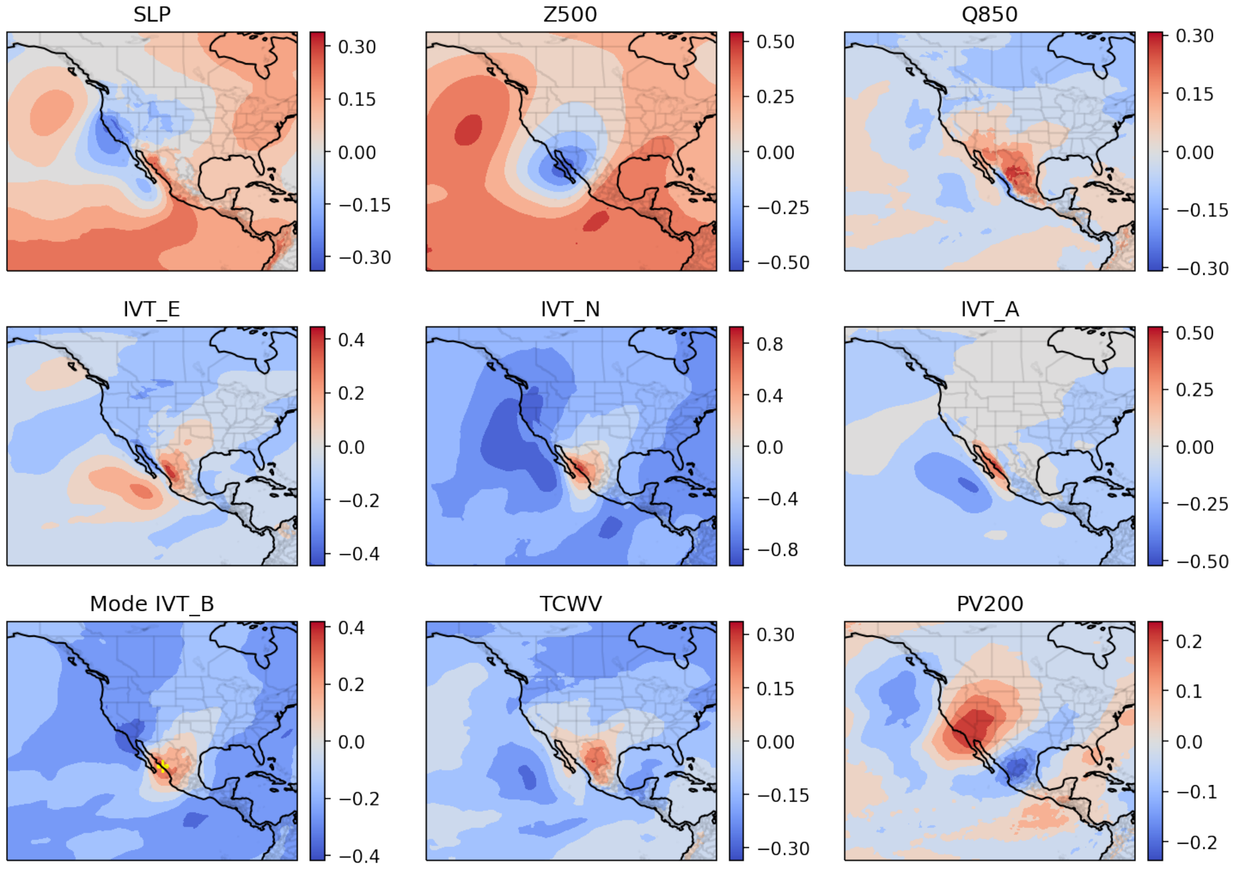

Figure S13: The second mode for 3-day averaged precipitation of subregion 6.

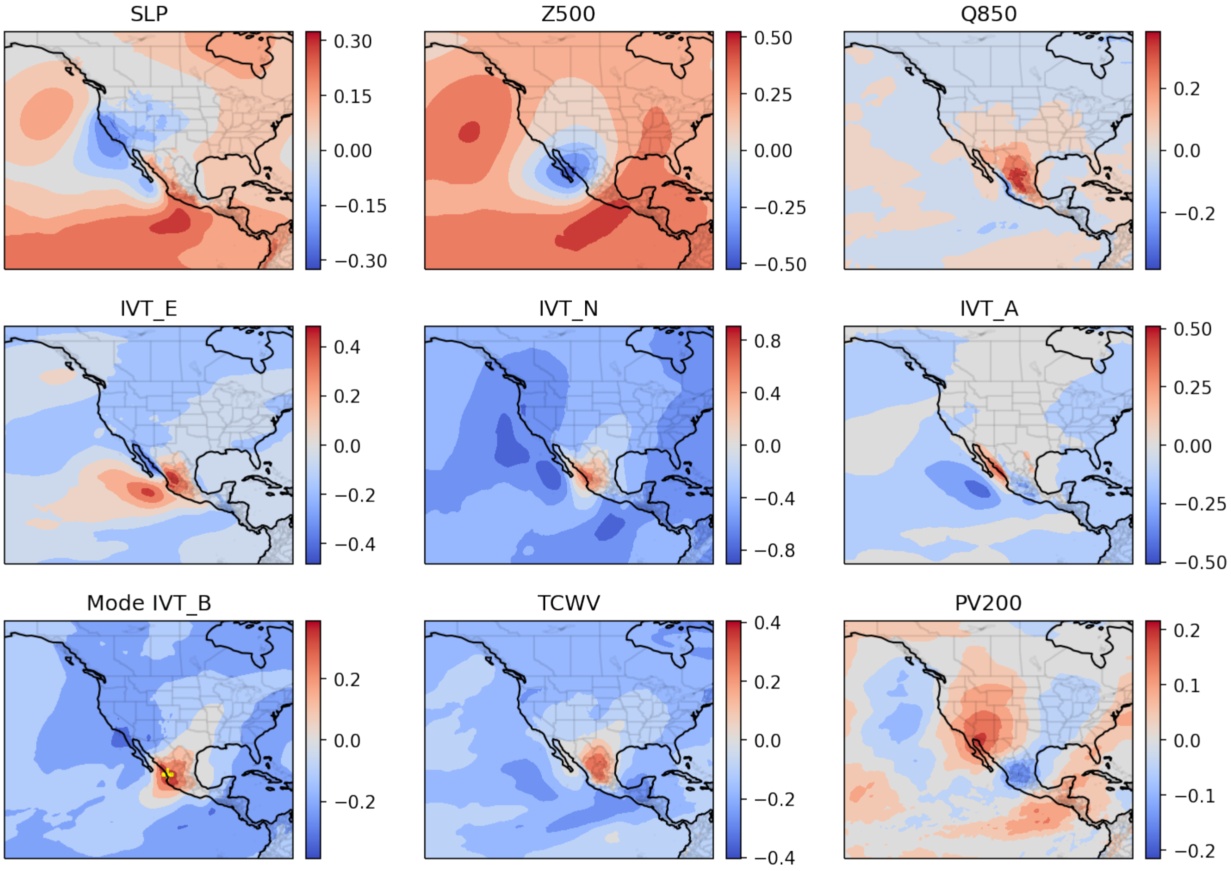

Figure S14: The second mode for 3-day averaged precipitation of subregion 7.

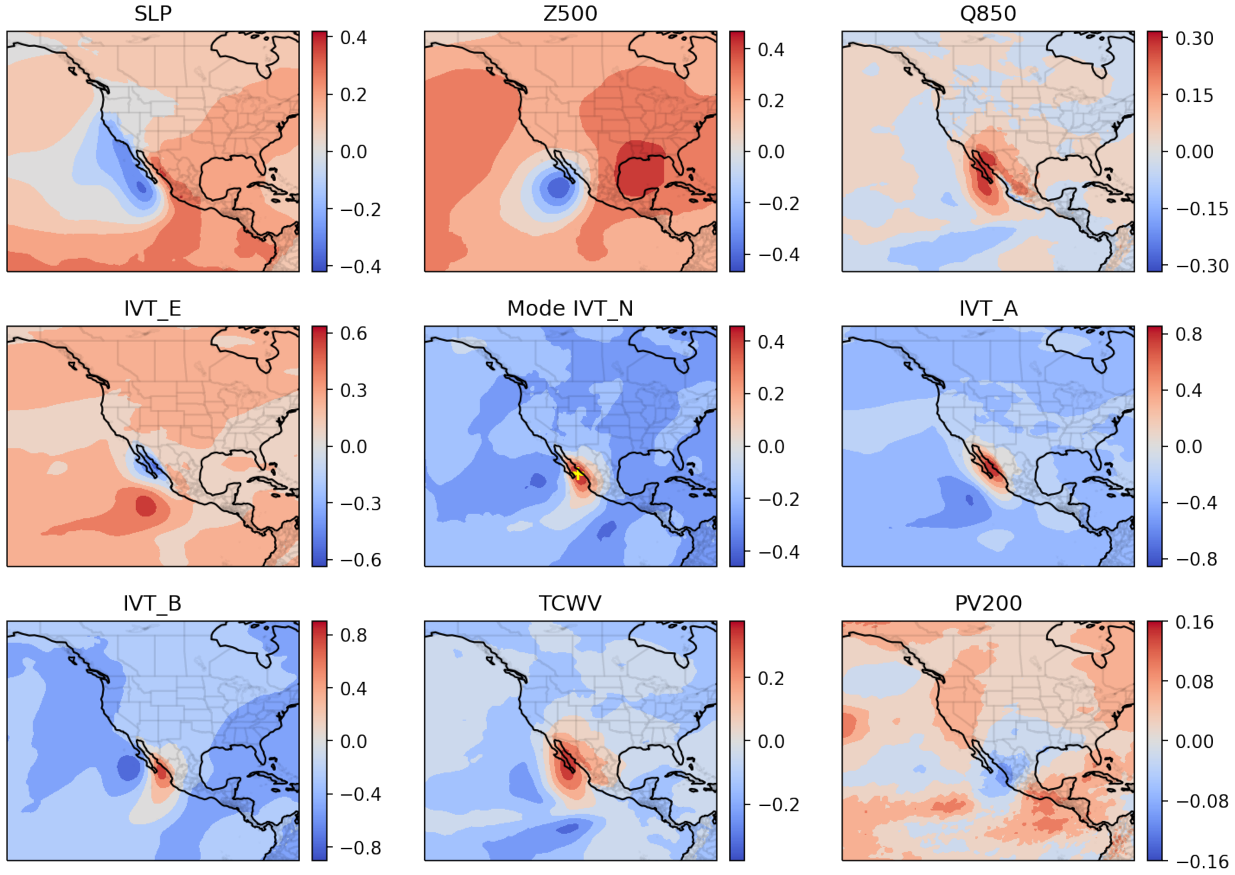

Figure S15: The first mode for daily precipitation of subregion 1.

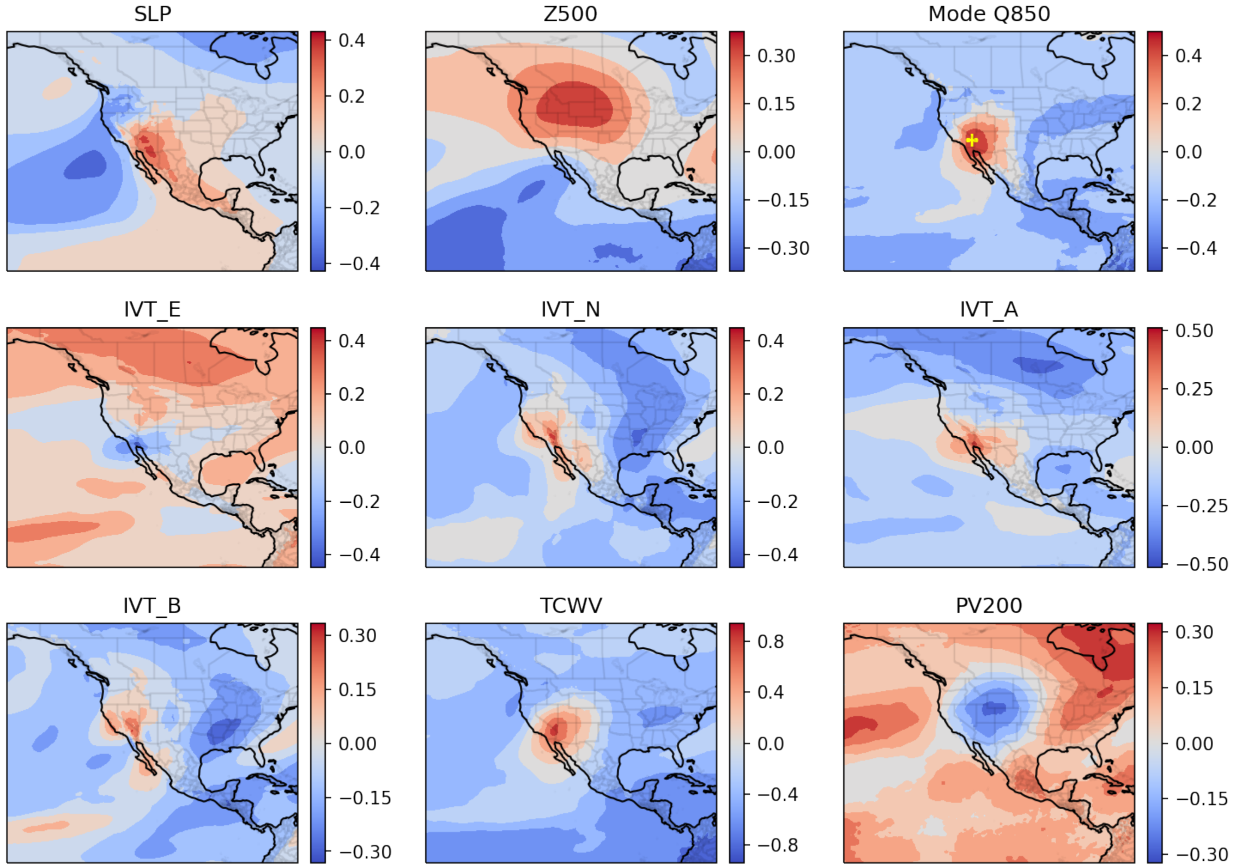

Figure S16: The first mode for daily precipitation of subregion 2.

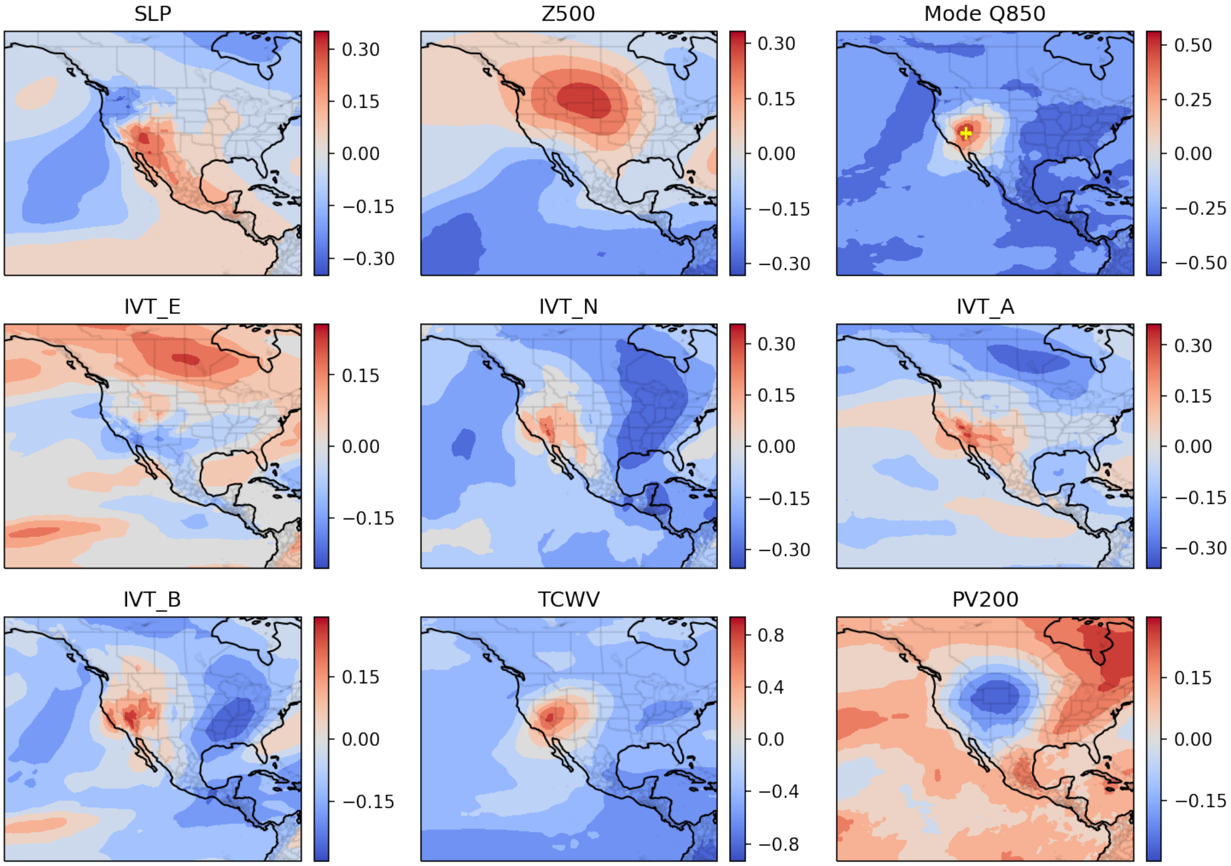

Figure S17: The first mode for daily precipitation of subregion 3.

