THESIS

ASSESSMENT OF NUMERICAL WEATHER PREDICTION MODEL RE-FORECASTS OF ATMOSPHERIC RIVERS ALONG THE WEST COAST OF NORTH AMERICA

Submitted by

Kyle M. Nardi

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2018

Master's Committee:

Advisor: Elizabeth A. Barnes

Russ S. Schumacher
Jay M. Ham

ABSTRACT

ASSESSMENT OF NUMERICAL WEATHER PREDICTION MODEL RE-FORECASTS OF ATMOSPHERIC RIVERS ALONG THE WEST COAST OF NORTH AMERICA

Atmospheric rivers (ARs) - narrow corridors of high atmospheric water vapor transport - occur globally and are associated with flooding and maintenance of the regional water supply. Therefore, it is important to improve forecasts of AR occurrence and characteristics. Although prior work has examined the skill of numerical weather prediction (NWP) models in forecasting ARs, these studies only cover several years of re-forecasts from a handful of models. Here, we expand this previous work and assess the performance of 10-30 years of wintertime (November-February) AR landfall re-forecasts from nine operational weather models, obtained from the International Subseasonal to Seasonal (S2S) Project Database. Model errors along the West Coast of North America at leads of 1-14 days are examined in terms of AR occurrence, intensity, and landfall location. We demonstrate that re-forecast performance varies across models, lead times, and geographical regions. Occurrence-based skill approaches that of climatology at 14 days, while models are, on average, more skillful at shorter leads in California, Oregon, and Washington compared to British Columbia and Alaska. We also find that the average magnitude of landfall Integrated Water Vapor Transport (IVT) error stays fairly constant across lead times, although over-prediction of IVT is more common at later lead times. We then show that northward landfall location errors are favored in California, Oregon, and Washington, although southward errors occur more often than expected from climatology. We next explore the link between the predictability of ARs at 1-14 days and synoptic-scale weather conditions by examining re-forecasts of 500-hPa geopotential height anomaly patterns conducive to landfalling ARs. Finally, the potential for skillful forecasts of IVT and precipitation at subseasonal to seasonal (S2S) leads is explored using an empirical forecast model based on the Madden-Julian oscillation (MJO) and the quasi-biennial oscillation (QBO).

Overall, these results highlight the need for model improvements at 1-14 days, while helping to identify factors that cause model errors as well as sources of additional predictability.

ACKNOWLEDGEMENTS

I would like to thank Professor Elizabeth Barnes for her guidance in the completion of this thesis. I also thank Professor Russ Schumacher and Professor Jay Ham for their roles on my graduate committee; Colorado State colleagues Dr. Cory Baggett, Dr. Bryan Mundhenk, Marie McGraw, and others for their assistance at various stages of this research; Ammon Redman, Matt Bishop, and Mostafa Elkady for their technical support; and the staff at the Center for Western Weather and Water Extremes for their frequent collaboration. This research was supported by a sub-award under the Forecast Informed Reservoir Operations (FIRO) project at the Center for Western Weather and Water Extremes (CW3E). The FIRO project is in partnership with local and federal agencies such as the Sonoma County Water Agency and the US Army Corps of Engineers. All model data used in this study come from ECMWF through the Subseasonal to Seasonal (S2S) International Project. Reanalysis data (ERA-Interim) also come from ECMWF. The empirical prediction model comes from Dr. Bryan Mundhenk, who developed the model while at Colorado State University.

This thesis is typeset using a document class provided by Colorado State University.

DEDICATION

I would like to dedicate this thesis to my family and friends who have supported me through the years.

TABLE OF CONTENTS

Chapter 1

Introduction

Atmospheric rivers (ARs) are narrow plumes of high water vapor transport in the atmosphere that account for a significant portion of total meridional moisture transport in the midlatitudes (Zhu and Newell, 1994, 1998). These filamentary structures in the lower troposphere are associated with transient extratropical storm systems, often appearing in the vicinity of the warm conveyor belt/low-level jet stream ahead of a cold front (e.g. Bao et al., 2006; Ralph et al., 2004, 2017), and can be formed from both local convergence of water vapor and direct transport from the tropics (e.g. Bao et al., 2006; Dacre et al., 2015).