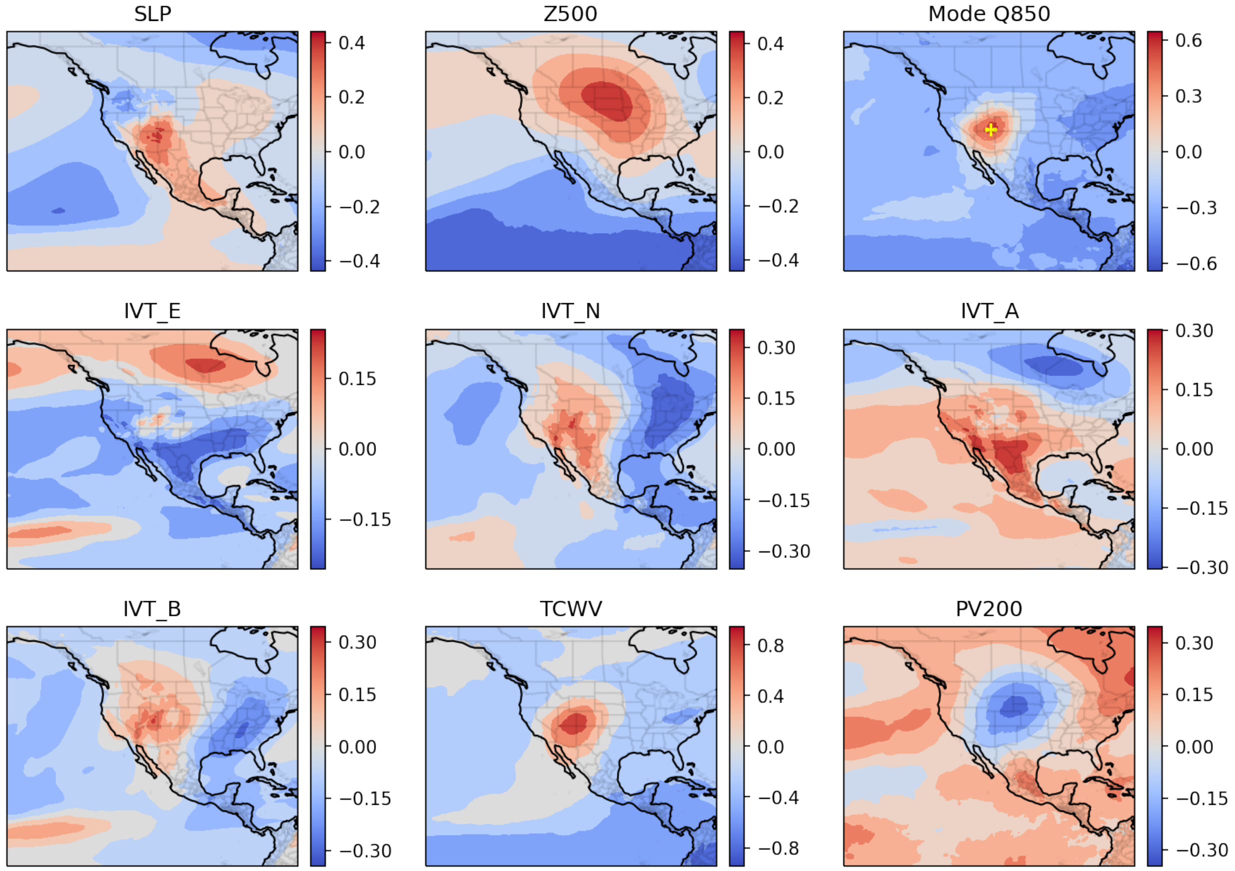

Figure S18: The first mode for daily precipitation of subregion 4.

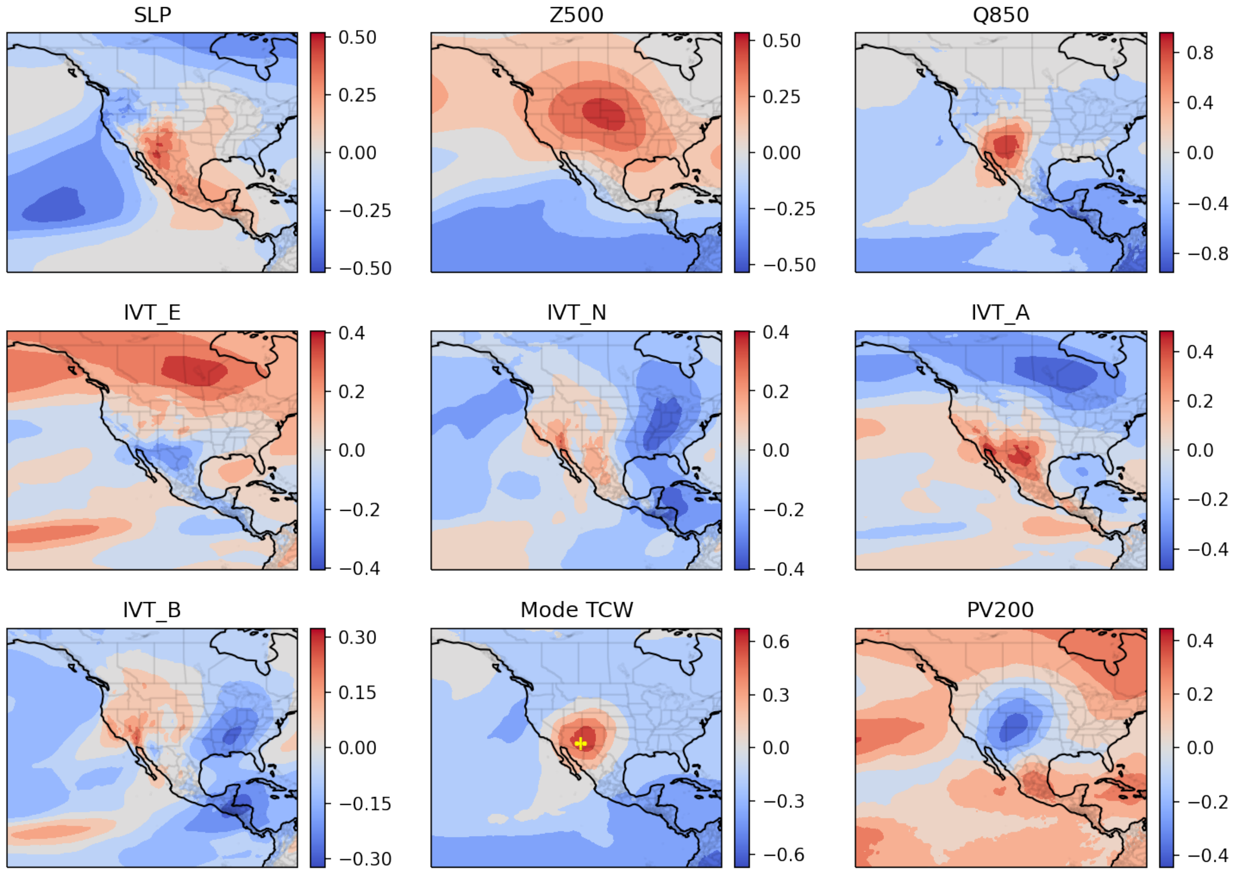

Figure S19: The first mode for daily precipitation of subregion 5.

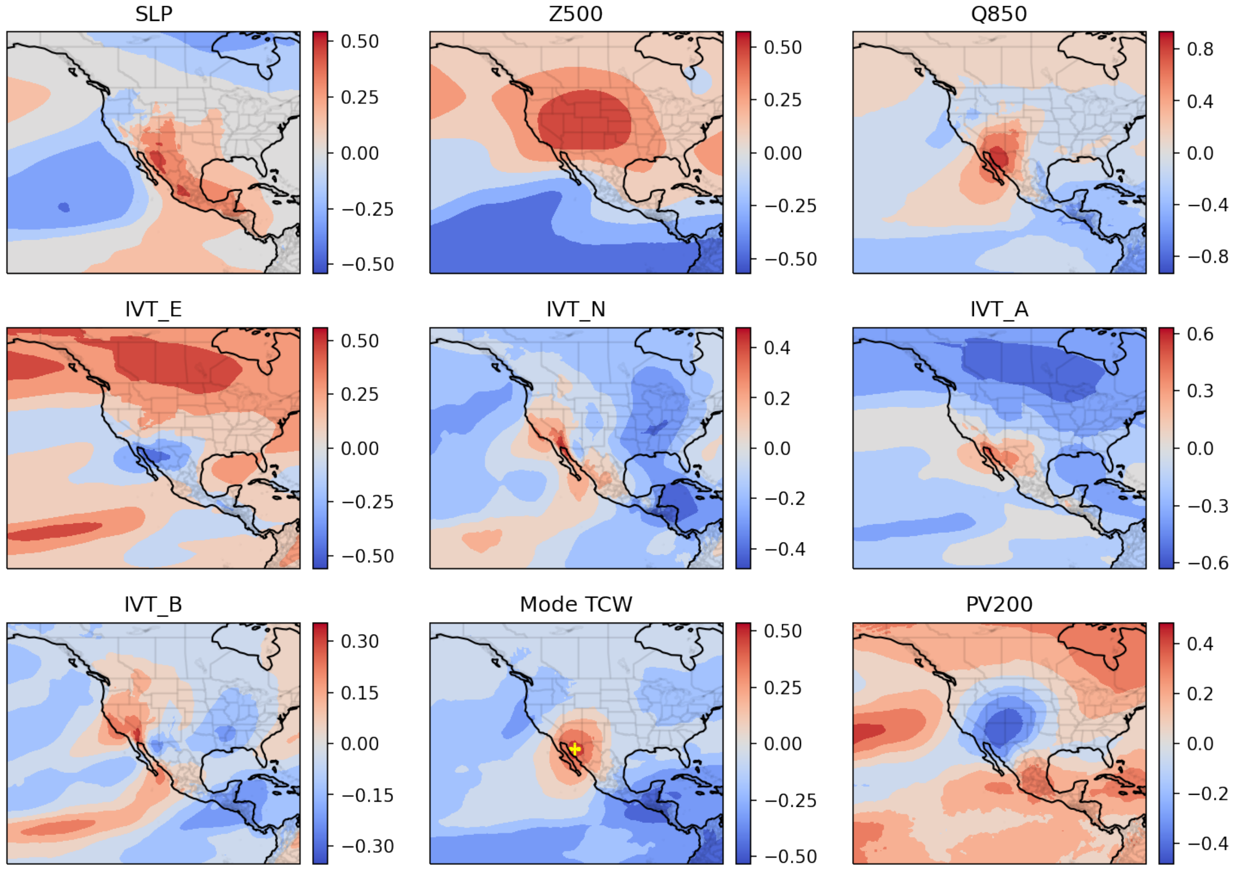

Figure S20: The first mode for daily precipitation of subregion 6.

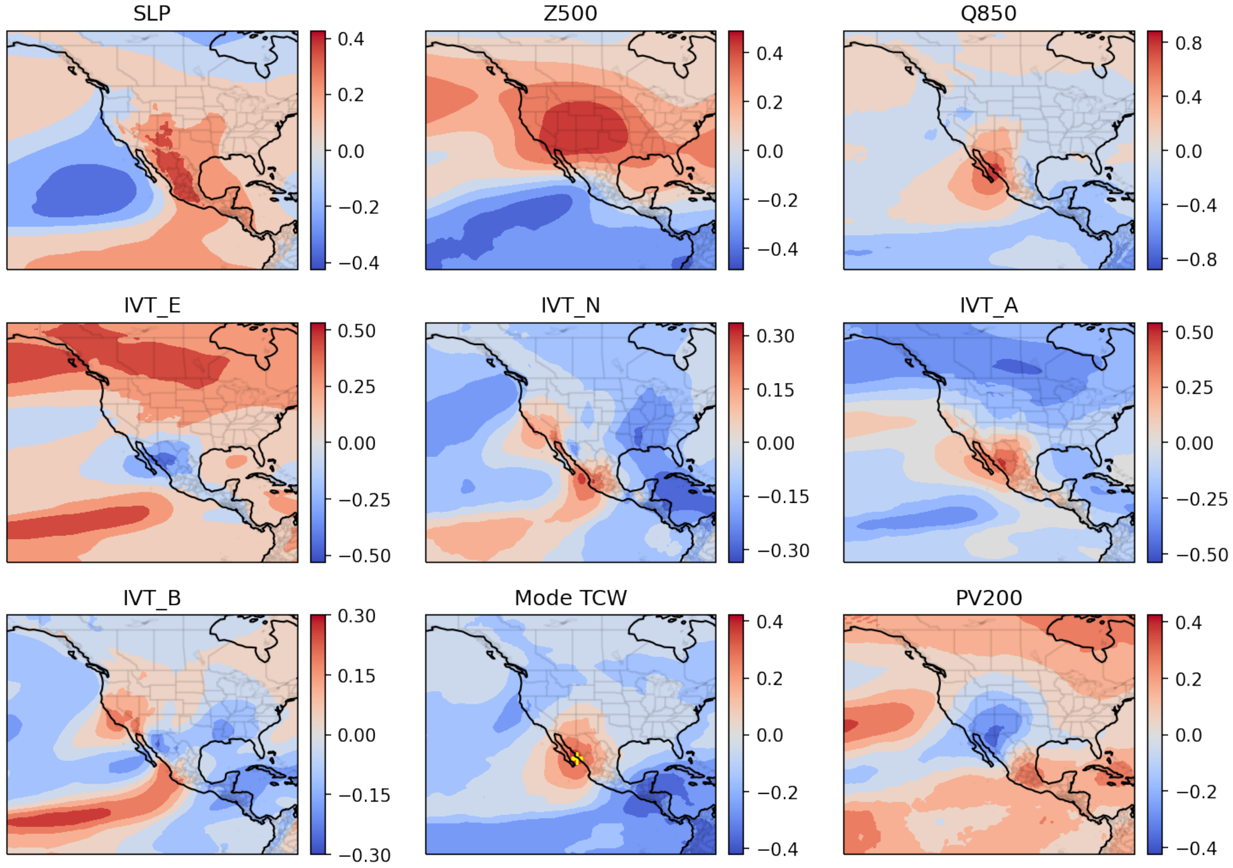

Figure S21: The first mode for daily precipitation of subregion 7.

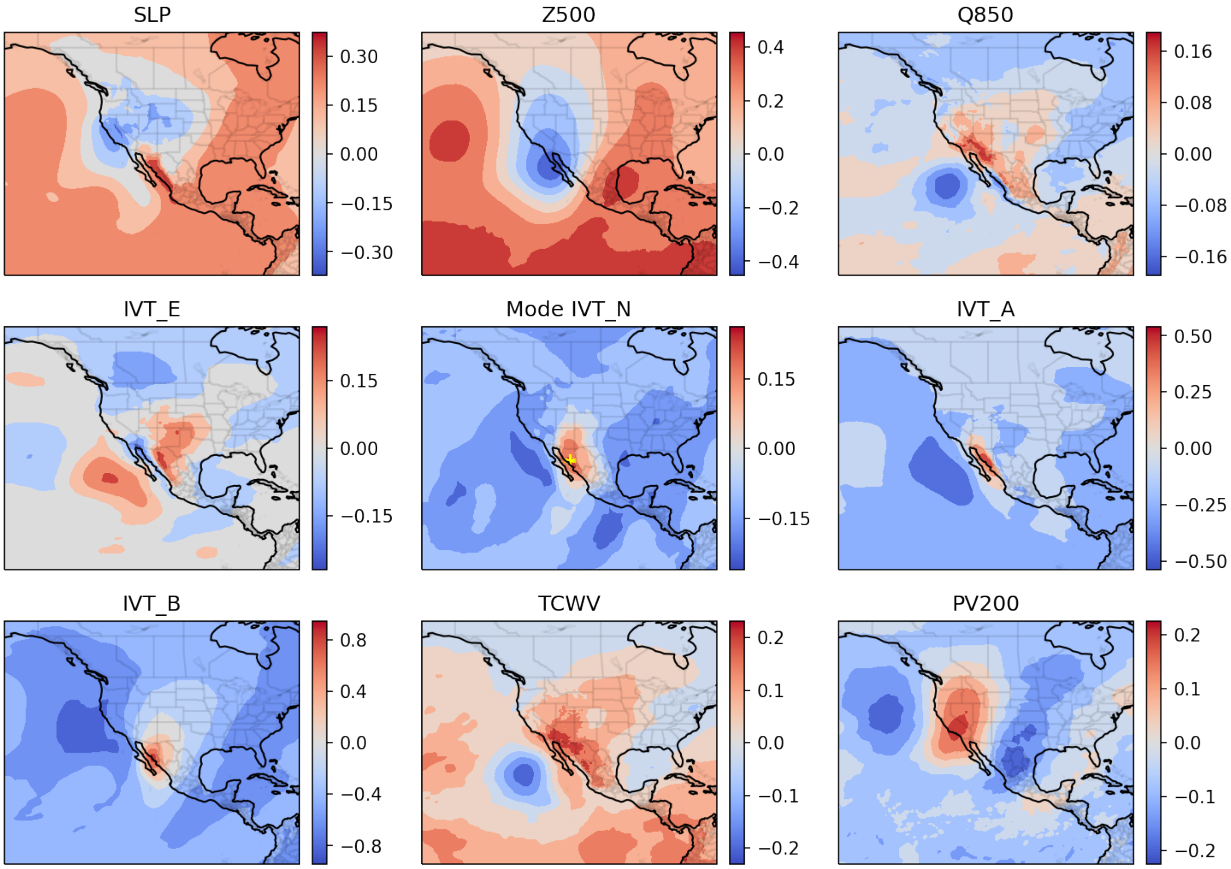

Figure S22: The second mode for daily precipitation of subregion 2.

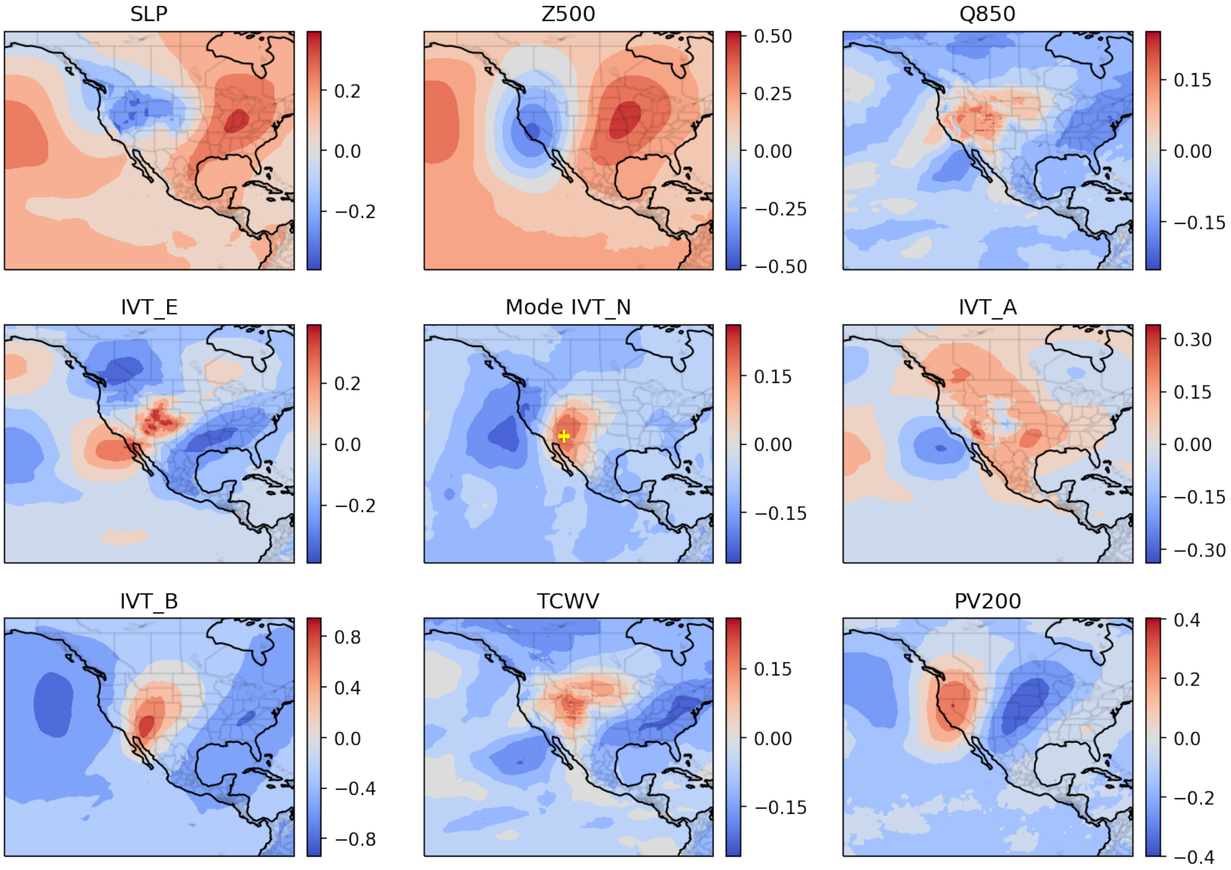

Figure S23: The second mode for daily precipitation of subregion 3.

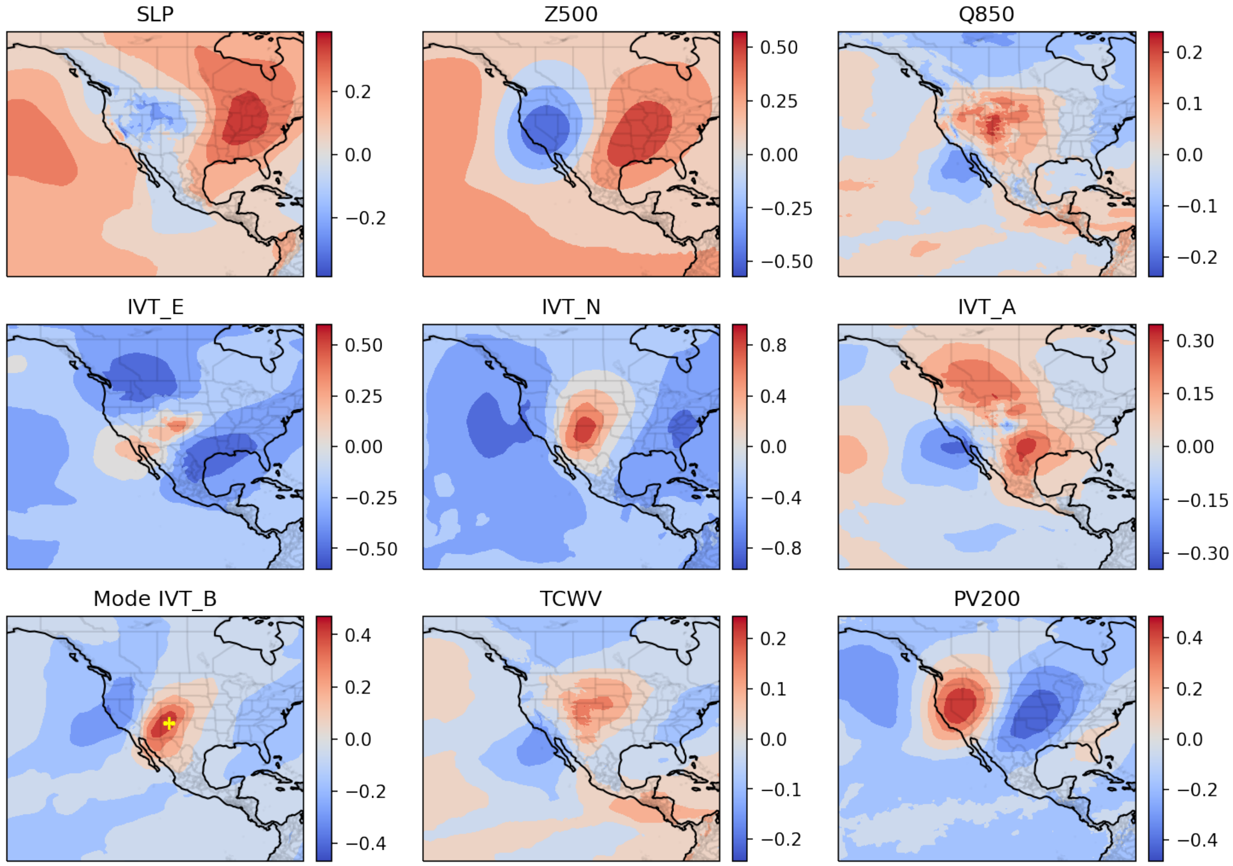

Figure S24: The second mode for daily precipitation of subregion 4.

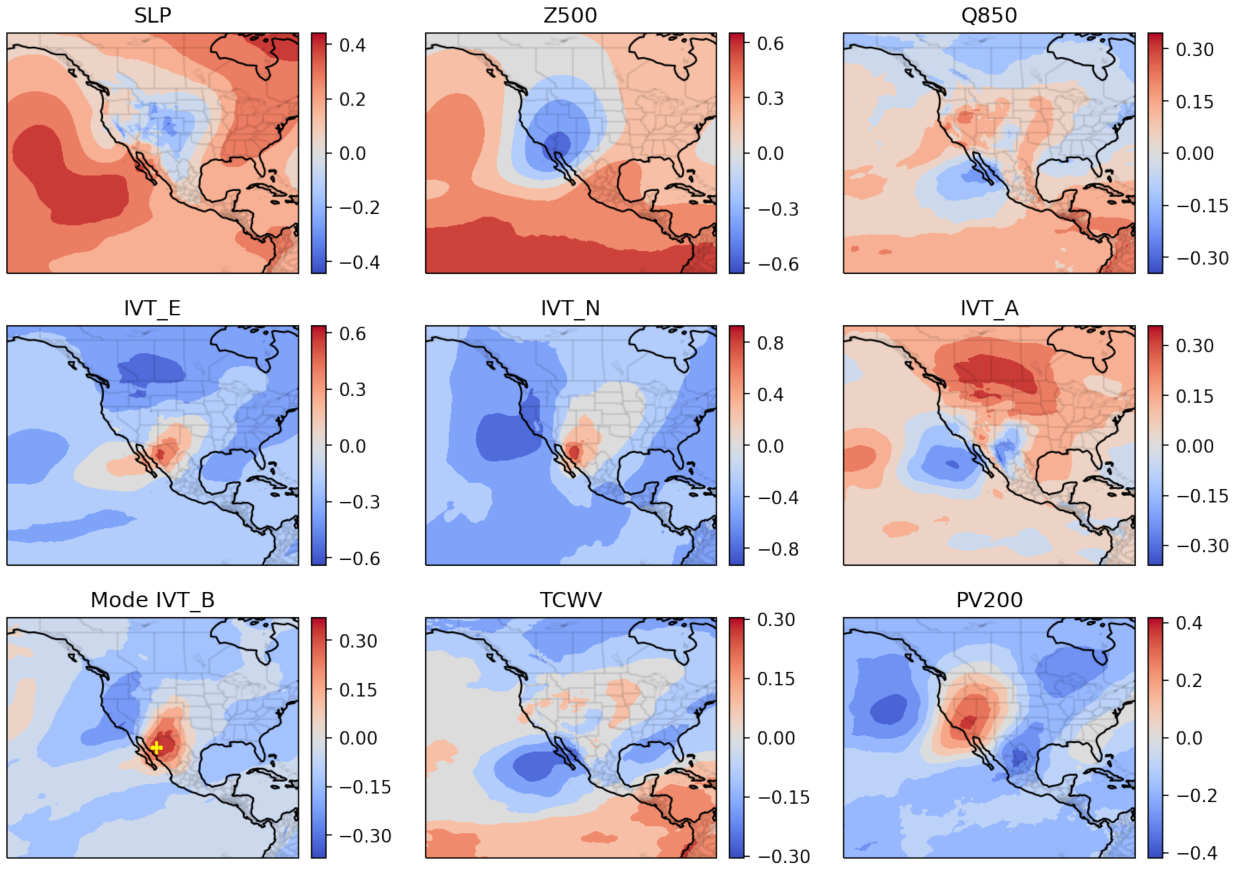

Figure S25: The second mode for daily precipitation of subregion 5.

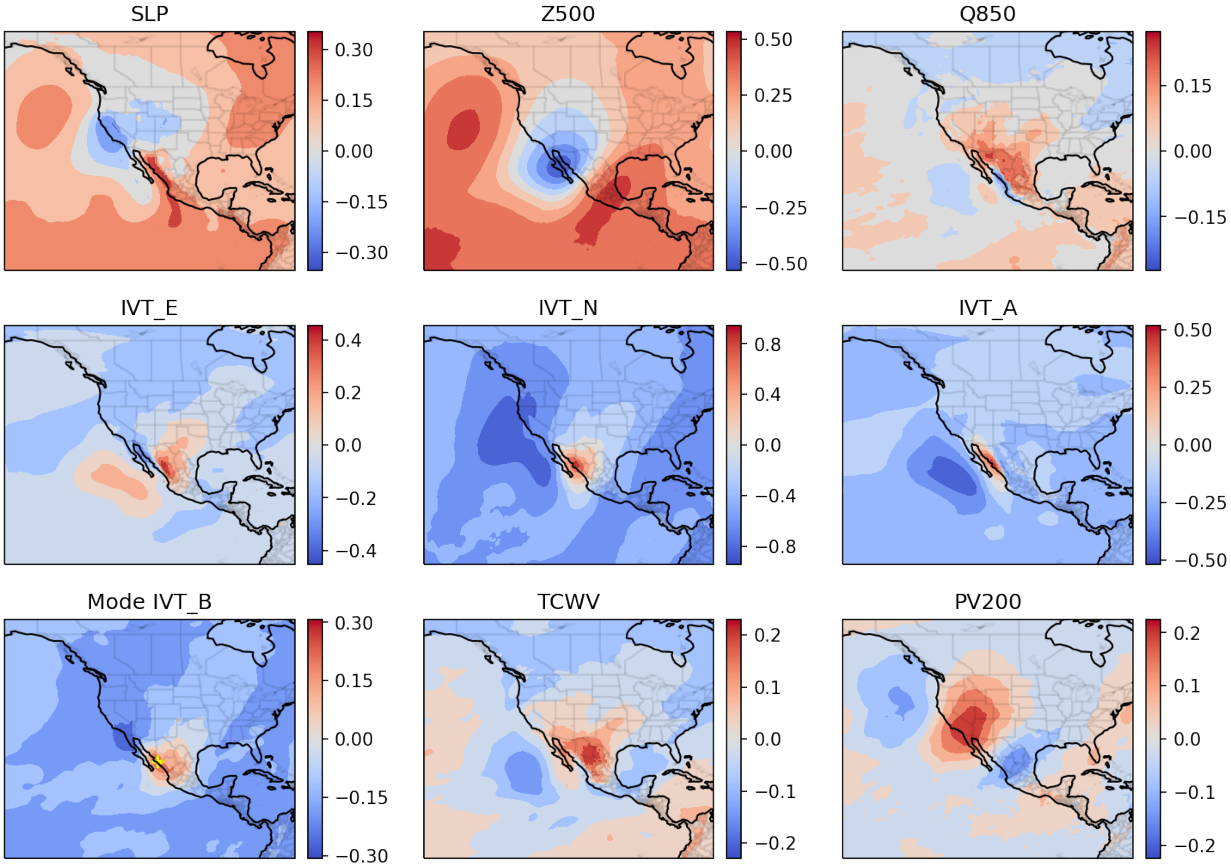

Figure S26: The second mode for daily precipitation of subregion 6.

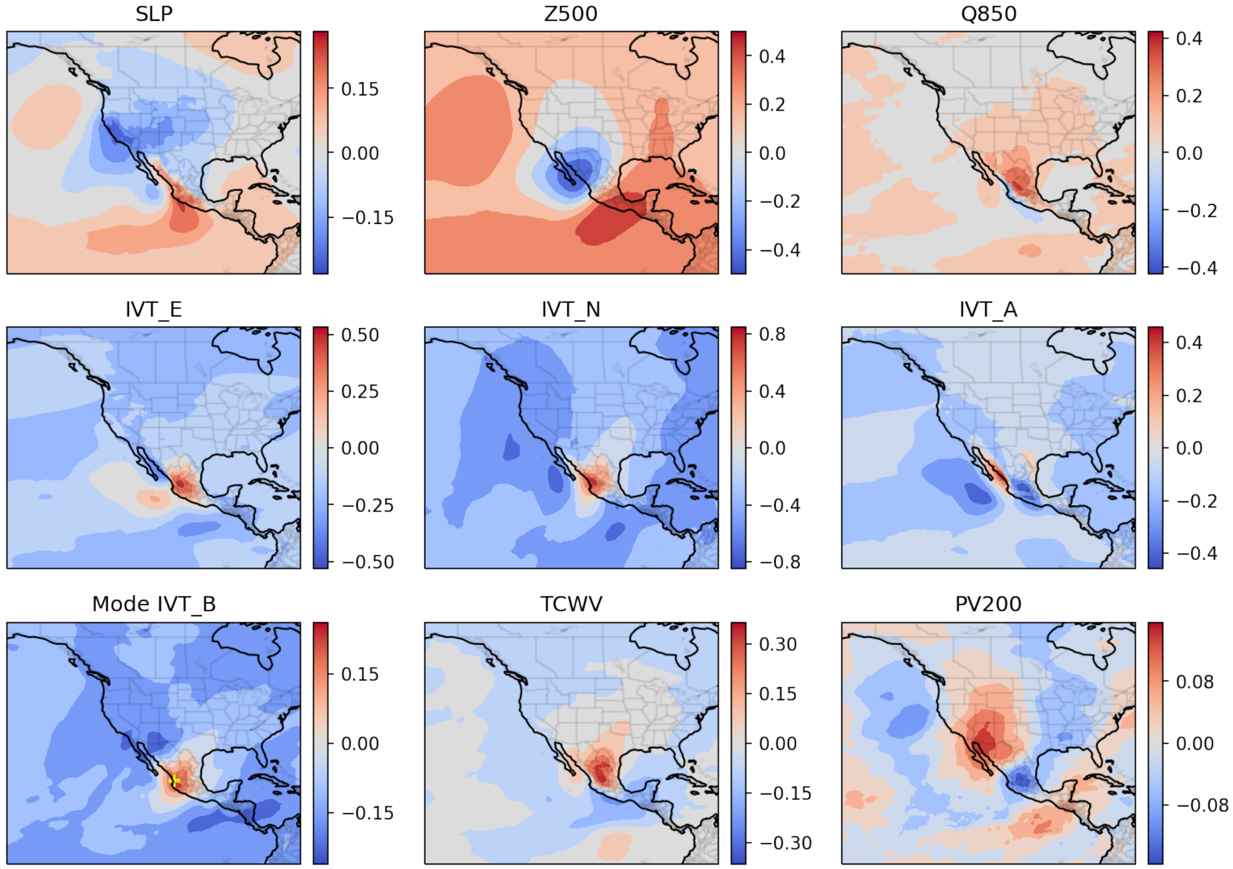

Figure S27: The second mode for daily precipitation of subregion 7.

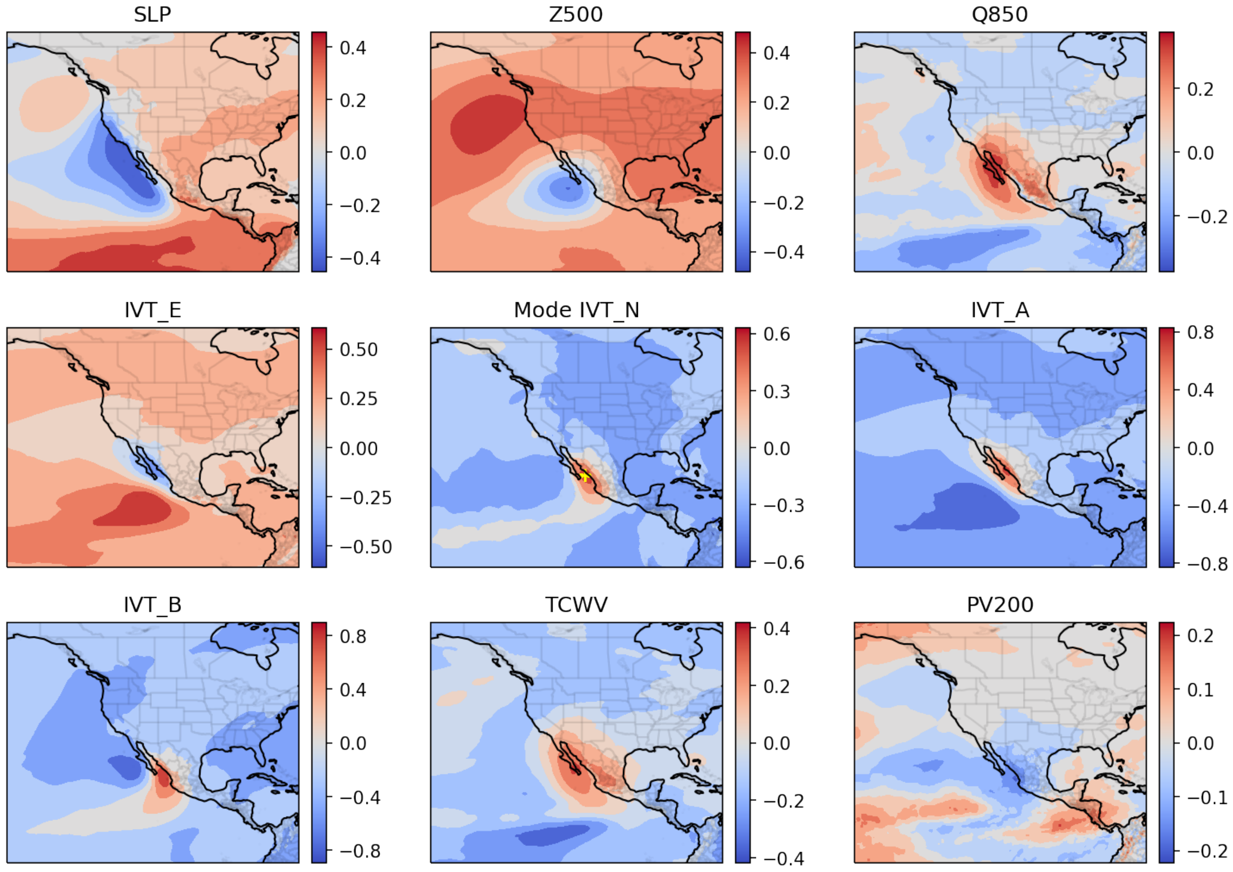

Figure S28: The first mode for 5-day averaged precipitation of subregion 1.