Upon making landfall, ARs interact with local topography, often producing enhanced upslope precipitation (e.g. Neiman et al., 2002; Ralph and Dettinger, 2012). Due to their ability to produce high precipitation totals, intense ARs bring a significant risk of flooding. For example, Neiman et al. (2008) and Ralph et al. (2006) found a strong relationship between observed ARs and recorded flooding events along California's Russian River, while Ralph et al. (2010) found that most of the extreme precipitation observations in California, Oregon, and Washington during the 2005/2006 cool season coincided with AR conditions. ARs are also crucial for the maintenance of the water supply, especially in the western United States. Dettinger et al. (2011) found that ARs account for 20-50% of the precipitation and streamflow in the state of California, while Guan et al. (2010) found that AR events, on average, generated about four times the daily snow water equivalent accumulation compared to non-AR events over the Sierra Nevada from 2004-2010. The high precipitation totals from ARs are often coupled with strong surface winds, and Waliser and Guan (2017) showed that landfalling ARs were coincident with approximately half of the extreme wind events recorded along the West Coast of North America between 1997 and 2014.

Many prior studies have focused on AR activity in California, where landfalling ARs account for a large portion of extreme precipitation events (e.g. Neiman et al., 2008; Ralph et al., 2006, 2010). However, ARs also impact locations throughout western North America. For example,

Neiman et al. (2011) found that 46 of 48 recent annual peak daily streamflows in western Washington coincided with landfalling ARs. Farther to the north, Lavers et al. (2014) revealed a link between landfalling ARs and significant flooding events in coastal British Columbia and Alaska in 2010 and 2012. Away from the coast, recent studies (e.g. Ralph and Galarneau, 2017; Rivera et al., 2014) highlighted a connection between ARs/easterly water vapor transport and extreme precipitation in Arizona, and Rutz and Steenburgh (2014) found that inland-penetrating ARs reach portions of the Intermountain West such as Idaho, Nevada, and Utah. In addition, Hatchett et al. (2017) found that between 25-65% of avalanche fatalities at various locations in the western United States coincided with AR conditions.

Since ARs are high-impact phenomena that impact much of western North America, it is important to accurately forecast their occurrence and characteristics. Prior model verification studies have examined the skill of numerical weather prediction (NWP) models in forecasting ARs. For example, Wick et al. (2013) examined the skill of AR re-forecasts from five dynamical models along the West Coast of the United States during three cool seasons from 2008-2009 through 2010-2011. These re-forecasts came from the The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble (TIGGE) dataset. The re-forecasts were then compared to satellite-derived fields of atmospheric water vapor from the Special Sensor Microwave Imager (SSM/I). In their study, Wick et al. (2013) demonstrated a general decrease of approximately 20-30% in model skill from initialization to 10 days in re-forecasting the occurrence of landfalling ARs in the western United States. Correlations between re-forecast and observed water vapor fields within the domain decreased from about 0.9 at initialization to about 0.6 at 10 days. In addition, Nayak et al. (2014) compared the AR re-forecast skill of models from the TIGGE dataset over the central United States for the time period from 2007 to 2013 to fields of water vapor transport from the National Aeronautics and Space Administration Modern-Era Retrospective Analysis for Research and Applications (MERRA) dataset. They showed that models have symmetric extremal dependence index (SEDI) skill scores that approach 0 at leads greater than 10 days, implying that models provide little additional forecast skill for ARs in the central United States at

such leads. Most recently, DeFlorio et al. (2018) examined European Centre for Medium-Range Weather Forecasts (ECMWF) AR re-forecasts at leads of 1 to 14 days and found a similar decrease in prediction skill past 10 days over the North Pacific and western United States.

In light of demonstrated model errors at short and medium-range leads, recent studies have also examined the link between AR activity along the West Coast of North America and synoptic-scale weather patterns across the North Pacific Ocean, which are important for the characteristics of landfalling ARs along the West Coast of North America (e.g. Hecht and Cordeira, 2017). Mundhenk et al. (2016b) found a clear modulation in AR frequency between Alaska and California based on the geopotential height field across the North Pacific. Specifically, they found that the presence of an anomalous blocking ridge at 500 hPa across the Northeast Pacific is favorable for increased AR activity in Alaska. However, prior work has shown that dynamical models, though improving, continue to struggle to predict the location and duration of atmospheric blocking events (e.g. D'Andrea and Coauthors, 1998; Davini and D'Andrea, 2016; Matsueda et al., 2011; Palmer et al., 2008). Therefore, synoptic-scale features such as geopotential height anomalies over the North Pacific have the potential to influence AR landfall forecasts along the West Coast of North America.