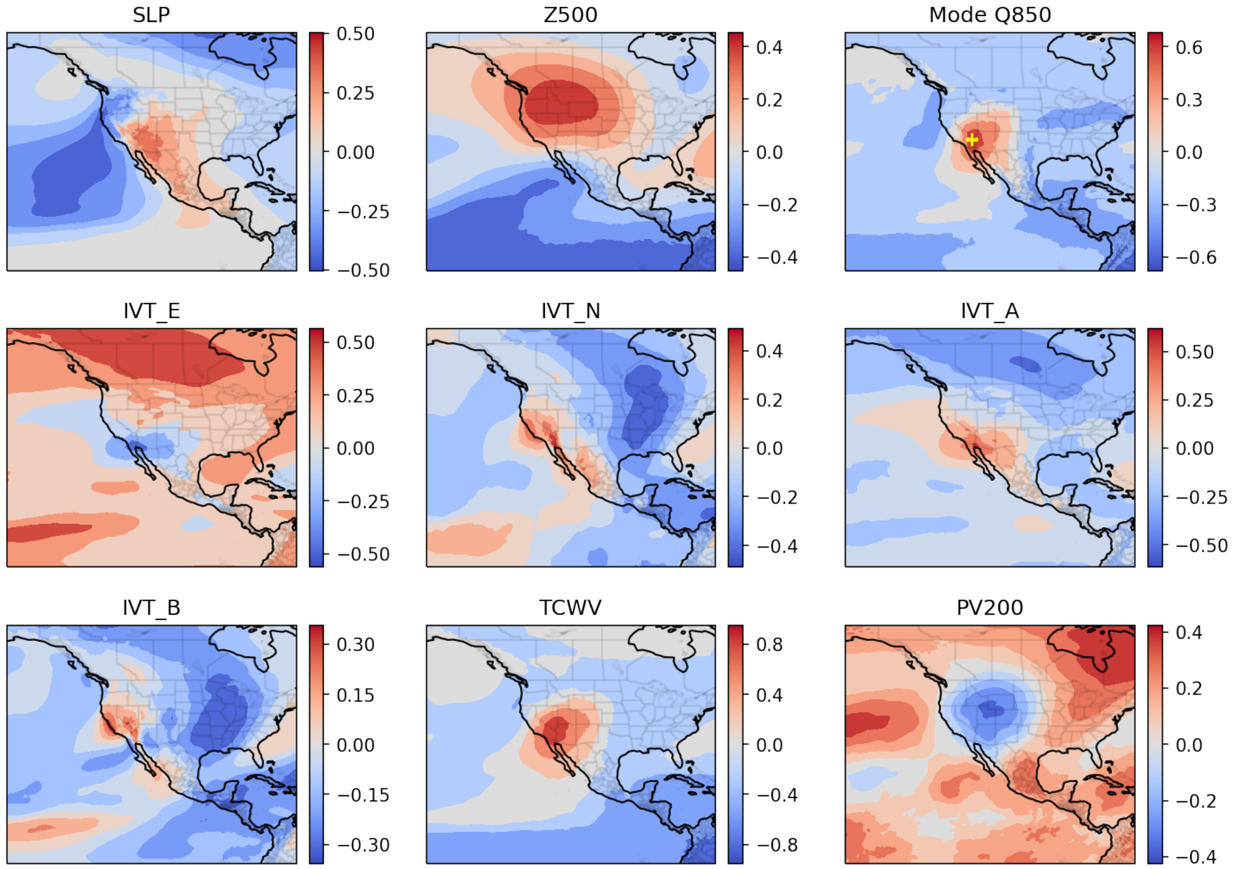

Figure S29: The first mode for 5-day averaged precipitation of subregion 2.

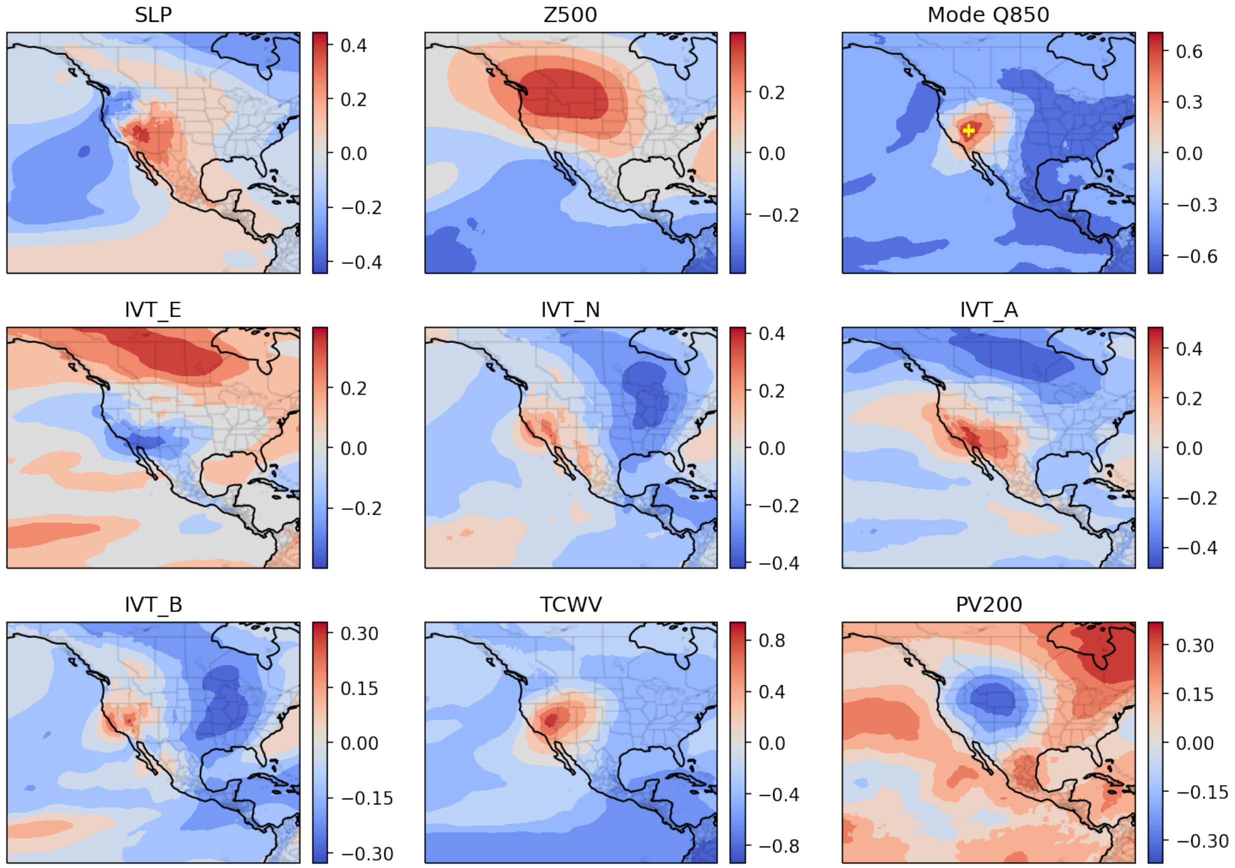

Figure S30: The first mode for 5-day averaged precipitation of subregion 3.

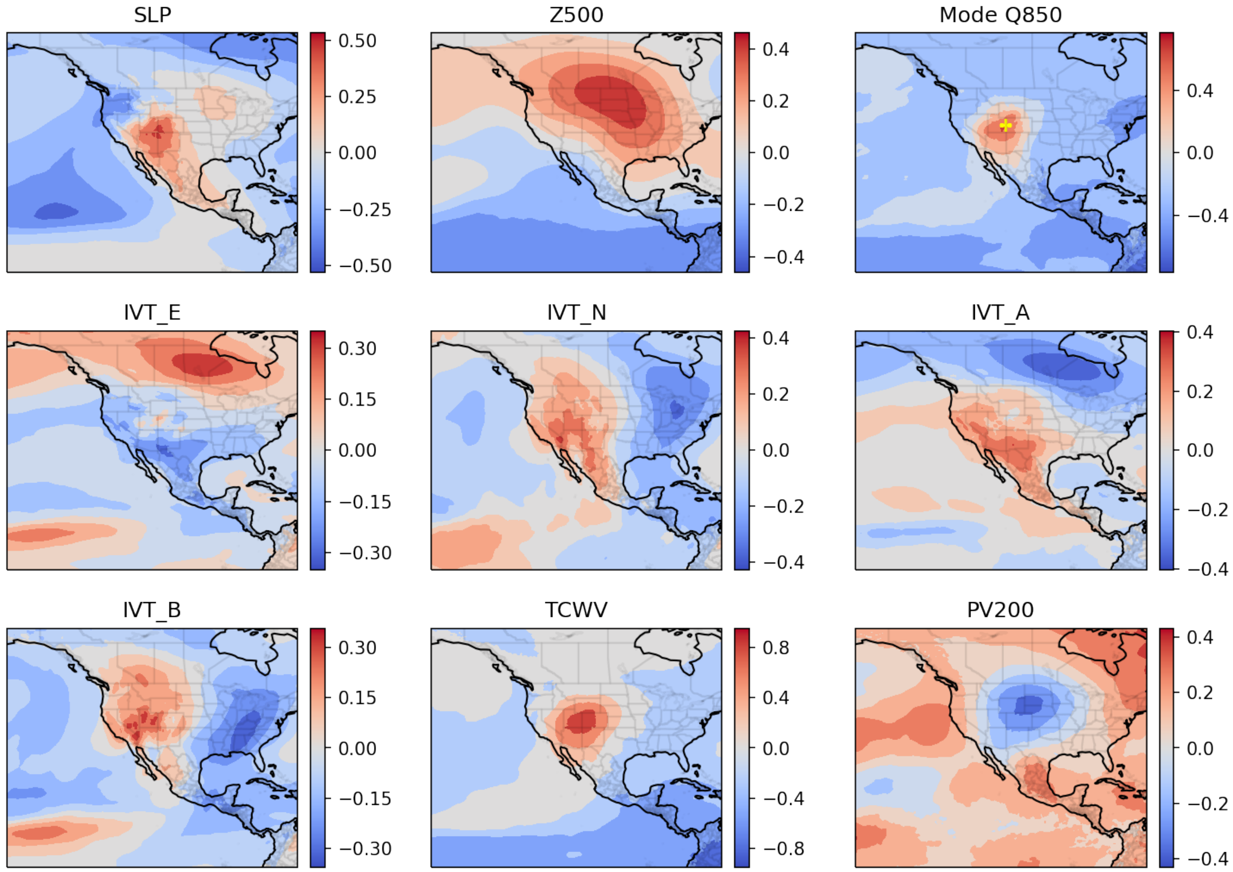

Figure S31: The first mode for 5-day averaged precipitation of subregion 4.

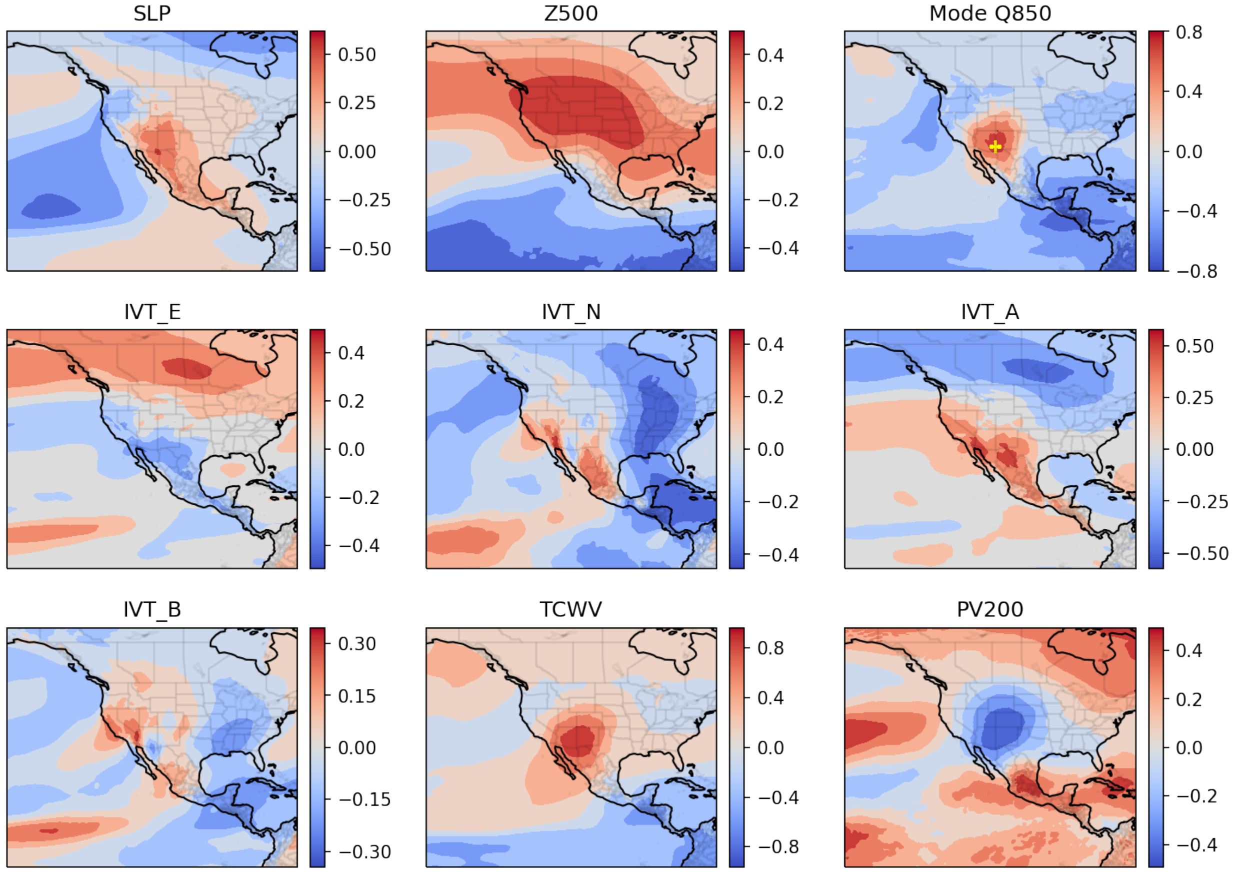

Figure S32: The first mode for 5-day averaged precipitation of subregion 5.

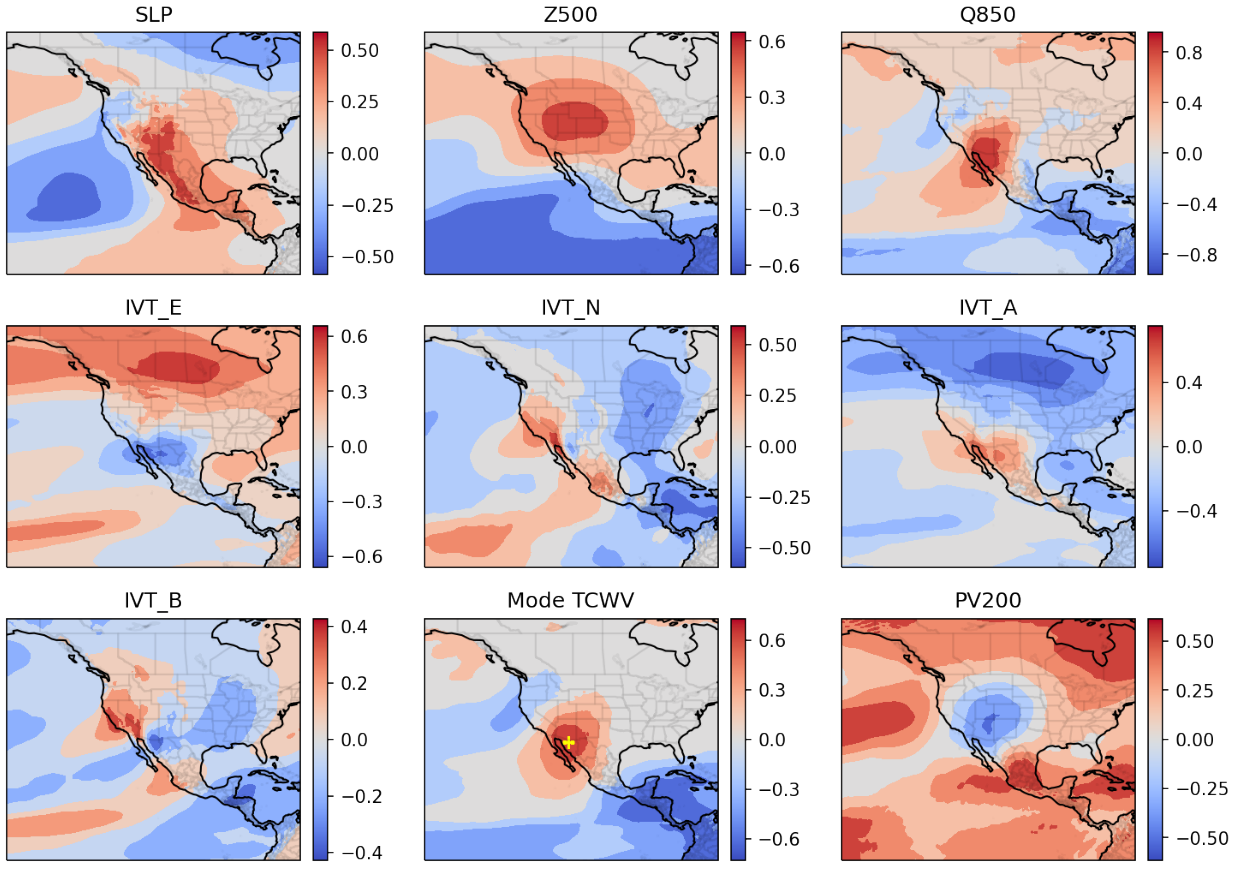

Figure S33: The first mode for 5-day averaged precipitation of subregion 6.

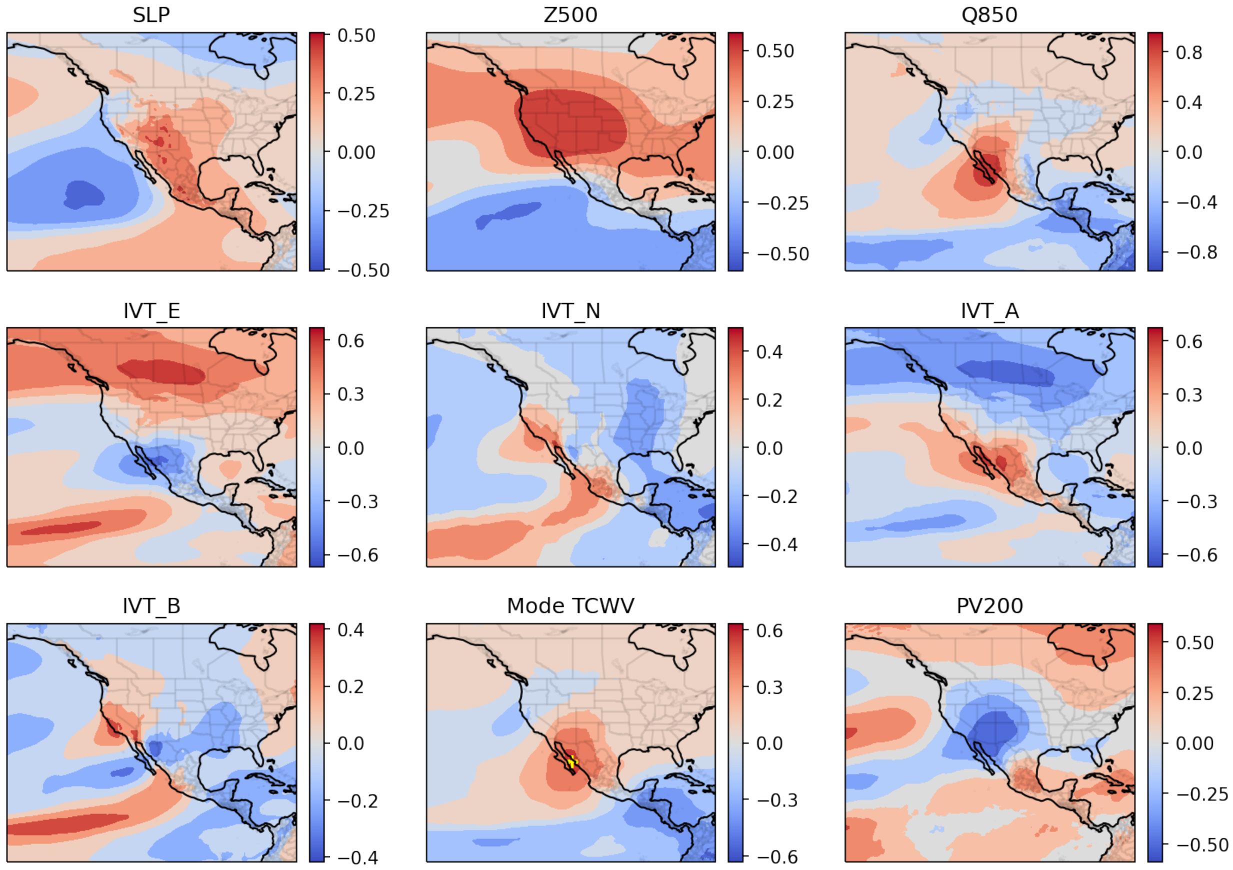

Figure S34: The first mode for 5-day averaged precipitation of subregion 7.

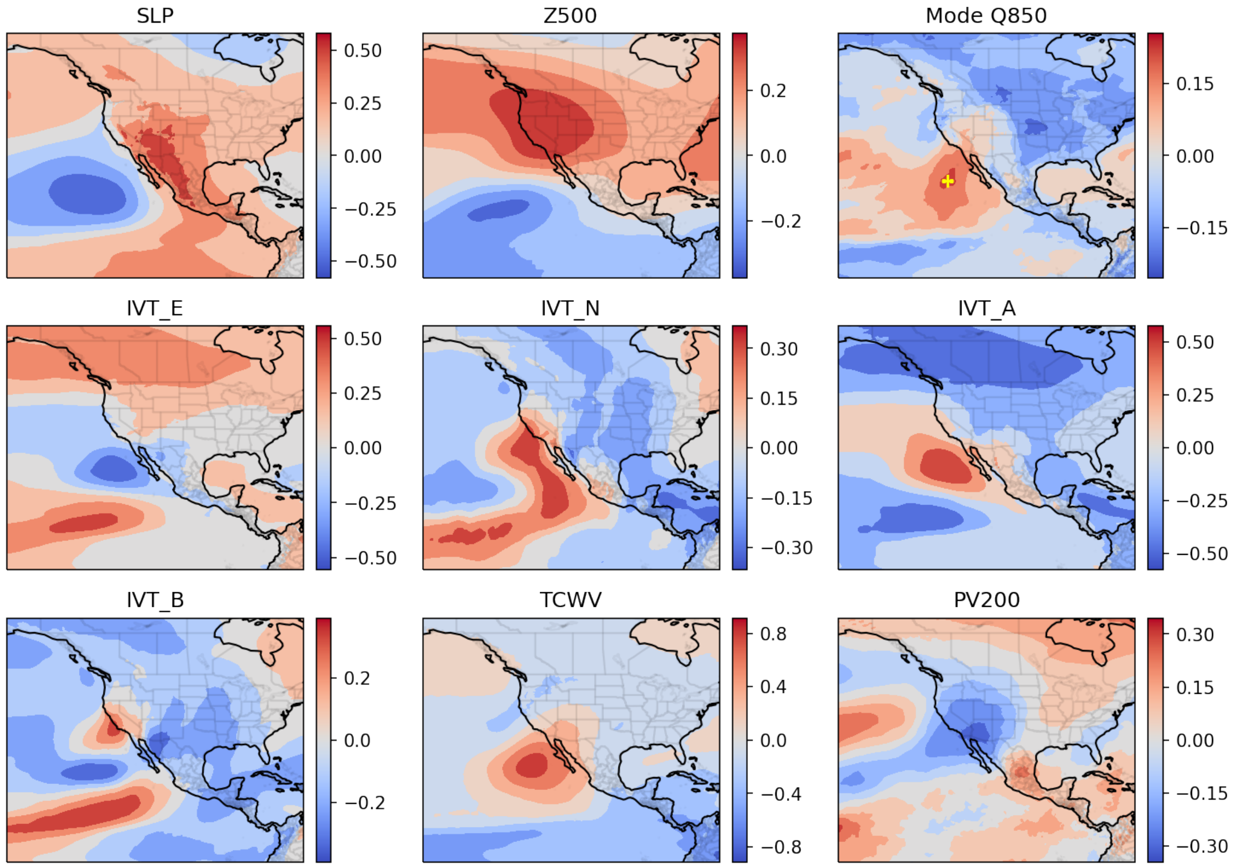

Figure S35: The second mode for 5-day averaged precipitation of subregion 1.

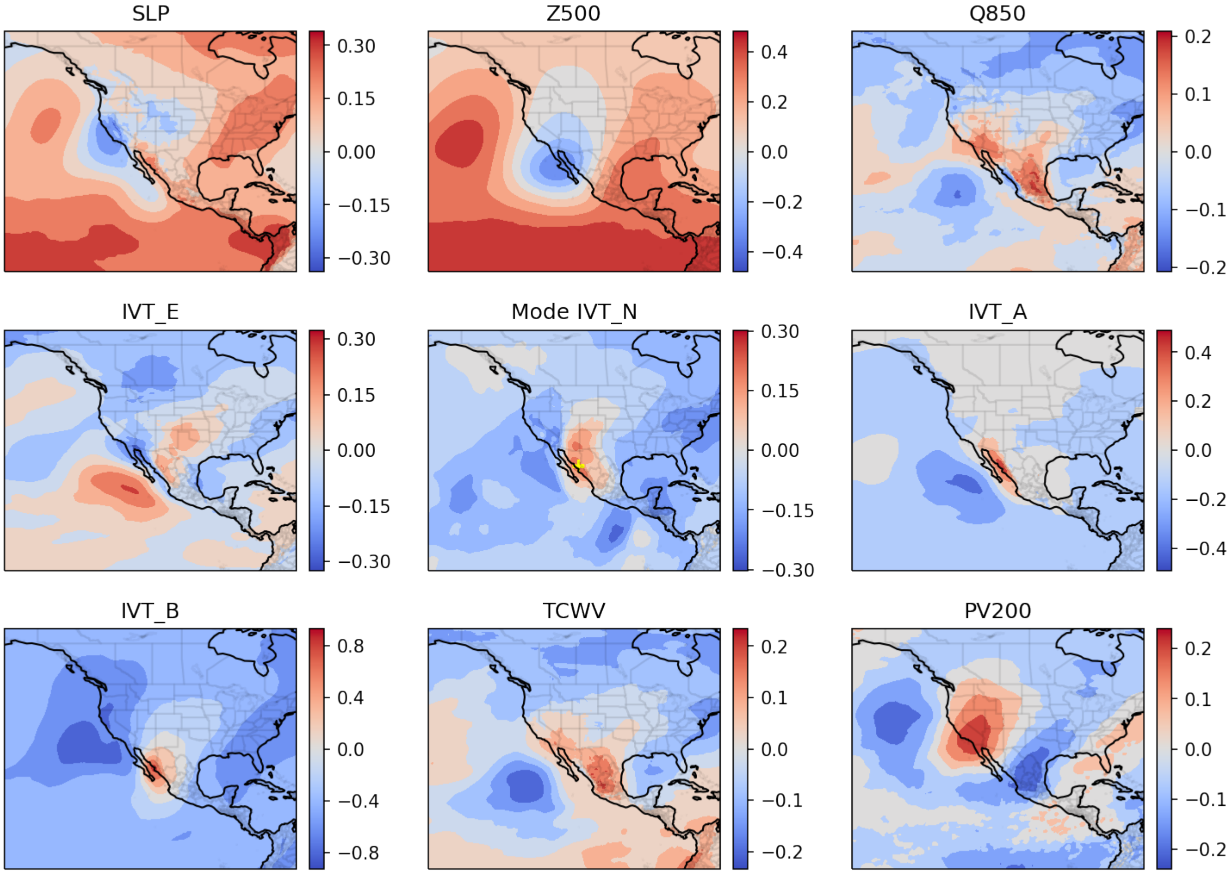

Figure S36: The second mode for 5-day averaged precipitation of subregion 2.

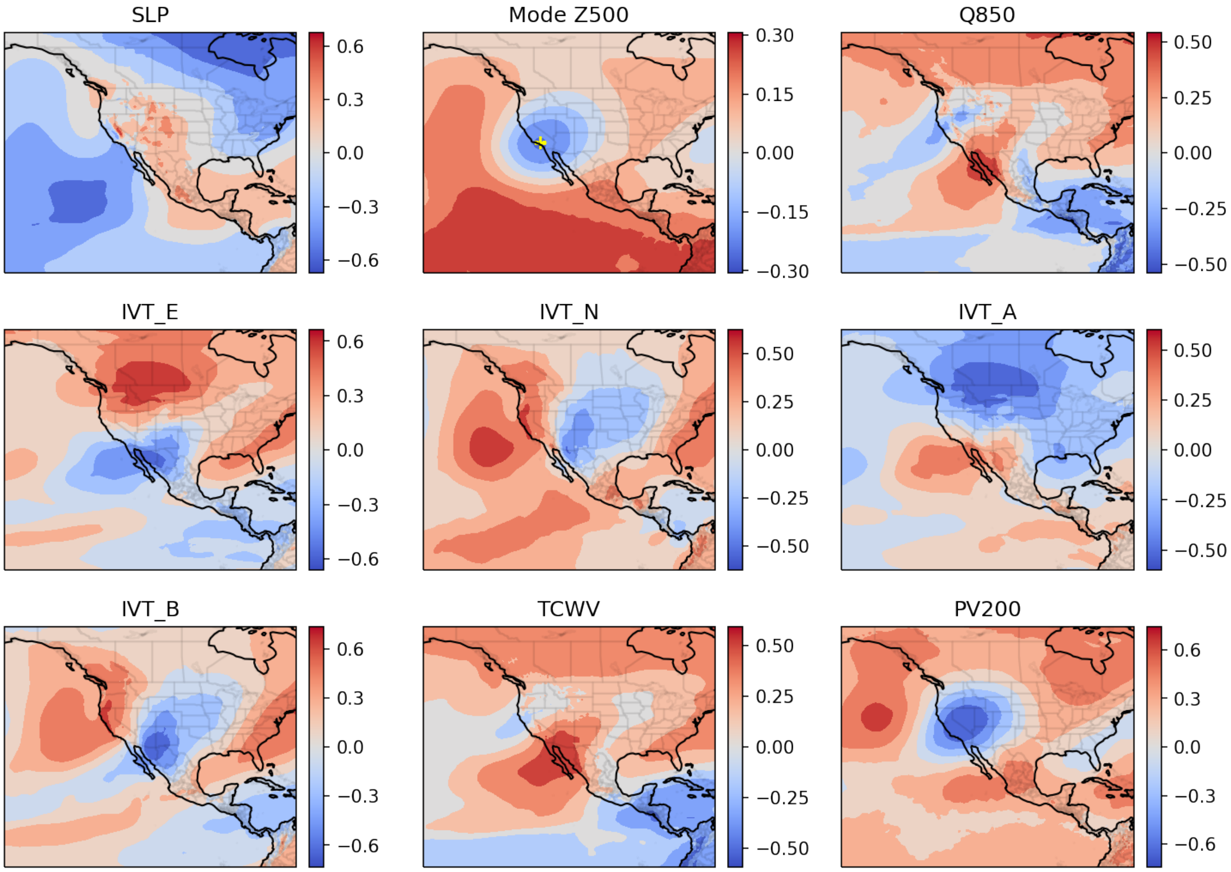

Figure S37: The second mode for 5-day averaged precipitation of subregion 3.

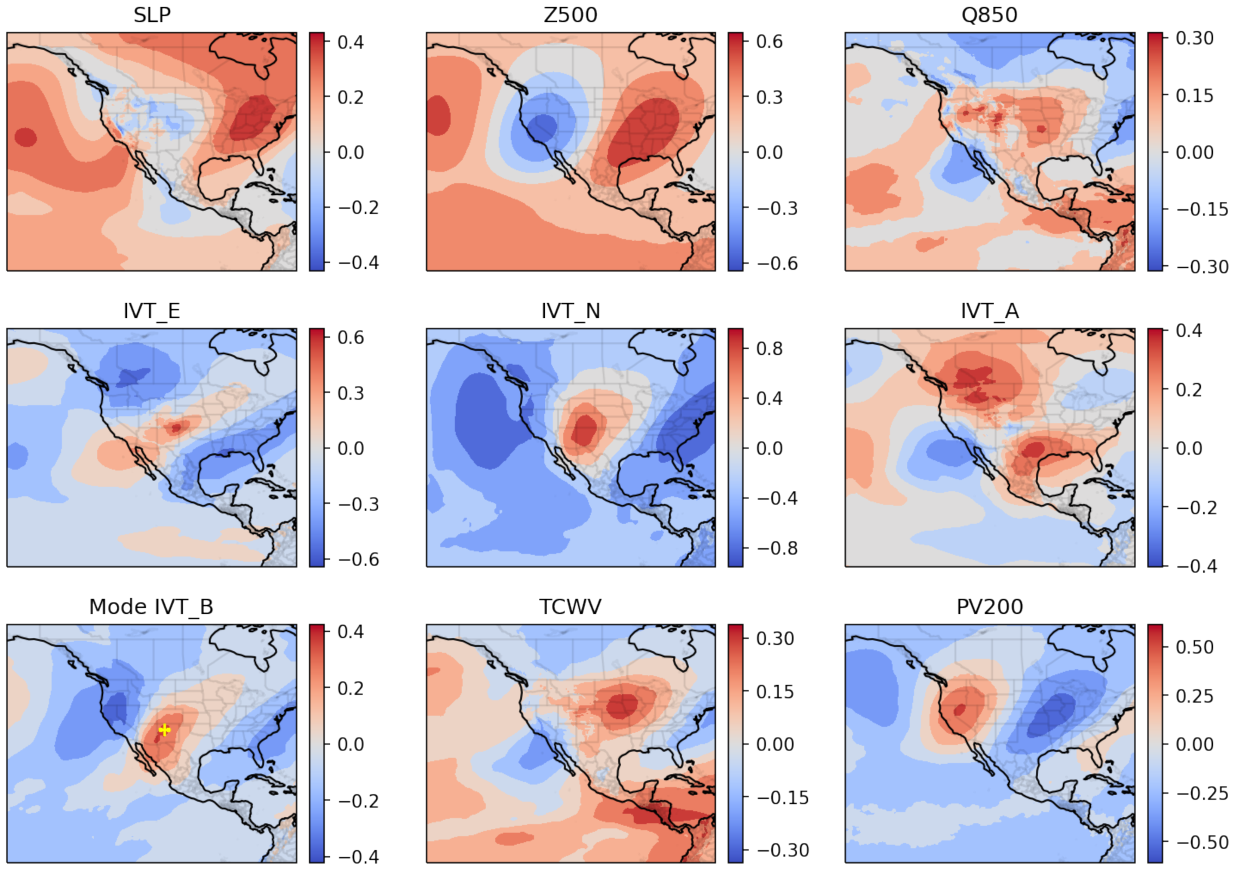

Figure S38: The second mode for 5-day averaged precipitation of subregion 4.

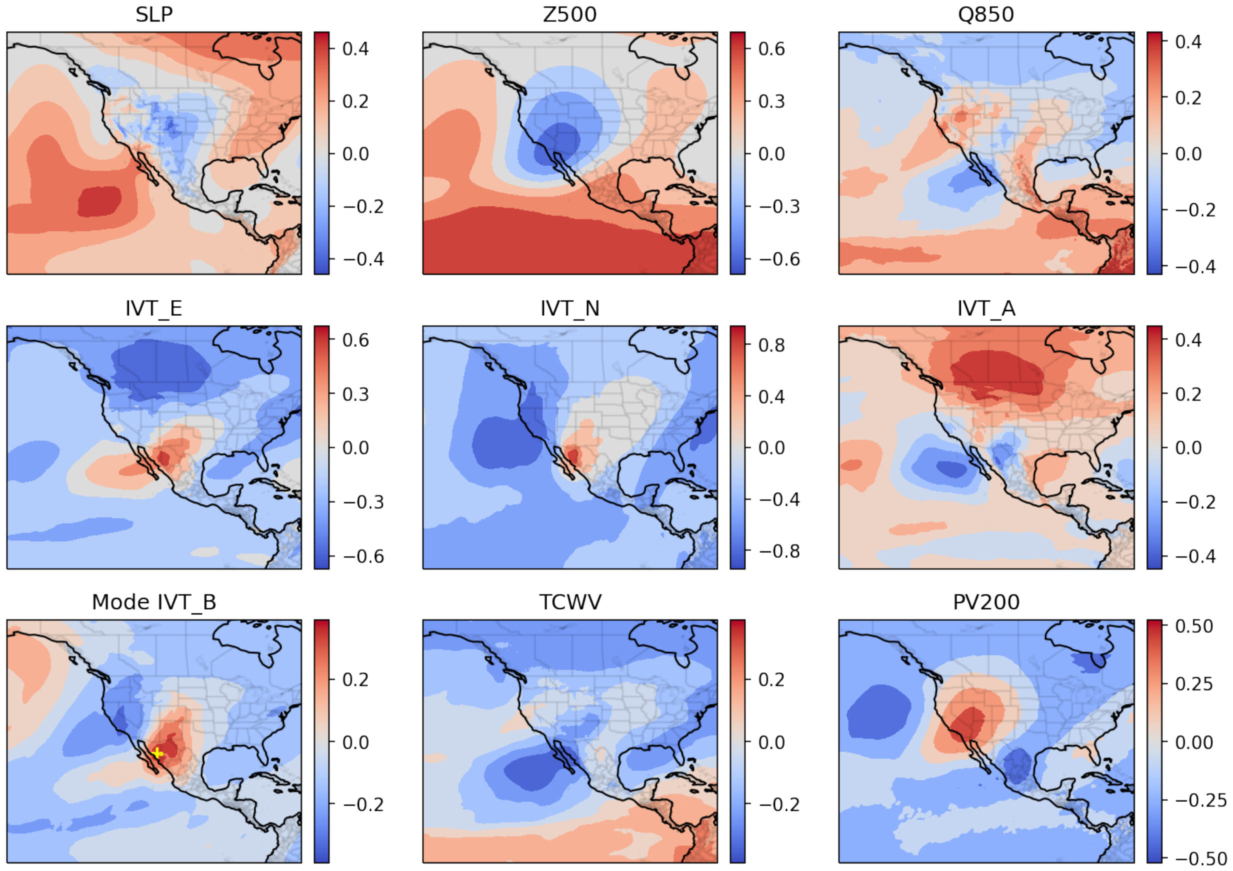

Figure S39: The second mode for 5-day averaged precipitation of subregion 5.