Meanwhile, recent efforts have shown the potential to extend AR predictive skill beyond 14 days into subseasonal to seasonal (S2S) leads (i.e. about 2 weeks to 2 months) by using climate teleconnection patterns as predictors. For example, Zhou and Kim (2017) examined AR re-forecasts over the North Pacific (from the North American Multi-Model Ensemble (NMME) dataset from 1981 through 2012) through the lens of the El Niño Southern Oscillation (ENSO), a quasi-periodic oscillation of significant sea surface temperature anomalies over the equatorial Pacific with a typical period of several years (e.g. Trenberth, 1997). They found improvements in predictions of seasonal AR frequency over the northeast Pacific during ENSO winters (i.e. winters with significant sea surface temperature anomalies). However, predictions of landfall frequencies along the West Coast of North America during ENSO winters were found to have less skill compared to predictions over the ocean. DeFlorio et al. (2018) compared global ECMWF en-

semble re-forecasts to European Reanalysis (ERA)-Interim (ERAI) at leads of 1 to 14 days from 1996 through 2013. They found that AR prediction skill was 15-20 % higher during boreal winter compared to boreal summer. They further showed that forecast skill was increased over the North Pacific and western United States during positive ENSO and Pacific-North America (PNA) teleconnection phases.

Recent work has identified the Madden-Julian oscillation (MJO) and quasi-biennial oscillation (QBO) as other potential sources of predictability at S2S leads for AR activity along the West Coast of North America. The MJO is an intraseasonal (approximately 30-90 days) oscillation in the pattern of atmospheric circulation and deep convection over equatorial regions (e.g. Zhang, 2005, 2013). The MJO has been shown to be a forcing mechanism for geopotential height patterns over the extratropical Pacific Ocean (e.g. Henderson et al., 2016; Tseng et al., 2017). In addition, both Guan and Waliser (2015) and Mundhenk et al. (2016a) demonstrated an associated modulation in AR frequency over the extratropical Pacific based on the MJO. By contrast, the QBO is defined as a downward propagation of easterly and westerly wind regimes in the equatorial stratosphere with a longer oscillation period of about 2-3 years (e.g. Baldwin and Coauthors, 2001). Recent studies have shown that the QBO modulates the MJO, with MJO activity higher during the easterly phase of the QBO due to influences on static stability and vertical wind shear (e.g. Hendon and Abhik, 2018; Marshall et al., 2017; Son et al., 2017; Yoo and Son, 2016; Zhang and Zhang, 2018). Therefore, the MJO and QBO have recently been used in tandem to assess predictability of AR activity (based on detected AR feature counts) along the West Coast of North America. For example, Baggett et al. (2017) demonstrated the potential to use the MJO and QBO for subseasonal to seasonal (S2S) forecasts of anomalous AR activity. To this same end, Mundhenk et al. (2018) studied the skill of an empirical prediction scheme, based on the MJO and QBO, in re-forecasting anomalous weekly AR activity along the West Coast of North America. The empirical model was compared to re-forecasts from ECMWF, and though ECMWF provided little additional skill at leads greater than 18 days, the empirical model was skillful in predicting anomalous AR activity at leads beyond 3 weeks.

Here, we expand on prior model verification studies by first examining re-forecasts (also known as hindcasts) of landfalling wintertime ARs from nine state-of-the-art weather models at leads of 1-14 days. Re-forecasts are retrospective forecasts from a particular numerical model for dates in the past. Re-forecasts allow for comparisons to reanalysis in order to assess model performance (Hamill et al., 2006). Re-forecast time periods range from about 10-30 years per model, with the total number of initializations ranging from about 200-2500 per model. This amounts to over 8000 initializations for analysis. The number of models and years covered by our study exceeds that of prior AR model verification studies, and our study utilizes present-day NWP models. We highlight three important components of an effective AR re-forecast (AR occurrence, AR intensity, and AR landfall location) and quantify each model's skill. The performance of the models with respect to these three components is analyzed at varying lead times for the West Coast of North America. In an attempt to link AR-related model errors to errors in predicting the synoptic-scale flow, we then examine the geopotential height anomaly patterns associated with AR landfalls along the West Coast of North America and assess the ability of the models to accurately predict such patterns at leads of 1-14 days. We conclude the study by exploring the possibility of "forecasts of opportunity" at S2S leads using an existing empirical model based on the MJO and QBO.