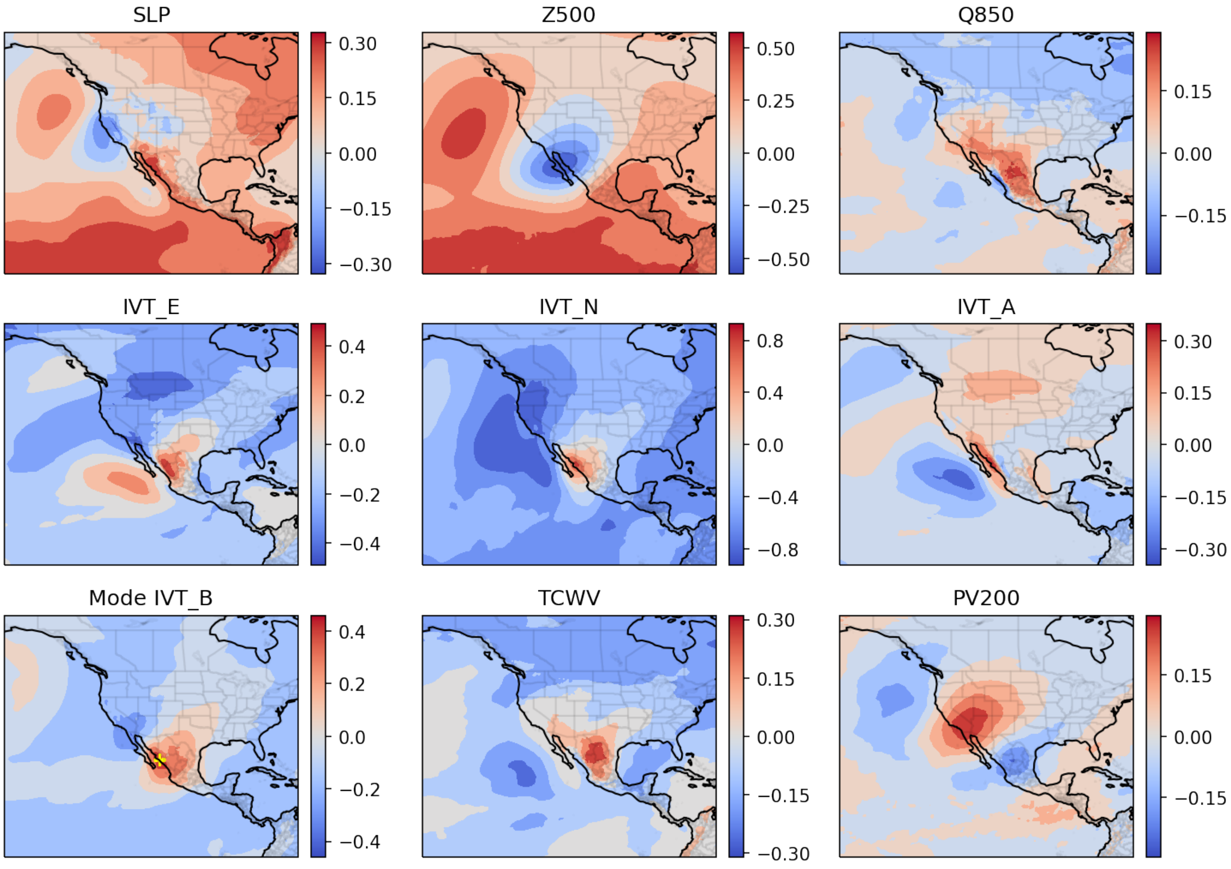

Figure S40: The second mode for 5-day averaged precipitation of subregion 6.

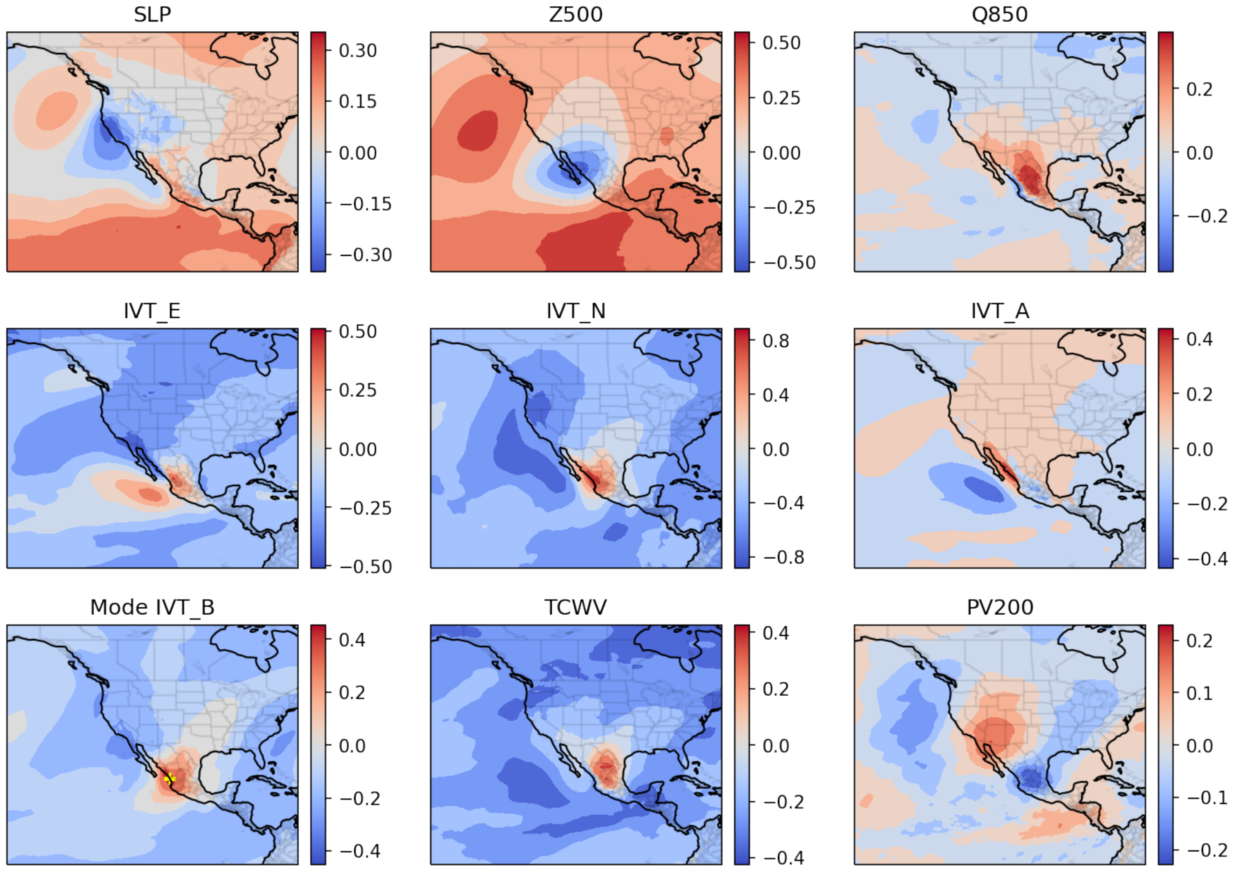

Figure S41: The second mode for 5-day averaged precipitation of subregion 7.

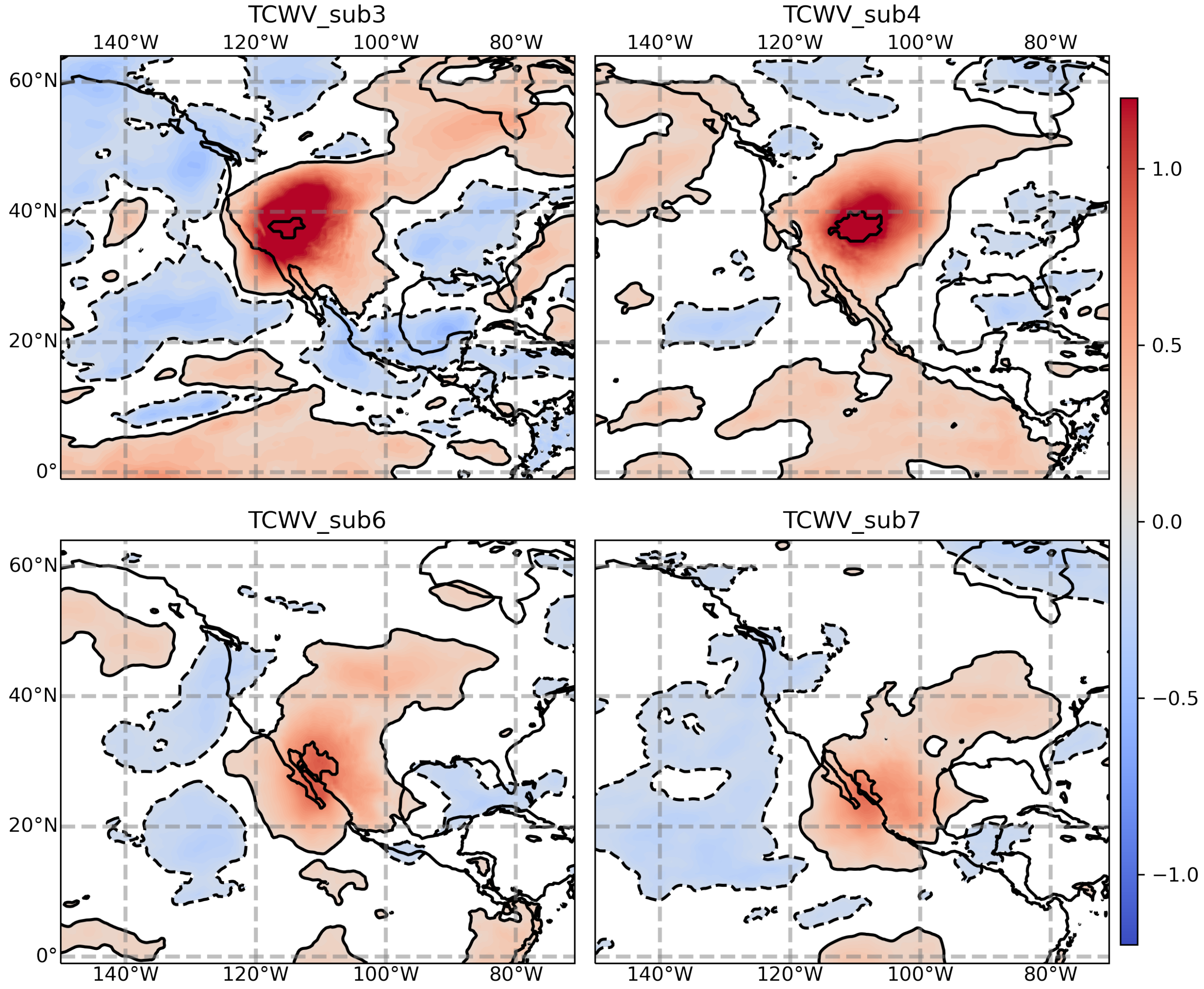

Figure S42: Total column water vapor standardized anomalies for EPEs in Sub3, Sub4, Sub6 and Sub7.

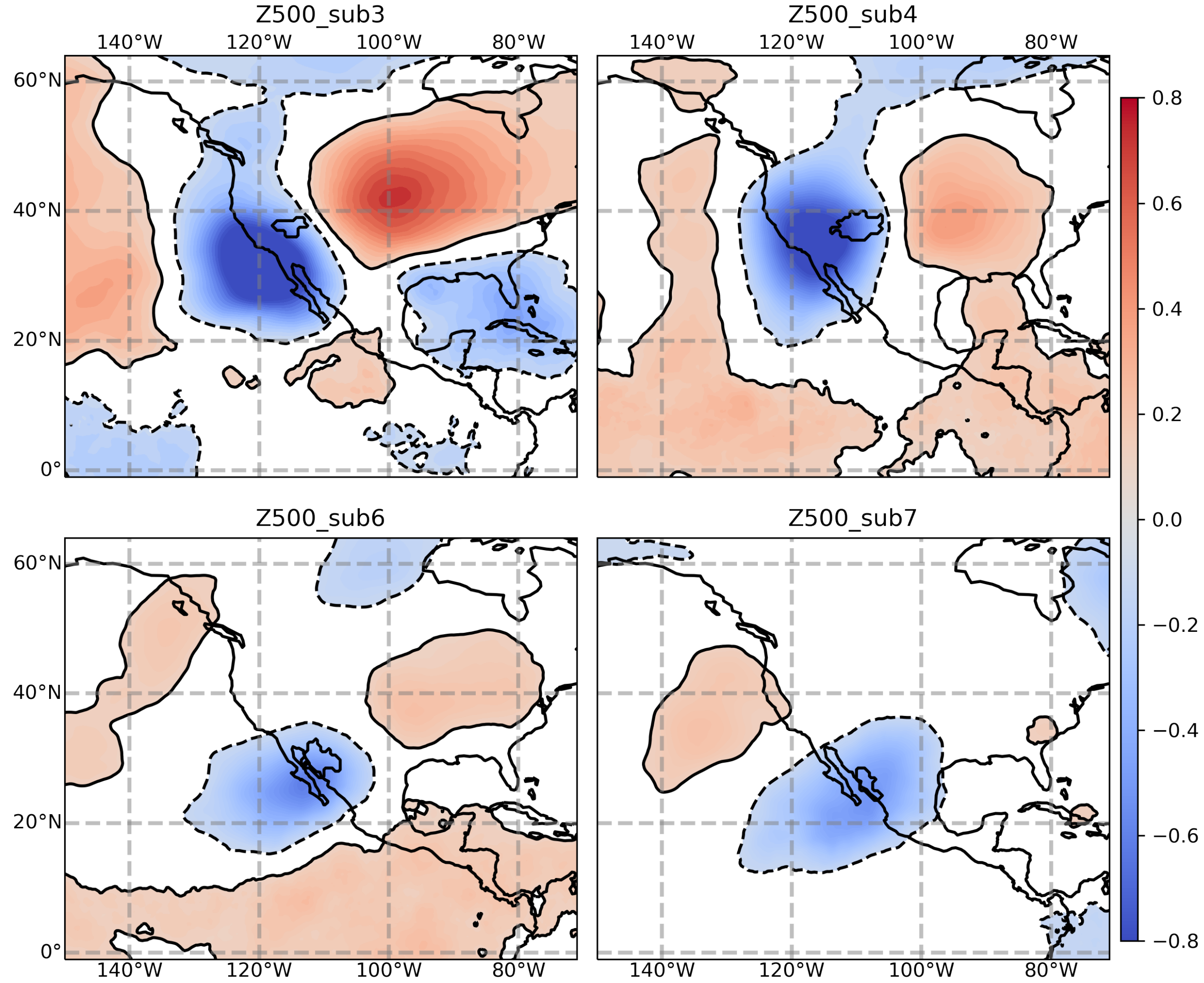

Figure S43: 500 hPa geopotential height standardized anomalies for EPEs in Sub3, Sub4, Sub6 and Sub7.

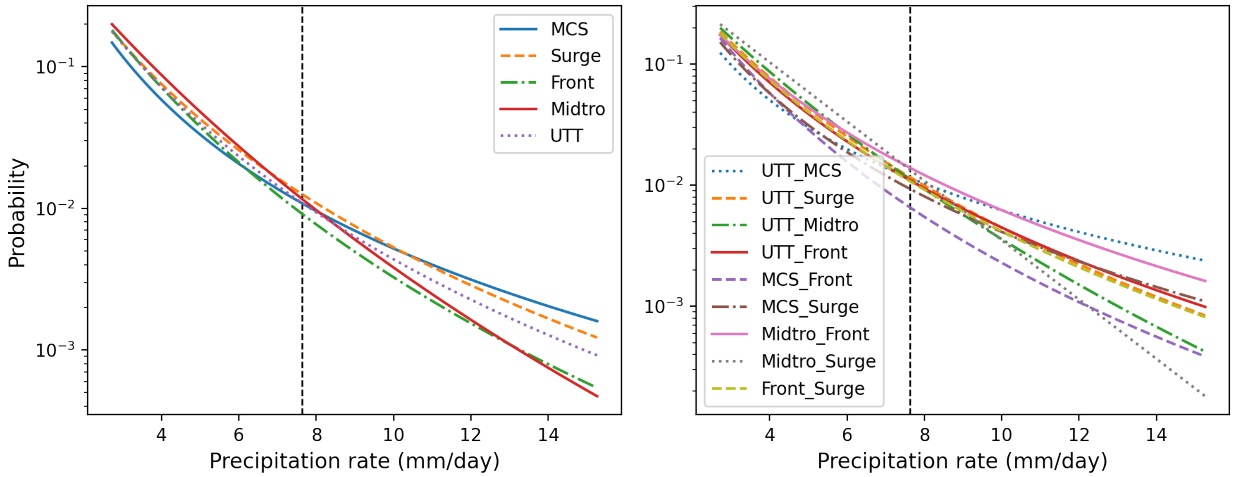

Figure S44: Precipitation rate distribution with respect to atmospheric drivers for Sub3. The dashed vertical line denotes the 95$^{\text{th}}$ percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

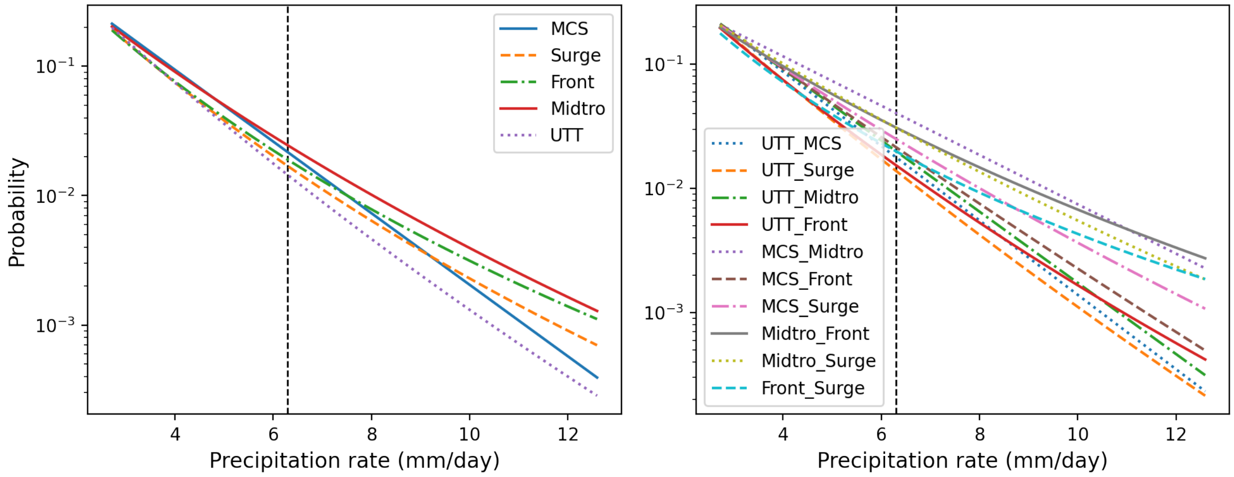

Figure S45: Precipitation rate distribution with respect to atmospheric drivers for Sub3. The dashed vertical line denotes the 95th percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

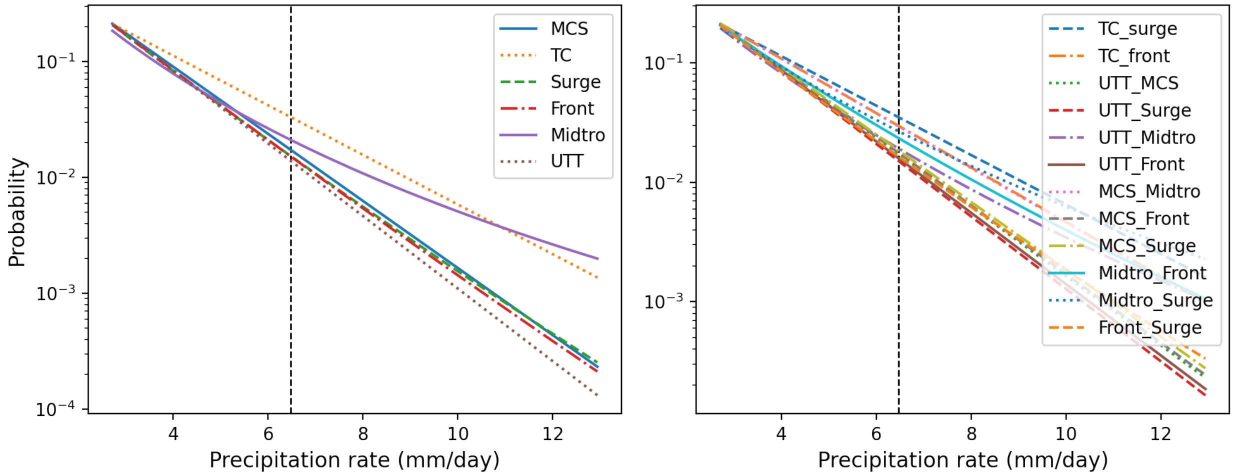

Figure S46: Precipitation rate distribution with respect to atmospheric drivers for Sub3. The dashed vertical line denotes the 95th percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

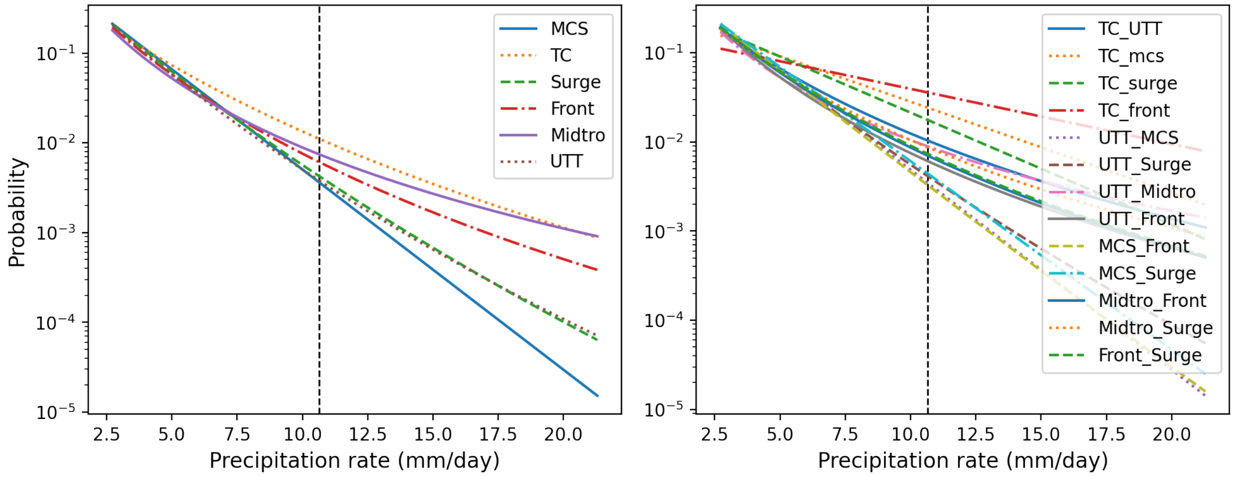

Figure S47: Precipitation rate distribution with respect to atmospheric drivers for Sub3. The dashed vertical line denotes the 95th percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

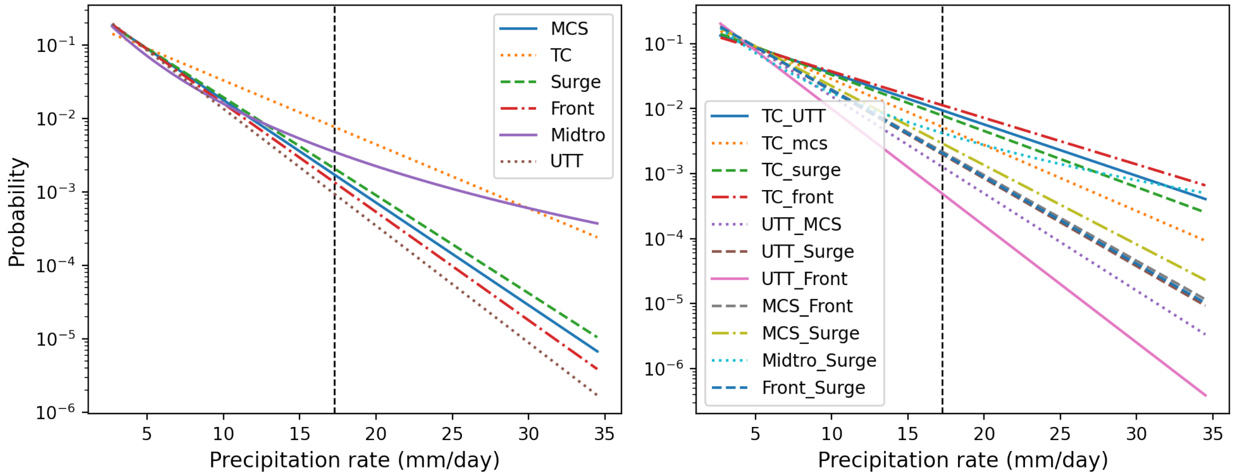

Figure S48: Precipitation rate distribution with respect to atmospheric drivers for Sub3. The dashed vertical line denotes the 95th percentile of precipitation rate. Left panel represents single drivers and the right for double drivers.

Reference

Abatzoglou, J. T.: Development of gridded surface meteorological data for ecological applications and modelling, International Journal of Climatology, 33, 121–131, 2013.

Acero, F. J., García, J. A., and Gallego, M. C.: Peaks-over-threshold study of trends in extreme rainfall over the Iberian Peninsula, Journal of Climate, 24, 1089–1105, 2011.

Adams, D. K. and Comrie, A. C.: The north American monsoon, Bulletin of the American Meteorological Society, 78, 2197–2214, 1997.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrology and Earth System Sciences, 21, 5293–5313, 2017.

Agel, L., Barlow, M., Feldstein, S. B., and Gutowski, W. J.: Identification of large-scale meteorological patterns associated with extreme precipitation in the US northeast, Climate dynamics, 50, 1819–1839, 2018.

Alaya, B., Zwiers, F., and Zhang, X.: An evaluation of block-maximum based estimation of very long return period precipitation extremes with a large ensemble climate simulation, Journal of Climate, 2020.

Anagnostopoulou, C. and Tolika, K.: Extreme precipitation in Europe: statistical threshold selection based on climatological criteria, Theoretical and Applied Climatology, 107, 479–489, 2012.

Atieh, M., Taylor, G., Sattar, A. M., and Gharabaghi, B.: Prediction of flow duration curves for ungauged basins, Journal of hydrology, 545, 383–394, 2017.

Bai, S., Kolter, J. Z., and Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, arXiv preprint arXiv:1803.01271, 2018.

Balestriero, R., Pesenti, J., and LeCun, Y.: Learning in High Dimension Always Amounts to Extrapolation, arXiv preprint arXiv:2110.09485, 2021.

Barlow, M., Gutowski, W. J., Gyakum, J. R., Katz, R. W., Lim, Y.-K., Schumacher, R. S., Wehner, M. F., Agel, L., Bosilovich, M., Collow, A., et al.: North American extreme precipitation events and related large-scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends, Climate Dynamics, 53, 6835–6875, 2019.

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., and Anderson, D.: Indicator patterns of forced change learned by an artificial neural network, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 195, 2020.

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J., and Kirchner, J. W.: The relative importance of different flood-generating mechanisms across Europe, Water Resources Research, 55, 4582–4593, 2019.

Bergstra, J. and Bengio, Y.: Random search for hyper-parameter optimization, Journal of machine learning research, 13, 281–305, 2012.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., and Gentine, P.: Enforcing analytic constraints in neural networks emulating physical systems, Physical Review Letters, 126, 098 302, 2021a.

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J., et al.: Climate-Invariant Machine Learning, arXiv preprint arXiv:2112.08440, 2021b.

Biard, J. C.: National Weather Service Coded Surface Bulletins, 2003- (netCDF format), https://doi.org/10.5281/zenodo.2651361, 2019.

Biard, J. C. and Kunkel, K. E.: Automated detection of weather fronts using a deep learning neural network, Advances in Statistical Climatology, Meteorology and Oceanography, 5, 147–160, 2019.

Bieda III, S. W., Castro, C. L., Mullen, S. L., Comrie, A. C., and Pytlak, E.: The relationship of transient upper-level troughs to variability of the North American monsoon system, Journal of Climate, 22, 4213–4227, 2009.

Böhner, J. and Antonić, O.: Land-surface parameters specific to topo-climatology, Developments in soil science, 33, 195–226, 2009.

Boos, W. and Pascale, S.: Mechanical forcing of the North American monsoon by orography, 2021.

Bordoni, S. and Stevens, B.: Principal component analysis of the summertime winds over the Gulf of California: A gulf surge index, Monthly weather review, 134, 3395–3414, 2006.

Breiman, L.: Bagging predictors, Machine learning, 24, 123–140, 1996.

Breña-Naranjo, J. A., Pedrozo-Acuña, A., Pozos-Estrada, O., Jiménez-López, S. A., and López-López, M. R.: The contribution of tropical cyclones to rainfall in Mexico, Physics and Chemistry of the Earth, Parts A/B/C, 83, 111–122, 2015.

Broxton, P. D., Dawson, N., and Zeng, X.: Linking snowfall and snow accumulation to generate spatial maps of SWE and snow depth, Earth and Space Science, 3, 246–256, 2016.

Catto, J., Jakob, C., Berry, G., and Nicholls, N.: Relating global precipitation to atmospheric fronts, Geophysical Research Letters, 39, 2012.

Chen, G. T.-J. and Chou, L.-F.: An investigation of cold vortices in the upper troposphere over the western North Pacific during the warm season, Monthly weather review, 122, 1436–1448, 1994.

Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., and Liu, Z.: Dynamic convolution: Attention over convolution kernels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 030–11 039, 2020.

Cook, B. I. and Seager, R.: The response of the North American Monsoon to increased greenhouse gas forcing, Journal of Geophysical Research: Atmospheres, 118, 1690–1699, 2013.

Cristea, N. C., Breckheimer, I., Raleigh, M. S., HilleRisLambers, J., and Lundquist, J. D.: An evaluation of terrain-based downscaling of fractional snow covered area data sets based on LiDAR-derived snow data and orthoimagery, Water Resources Research, 53, 6802–6820, 2017.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J., and Pasteris, P. P.: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States, International Journal of Climatology: a Journal of the Royal Meteorological Society, 28, 2031–2064, 2008.

Dargahi-Noubary, G.: On tail estimation: an improved method, Mathematical Geology, 21, 829–842, 1989.

de Carvalho, L. M. V. and Jones, C., eds.: The Monsoons and Climate Change, Springer International Publishing, https://doi.org/10.1007/978-3-319-21650-8, 2016.

Dettinger, M. D. and Diaz, H. F.: Global characteristics of stream flow seasonality and variability, Journal of hydrometeorology, 1, 289–310, 2000.

Díaz, S., Salinas-Zavala, C., and Hernández-Vázquez, S.: Variability of rainfall from tropical cyclones in northwestern México and its relation to SOI and PDO, Atmósfera, 21, 213–223, 2008.

Dominguez, C. and Magaña, V.: The role of tropical cyclones in precipitation over the tropical and subtropical North America, Frontiers in Earth Science, 6, 19, 2018.

Dominguez, F., Miguez-Macho, G., and Hu, H.: WRF with water vapor tracers: A study of moisture sources for the North American monsoon, Journal of Hydrometeorology, 17, 1915–1927, 2016.

Douglas, A. V. and Englehart, P. J.: A climatological perspective of transient synoptic features during NAME 2004, Journal of climate, 20, 1947–1954, 2007.

Duan, S., Ullrich, P., and Shu, L.: Using convolutional neural networks for streamflow projection in california, Frontiers in Water, 2, 28, 2020a.

Duan, S., Ullrich, P., and Shu, L.: California Streamflow Projection Dataset, https://doi.org/10.5281/zenodo.3823273, 2020b.

Englehart, P. J. and Douglas, A. V.: The role of eastern North Pacific tropical storms in the rainfall climatology of western Mexico, International Journal of Climatology: A Journal of the Royal Meteorological Society, 21, 1357–1370, 2001.

Farfán, L. M., D'Sa, E. J., Liu, K.-b., and Rivera-Monroy, V. H.: Tropical cyclone impacts on coastal regions: the case of the Yucatán and the Baja California Peninsulas, Mexico, Estuaries and coasts, 37, 1388–1402, 2014.

Feng, D., Fang, K., and Shen, C.: Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales, arXiv preprint arXiv:1912.08949, 2019.

Feng, Z., Leung, L. R., Liu, N., Wang, J., Houze Jr, R. A., Li, J., Hardin, J. C., Chen, D., and Guo, J.: A Global High-Resolution Mesoscale Convective System Database Using Satellite-Derived Cloud Tops, Surface Precipitation, and Tracking, Journal of Geophysical Research: Atmospheres, 126, e2020JD034 202, 2021.

Finch, Z. O. and Johnson, R. H.: Observational analysis of an upper-level inverted trough during the 2004 north american monsoon experiment, Monthly Weather Review, 138, 3540–3555, https://doi.org/10.1175/2010MWR3369.1, 2010a.

Finch, Z. O. and Johnson, R. H.: Observational analysis of an upper-level inverted trough during the 2004 North American Monsoon Experiment, Monthly weather review, 138, 3540–3555, 2010b.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., and Thompson, G.: Interpretable deep learning for spatial analysis of severe hailstorms, Monthly Weather Review, 147, 2827–2845, 2019.

Gao, C., Gemmer, M., Zeng, X., Liu, B., Su, B., and Wen, Y.: Projected streamflow in the Huaihe River Basin (2010–2100) using artificial neural network, Stochastic Environmental Research and Risk Assessment, 24, 685–697, 2010.

Gochis, D., Barlage, M., Cabell, R., Dugger, A., Fanfarillo, A., FitzGerald, K., McAllister, M., McCreight, J., RafieeiNasab, A., Read, L., Frazier, N., Johnson, D., Mattern, J. D., Karsten, L., Mills, T. J., and Fersch, B.: WRF-Hydro® v5.1.1, https://doi.org/10.5281/ZENODO.3625238, 2020.

Goswami, B. N., Krishnamurthy, V., and Annmalai, H.: A broad-scale circulation index for the interannual variability of the Indian summer monsoon, Quarterly Journal of the Royal Meteorological Society, 125, 611–633, 1999.

He, K., Zhang, X., Ren, S., and Sun, J.: Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Hewson, T. D.: Objective fronts, Meteorological Applications, 5, 37–65, 1998.

Higgins, R., Yao, Y., and Wang, X.: Influence of the North American monsoon system on the US summer precipitation regime, Journal of climate, 10, 2600–2622, 1997.

Higgins, R., Shi, W., and Hain, C.: Relationships between Gulf of California moisture surges and precipitation in the southwestern United States, Journal of Climate, 17, 2983–2997, 2004.

Higgins, W., Ahijevych, D., Amador, J., Barros, A., Berbery, E. H., Caetano, E., Carbone, R., Ciesielski, P., Cifelli, R., Cortez-Vazquez, M., et al.: The NAME 2004 field campaign and modeling strategy, Bulletin of the American Meteorological Society, 87, 79–94, 2006.

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., Hochreiter, S., and Klambauer, G.: MC-LSTM: Mass-Conserving LSTM, arXiv preprint arXiv:2101.05186, 2021.

Hornik, K., Stinchcombe, M., and White, H.: Multilayer feedforward networks are universal approximators, Neural networks, 2, 359–366, 1989.

Houssos, E., Lolis, C., and Bartzokas, A.: Atmospheric circulation patterns associated with extreme precipitation amounts in Greece, Advances in Geosciences, 17, 5–11, 2008.

Huang, S., Chang, J., Huang, Q., and Chen, Y.: Monthly streamflow prediction using modified EMD-based support vector machine, Journal of Hydrology, 511, 764–775, 2014.

Huang, X. and Ullrich, P. A.: The changing character of twenty-first-century precipitation over the western United States in the variable-resolution CESM, Journal of Climate, 30, 7555–7575, 2017.

Hung, C.-W. and Yanai, M.: Factors contributing to the onset of the Australian summer monsoon, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 130, 739–758, 2004.

Igel, M. R., Ullrich, P. A., and Boos, W. R.: Upper-tropospheric troughs and North American monsoon rainfall in a long-term track dataset, Journal of Geophysical Research: Atmospheres, p. e2021JD034541, 2021.