The following chapter (Chapter 2) discusses the data and methods employed in this study. Chapter 3 discusses the model skill for AR re-forecasts along the West Coast of North America at 1-14 days. Chapter 4 details the skill of geopotential height anomaly re-forecasts at 1-7 days and relates these errors to occurrence-based AR re-forecasts. Chapter 5 summarizes our analysis of AR activity forecasts at S2S leads using an empirical model. Chapters 6 and 7 conclude our analysis and discuss possible avenues of future research. All results related to AR re-forecasts along the West Coast of North America come from Nardi et al. (2018). Contents related to S2S forecasting of AR-related activity are in support of an S2S forecasting initiative at the Center for Western Weather and Water Extremes (CW3E). All material discussed herein is also in support of the Forecast-Informed Reservoir Operations (FIRO) project at CW3E. This project aims to leverage

the predictability of AR landfalls at various time scales in order to better manage water levels in reservoirs throughout the western United States.

Chapter 2

Data and Methods

2.1 S2S Database

In order to assess the performance of modern operational weather prediction models in the prediction of ARs, re-forecasts provided by the Subseasonal to Seasonal (S2S) International Project database (Vitart et al., 2017) are used. From this database, control runs from nine different models are analyzed at leads of 1-14 days (Table 2.1). As seen in Table 2.1, both the temporal range and frequency of initializations vary by model. For all models, re-forecast fields of specific humidity and horizontal wind are analyzed at a horizontal resolution of 1.5° latitude x 1.5° longitude and at constant pressure levels of 1000, 850, 700, 500, 300, 200, 100, 50, and 10 hPa. Geopotential is also retrieved at 500 hPa and then divided by the gravitational acceleration in order to obtain geopotential height. For a given initialization, re-forecasts are valid at 0000 UTC each day. Re-forecast data from four of the models (BOM, ECCC, JMA, and NCEP) do not include output for Day 0 (i.e. the initialization time).

2.2 Reanalysis data

As a means of providing verification for the re-forecast models, ERAI reanalysis data are used to approximate the observations of ARs. Though Wick et al. (2013) used satellite observations for model verification of ARs, reanalysis data have been widely used in observational (e.g. Guan and Waliser, 2015; Lavers and Villarini, 2013; Lavers et al., 2012; Mundhenk et al., 2016a) and model verification (e.g. Baggett et al., 2017; DeFlorio et al., 2018; Nayak et al., 2014) studies of AR activity. Furthermore, prior studies (e.g. Jackson et al., 2016; Lavers et al., 2012) found fairly good agreement between different reanalysis products in the depiction of ARs. An advantage of using reanalysis over satellite data is the ability to continuously depict the low-level transport of water vapor in the troposphere. Instantaneous fields of specific humidity and horizontal wind from

Table 2.1: Characteristics of the nine numerical weather prediction models assessed in this study. Note that models have different initialization frequencies. Additional information can be found in (Vitart et al., 2017). Modeling centers are as follows: Bureau of Meteorology, Chinese Meteorological Administration, National Centre for Meteorological Research, Environment and Climate Change Canada, European Centre for Medium-Range Forecasts, Hydrometcentre of Russia, Japan Meteorological Agency, National Centers for Environmental Prediction, and United Kingdom Met Office.

Modeling Center	Initialization Years	Num. Initializations in NDJF	Missing Day 0?
BOM	1981-2013	792	Yes
CMA	1994-2014	2520	No
CNRM	1994-2006	100	No
ECCC	1995-2014	340	Yes
ECMWF	1995-2016	1380	No
HMCR	1985-2010	858	No
JMA	1981-2010	360	Yes
NCEP	1999-2010	1439	Yes
UKMO	1996-2009	224	No

reanalysis data are analyzed at a horizontal resolution (namely, 1.5° x 1.5°) and vertical resolution that matches the re-forecast model data described above. In addition, the reanalysis data have a similar temporal resolution of 24 hours and are also valid at 0000 UTC for a 38-year period spanning from 01 January 1979 through 31 December 2016. Geopotential height is also obtained at 500 hPa from ERAI at the same resolution by retrieving geopotential and then dividing by the gravitational acceleration.