Jiang, H. and Zipser, E. J.: Contribution of tropical cyclones to the global precipitation from eight seasons of TRMM data: Regional, seasonal, and interannual variations, Journal of climate, 23, 1526–1543, 2010.

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S., Azizzadenesheli, K., Wang, R., Chattopadhyay, A., Singh, A., et al.: Physics-informed machine learning: case studies for weather and climate modelling, Philosophical Transactions of the Royal Society A, 379, 20200 093, 2021.

Kelley, W. E. and Mock, D. R.: A diagnostic study of upper tropospheric cold lows over the western North Pacific, Monthly Weather Review, 110, 471–480, 1982.

King, F., Erler, A. R., Frey, S. K., and Fletcher, C. G.: Application of machine learning techniques for regional bias correction of snow water equivalent estimates in Ontario, Canada, Hydrology and Earth System Sciences, 24, 4887–4902, 2020.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kirkham, J. D., Koch, I., Saloranta, T. M., Litt, M., Stigter, E. E., Møen, K., Thapa, A., Melvold, K., and Immerzeel, W. W.: Near real-time measurement of snow water equivalent in the Nepal Himalayas, Frontiers in Earth Science, 7, 177, 2019.

Kisi, O. and Cimen, M.: A wavelet-support vector machine conjunction model for monthly streamflow forecasting, Journal of Hydrology, 399, 132–140, 2011.

Klotzbach, P. J., Wood, K. M., Schreck III, C. J., Bowen, S. G., Patricola, C. M., and Bell, M. M.: Trends in Global Tropical Cyclone Activity: 1990–2021, Geophysical Research Letters, 49, e2021GL095 774, 2022.

Knapp, K. R., Kruk, M. C., Levinson, D. H., Diamond, H. J., and Neumann, C. J.: The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data, Bulletin of the American Meteorological Society, 91, 363–376, 2010.

Koirala, S., Hirabayashi, Y., Mahendran, R., and Kanae, S.: Global assessment of agreement among streamflow projections using CMIP5 model outputs, Environmental Research Letters, 9, 064 017, 2014.

Konapala, G., Kao, S.-C., Painter, S. L., and Lu, D.: Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US, Environmental Research Letters, 15, 104 022, 2020.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Benchmarking a catchment-aware Long Short-Term Memory Network (LSTM) for large-scale hydrological modeling, arXiv preprint arXiv:1907.08456, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, 2019b.

Krzywinski, M. and Altman, N.: Visualizing samples with box plots: use box plots to illustrate the spread and differences of samples, Nature Methods, 11, 119–121, 2014.

Kunkel, K. E., Easterling, D. R., Kristovich, D. A., Gleason, B., Stoecker, L., and Smith, R.: Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States, Journal of Hydrometeorology, 13, 1131–1141, 2012.

Le, X.-H., Ho, H. V., Lee, G., and Jung, S.: Application of long short-term memory (LSTM) neural network for flood forecasting, Water, 11, 1387, 2019.

Lee, J.-Y. and Wang, B.: Future change of global monsoon in the CMIP5, Climate Dynamics, 42, 101–119, 2014.

Li, D., Wrzesien, M. L., Durand, M., Adam, J., and Lettenmaier, D. P.: How much runoff originates as snow in the western United States, and how will that change in the future?, Geophysical Research Letters, 44, 6163–6172, 2017.

Liang, X., Lettenmaier, D. P., Wood, E. F., and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, Journal of Geophysical Research: Atmospheres, 99, 14 415–14 428, 1994.

Lin, T., Wang, Y., Liu, X., and Qiu, X.: A Survey of Transformers, arXiv preprint arXiv:2106.04554, 2021.

Liu, F., Chai, J., Wang, B., Liu, J., Zhang, X., and Wang, Z.: Global monsoon precipitation responses to large volcanic eruptions, Scientific reports, 6, 1–11, 2016.

Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., Cayan, D. R., and Brekke, L.: A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada 1950–2013, Scientific data, 2, 1–12, 2015.

Manepalli, A., Albert, A., Rhoades, A., Feldman, D., and Jones, A. D.: Emulating numeric hydroclimate models with physics-informed cGANs, in: AGU Fall Meeting 2019, AGU, 2019.

Martinez-Villalobos, C. and Neelin, J. D.: Why do precipitation intensities tend to follow gamma distributions?, Journal of the Atmospheric Sciences, 76, 3611–3631, 2019.

Martius, O., Zenklusen, E., Schwierz, C., and Davies, H. C.: Episodes of Alpine heavy precipitation with an overlying elongated stratospheric intrusion: A climatology, International Journal of Climatology: A Journal of the Royal Meteorological Society, 26, 1149–1164, 2006.

McCrary, R., Mearns, L., Hughes, M., Biner, S., and Bukovsky, M.: Projections of North American snow from NA-CORDEX and their uncertainties, with a focus on model resolution, Climatic Change, 170, 1–25, 2022.

McCrary, R. R. and Mearns, L. O.: Quantifying and diagnosing sources of uncertainty in midcentury changes in North American snowpack from NARCCAP, Journal of Hydrometeorology, 20, 2229–2252, 2019.

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the black box more transparent: Understanding the physical implications of machine learning, Bulletin of the American Meteorological Society, 100, 2175–2199, 2019.

Mejia, J. F., Douglas, M. W., and Lamb, P. J.: Observational investigation of relationships between moisture surges and mesoscale-to large-scale convection during the North American monsoon, International Journal of Climatology, 36, 2555–2569, 2016.

Meyal, A. Y., Versteeg, R., Alper, E., Johnson, D., Rodzianko, A., Franklin, M., and Wainwright, H.: Automated cloud based Long Short-Term Memory neural network based SWE prediction, Frontiers in Water, 2, 2020.

Mohtadi, M., Prange, M., and Steinke, S.: Palaeoclimatic insights into forcing and response of monsoon rainfall, Nature, 533, 191–199, 2016.

Moore, B. J., Keyser, D., and Bosart, L. F.: Linkages between extreme precipitation events in the central and eastern United States and Rossby wave breaking, Monthly Weather Review, 147, 3327–3349, 2019.

Murakami, H., Delworth, T. L., Cooke, W. F., Zhao, M., Xiang, B., and Hsu, P.-C.: Detected climatic change in global distribution of tropical cyclones, Proceedings of the National Academy of Sciences, 117, 10 706–10 714, 2020.

Myhre, G., Alterskjær, K., Stjern, C. W., Hodnebrog, Ø., Marelle, L., Samset, B. H., Sillmann, J., Schaller, N., Fischer, E., Schulz, M., et al.: Frequency of extreme precipitation increases extensively with event rareness under global warming, Scientific reports, 9, 1–10, 2019.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I—A discussion of principles, Journal of hydrology, 10, 282–290, 1970.

Newman, A. and Johnson, R. H.: Mechanisms for precipitation enhancement in a North American Monsoon upper-tropospheric trough, Journal of the Atmospheric Sciences, 69, 1775–1792, 2012.

Newman, A., Sampson, K., Clark, M., Bock, A., Viger, R., and Blodgett, D.: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA, Boulder, CO, UCAR/NCAR, 2014.

Noori, N. and Kalin, L.: Coupling SWAT and ANN models for enhanced daily streamflow prediction, Journal of Hydrology, 533, 141–151, 2016.

Ntokas, K. F., Odry, J., Boucher, M.-A., and Garnaud, C.: Investigating ANN architectures and training to estimate snow water equivalent from snow depth, Hydrology and Earth System Sciences, 25, 3017–3040, 2021.

Odry, J., Boucher, M., Cantet, P., Lachance-Cloutier, S., Turcotte, R., and St-Louis, P.: Using artificial neural networks to estimate snow water equivalent from snow depth, Canadian Water Resources Journal/Revue canadienne des ressources hydriques, 45, 252–268, 2020.

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., Marks, D., Mattmann, C., McGurk, B., Ramirez, P., Richardson, M., and Skiles, S. M.: ASO L4 Lidar Snow Water Equivalent 50m UTM Grid, Version 1, https://doi.org/10.5067/M4TUH28NHL4Z, 2018.

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., Tian, Y., and Ma, H.-Y.: Learning to correct climate projection biases, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 509, 2021.

Papin, P. P., Bosart, L. F., and Torn, R. D.: A feature-based approach to classifying summertime potential vorticity streamers linked to Rossby wave breaking in the North Atlantic basin, Journal of Climate, 33, 5953–5969, 2020.

Parfitt, R., Czaja, A., and Seo, H.: A simple diagnostic for the detection of atmospheric fronts, Geophysical Research Letters, 44, 4351–4358, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, edited by Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., pp. 8024–8035, Curran Associates, Inc., URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Pazos, M. and Mendoza, B.: Landfalling tropical cyclones along the Eastern Pacific coast between the sixteenth and twentieth centuries, Journal of climate, 26, 4219–4230, 2013.

Peng, T., Zhou, J., Zhang, C., and Fu, W.: Streamflow forecasting using empirical wavelet transform and artificial neural networks, Water, 9, 406, 2017.

Perneger, T. V.: What's wrong with Bonferroni adjustments, Bmj, 316, 1236–1238, 1998.

Pierce, D. W., Cayan, D. R., and Thrasher, B. L.: Statistical downscaling using localized constructed analogs (LOCA), Journal of Hydrometeorology, 15, 2558–2585, 2014.

Pierce, D. W., Kalansky, J. F., and Cayan, D. R.: Climate, drought, and sea level rise scenarios for

California's fourth climate change assessment, Tech. rep., Technical Report CCCA4-CEC-2018-006, California Energy Commission, 2018.

Ramage, C. S.: Monsoon meteorology, Tech. rep., 1971.

Rasmussen, R., Ikeda, K., Liu, C., Gochis, D., Clark, M., Dai, A., Gutmann, E., Dudhia, J., Chen, F., Barlage, M., et al.: Climate change impacts on the water balance of the Colorado headwaters: high-resolution regional climate model simulations, Journal of Hydrometeorology, 15, 1091–1116, 2014.

Rasouli, K., Hsieh, W. W., and Cannon, A. J.: Daily streamflow forecasting by machine learning methods with weather and climate inputs, Journal of Hydrology, 414, 284–293, 2012.

Rhoades, A. M., Jones, A. D., and Ullrich, P. A.: Assessing mountains as natural reservoirs with a multimetric framework, Earth's Future, 6, 1221–1241, 2018a.

Rhoades, A. M., Jones, A. D., and Ullrich, P. A.: The Changing Character of the California Sierra Nevada as a Natural Reservoir, Geophysical Research Letters, 45, 13,008–13,019, https://doi.org/https://doi.org/10.1029/2018GL080308, 2018b.

Rhoades, A. M., Ullrich, P. A., and Zarzycki, C. M.: Projecting 21st century snowpack trends in western USA mountains using variable-resolution CESM, Climate Dynamics, 50, 261–288, 2018c.

Roberts, D. W. and Cooper, S. V.: Concepts and techniques of vegetation mapping, Land classifications based on vegetation: applications for resource management, pp. 90–96, 1989.

Ryoo, J.-M., Kaspi, Y., Waugh, D. W., Kiladis, G. N., Waliser, D. E., Fetzer, E. J., and Kim, J.: Impact of Rossby wave breaking on US West Coast winter precipitation during ENSO events, Journal of Climate, 26, 6360–6382, 2013.

Seastrand, S., Serra, Y., Castro, C., and Ritchie, E.: The dominant synoptic-scale modes of North American monsoon precipitation, International Journal of Climatology, 35, 2019–2032, 2015.

Shanker, M., Hu, M. Y., and Hung, M. S.: Effect of data standardization on neural network training, Omega, 24, 385–397, 1996.

Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists, Water Resources Research, 54, 8558–8593, 2018.

Sierks, M. D., Kalansky, J., Cannon, F., and Ralph, F.: Characteristics, origins, and impacts of summertime extreme precipitation in the Lake Mead Watershed, Journal of Climate, 33, 2663–2680, 2020.

Siirila-Woodburn, E. R., Rhoades, A. M., Hatchett, B. J., Huning, L. S., Szinai, J., Tague, C., Nico, P. S., Feldman, D. R., Jones, A. D., Collins, W. D., et al.: A low-to-no snow future and its impacts on water resources in the western United States, Nature Reviews Earth & Environment, 2, 800–819, 2021.

Snauffer, A. M., Hsieh, W. W., Cannon, A. J., and Schnorbus, M. A.: Improving gridded snow water equivalent products in British Columbia, Canada: multi-source data fusion by neural network models, The Cryosphere, 12, 891–905, 2018.

Stidd, C.: Cube-root-normal precipitation distributions, Eos, Transactions American Geophysical Union, 34, 31–35, 1953.

Sukhdeo, R., Ullrich, P. A., and Grotjahn, R.: Assessing the large-scale drivers of precipitation in the northeastern United States via linear orthogonal decomposition, Climate Dynamics, pp. 1–25, 2022.

Swain, D. L., Langenbrunner, B., Neelin, J. D., and Hall, A.: Increasing precipitation volatility in twenty-first-century California, Nature Climate Change, 8, 427–433, 2018.

Swenson, L. M. and Grotjahn, R.: Using Self-Organizing Maps to Identify Coherent CONUS Precipitation Regions, Journal of Climate, 32, 7747–7761, 2019.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D.: Efficient transformers: A survey, arXiv preprint arXiv:2009.06732, 2020.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically interpretable neural networks for the geosciences: Applications to earth system variability, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 002, 2020.

Tsymbal, A.: The problem of concept drift: definitions and related work, Computer Science Department, Trinity College Dublin, 106, 58, 2004.

Turrent, C. and Cavazos, T.: Role of the land-sea thermal contrast in the interannual modulation of the North American Monsoon, Geophysical Research Letters, 36, 2009.

Ullrich, P. A. and Taylor, M. A.: Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part I, Monthly Weather Review, 143, 2419–2440, 2015.

Ullrich, P. A., Devendran, D., and Johansen, H.: Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part II, Monthly Weather Review, 144, 1529–1549, 2016.

Ullrich, P. A., Xu, Z., Rhoades, A. M., Dettinger, M. D., Mount, J. F., Jones, A. D., and Vahmani, P.: California's drought of the future: A midcentury recreation of the exceptional conditions of 2012–2017, Earth's future, 6, 1568–1587, 2018.

Varuolo-Clarke, A. M., Reed, K. A., and Medeiros, B.: Characterizing the North American monsoon in the Community Atmosphere Model: Sensitivity to resolution and topography, Journal of Climate, 32, 8355–8372, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, Advances in neural information processing systems, 30, 2017.

Vera, C., Higgins, W., Amador, J., Ambrizzi, T., Garreaud, R., Gochis, D., Gutzler, D., Lettenmaier, D., Marengo, J., Mechoso, C., et al.: Toward a unified view of the American monsoon systems, Journal of climate, 19, 4977–5000, 2006.

Wang, B. and Fan, Z.: Choice of South Asian summer monsoon indices, Bulletin of the American Meteorological Society, 80, 629–638, 1999.

Wang, B., Li, J., Cane, M. A., Liu, J., Webster, P. J., Xiang, B., Kim, H.-M., Cao, J., and Ha, K.-J.: Toward predicting changes in the land monsoon rainfall a decade in advance, Journal of Climate, 31, 2699–2714, 2018.

Wang, F., Shao, W., Yu, H., Kan, G., He, X., Zhang, D., Ren, M., and Wang, G.: Re-evaluation of the power of the mann-kendall test for detecting monotonic trends in hydrometeorological time series, Frontiers in Earth Science, p. 14, 2020a.

Wang, X., Kondratyuk, D., Christiansen, E., Kitani, K. M., Alon, Y., and Eban, E.: Wisdom of Committees: An Overlooked Approach To Faster and More Accurate Models, 2021.

Wang, Z., Zhang, G., Dunkerton, T. J., and Jin, F.-F.: Summertime stationary waves integrate tropical and extratropical impacts on tropical cyclone activity, Proceedings of the National Academy of Sciences, 117, 22 720–22 726, 2020b.

Watterson, I. and Dix, M.: Simulated changes due to global warming in daily precipitation means and extremes and their interpretation using the gamma distribution, Journal of Geophysical Research: Atmospheres, 108, 2003.

Webster, P. J. and Yang, S.: Monsoon and ENSO: Selectively interactive systems, Quarterly Journal of the Royal Meteorological Society, 118, 877–926, 1992.

Wehner, M., Risser, M., Ullrich, P., and Duan, S.: Exploring variability in seasonal average and extreme precipitation using unsupervised machine learning., Tech. rep., Artificial Intelligence for Earth System Predictability (AI4ESP . . . , 2021.

White, A. B., Moore, B. J., Gottas, D. J., and Neiman, P. J.: Winter storm conditions leading to excessive runoff above California's Oroville Dam during January and February 2017, Bulletin of the American Meteorological Society, 100, 55–70, 2019.

Wibig, J.: Precipitation in Europe in relation to circulation patterns at the 500 hPa level, International Journal of Climatology: A Journal of the Royal Meteorological Society, 19, 253–269, 1999.

Wrzesien, M. L. and Pavelsky, T. M.: Projected changes to extreme runoff and precipitation events from a downscaled simulation over the western United States, Frontiers in Earth Science, p. 355, 2020.

Xie, P., Chen, M., and Shi, W.: CPC unified gauge-based analysis of global daily precipitation, in: Preprints, 24th Conf. on Hydrology, Atlanta, GA, Amer. Meteor. Soc, vol. 2, 2010.

Xu, G., Ren, T., Chen, Y., and Che, W.: A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis, Frontiers in Neuroscience, 14, 1253, 2020.

Yan, L., Feng, J., and Hang, T.: Small watershed stream-flow forecasting based on LSTM, in: International Conference on Ubiquitous Information Management and Communication, pp. 1006–1014, Springer, 2019.

Yaseen, Z. M., Jaafar, O., Deo, R. C., Kisi, O., Adamowski, J., Quilty, J., and El-Shafie, A.: Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq, Journal of Hydrology, 542, 603–614, 2016.

Zavadoff, B. L. and Kirtman, B. P.: North Atlantic summertime anticyclonic Rossby wave breaking: Climatology, impacts, and connections to the Pacific decadal oscillation, Journal of Climate, 32, 485–500, 2019.

Zeng, X., Broxton, P., and Dawson, N.: Snowpack change from 1982 to 2016 over conterminous United States, Geophysical Research Letters, 45, 12–940, 2018.

Zhang, G., Wang, Z., Dunkerton, T. J., Peng, M. S., and Magnusdottir, G.: Extratropical impacts on Atlantic tropical cyclone activity, Journal of the Atmospheric Sciences, 73, 1401–1418, 2016.

Zhang, G., Wang, Z., Peng, M. S., and Magnusdottir, G.: Characteristics and impacts of extratropical Rossby wave breaking during the Atlantic hurricane season, Journal of Climate, 30, 2363–2379, 2017.

Zhao, M.: A Study of AR-, TS-, and MCS-Associated Precipitation and Extreme Precipitation in Present and Warmer Climates, Journal of Climate, 35, 479–497, 2022.