For the purposes of running the S2S empirical model discussed in Section 2.8, historical data for the MJO and QBO are obtained and applied in a similar way to Mundhenk et al. (2018). Namely, the MJO is defined using the real-time multivariate MJO (RMM) index. This index combines the two leading principal components (PC1 and PC2) from an empirical orthogonal function (EOF) analysis of equatorially-averaged outgoing longwave radiation and zonal winds at 200 and 850 hPa over the tropics. From the PC1 and PC2 time series, an MJO index that provides a phase (1 through 8) and amplitude is derived. The QBO is defined by mean standardized monthly zonal wind anomalies at 50 hPa. Data are provided by the National Oceanic and Atmospheric Administration (NOAA) National Weather Service (NWS) Climate Prediction Center (CPC). To remain

consistent with Mundhenk et al. (2018), we use MJO and QBO data for December-March (DJFM) for the time period 1980-2016. For further information about MJO and QBO inputs, please see Mundhenk et al. (2018).

Historical low-level water vapor transport data for the empirical model are obtained from ERAI, as described above, for DJFM for the period 1980-2016. Precipitation data for the same period come from the CPC Global Unified Precipitation dataset provided by NOAA's Earth System Research Laboratory (ESRL) Physical Sciences Division (PSD) (Chen et al., 2008; Xie et al., 2007). This gauge-based dataset provides global daily precipitation accumulation at a spatial resolution of 0.5° latitude x 0.5° longitude.

2.3 AR detection

The aforementioned isobaric fields of specific humidity and horizontal wind are used to calculate integrated water vapor transport (IVT) across the globe. IVT is a vector quantity, but here we use IVT to refer only to IVT magnitude, which is calculated based on the formula from Lavers et al. (2012):

$$IVT = \sqrt{\left(\frac{1}{g}\int_{1000}^{300} qu\mathrm{d}p\right)^2 + \left(\frac{1}{g}\int_{1000}^{300} qv\mathrm{d}p\right)^2} \tag{2.1}$$

Here, g is gravitational acceleration, q is specific humidity of water vapor, u is zonal wind, v is meridional wind, and p is pressure. The integration bounds are pressure levels in units of hPa.

IVT calculations are performed for both the re-forecast and reanalysis data. Calculated fields of IVT are then input into a modified version of the AR detection algorithm described in Mundhenk et al. (2016a). This algorithm scans the IVT field for grid cells that exceed a specific intensity threshold that is based on IVT. This results in a number of candidate AR objects that are subsequently put through various geometric tests in order to obtain plume-like corridors of high atmospheric water vapor transport. The modified version of the algorithm used in this study incorporates an instantaneous absolute IVT threshold of 500 kg $\cdot$ m^{-1} $\cdot$ s^{-1}, compared to the IVT anomaly threshold of 250 kg $\cdot$ m^{-1} $\cdot$ s^{-1} used in Mundhenk et al. (2016a). A threshold that incor-

porates absolute IVT, rather than anomalous IVT, is preferred in this study since the calculation of anomalous IVT requires the calculation of the seasonal cycle of IVT. In order to consistently detect AR features in fields of IVT for multiple models with different background climatologies, a choice of a single seasonal cycle would be required. However, such a choice may introduce an inherent bias toward a particular model. Likewise, using a climatology based on reanalysis may introduce a bias. Thus, we use instantaneous absolute IVT to define ARs in this study.

Figure 2.1 shows an example of AR features detected in a field of IVT over the Pacific Ocean. It is clear that the detection algorithm is adequate in identifying long, plumelike corridors of high water vapor transport. A detailed description of all modifications to the Mundhenk et al. (2016a) algorithm can be found in Appendix A. AR detection in the model data is confined to initializations in the Northern Hemisphere wintertime, defined here as the November-February (NDJF) time period. This time period is chosen in order to capture periods of high AR activity along the climatologically-diverse West Coast of North America (e.g. Mundhenk et al., 2016a,b).

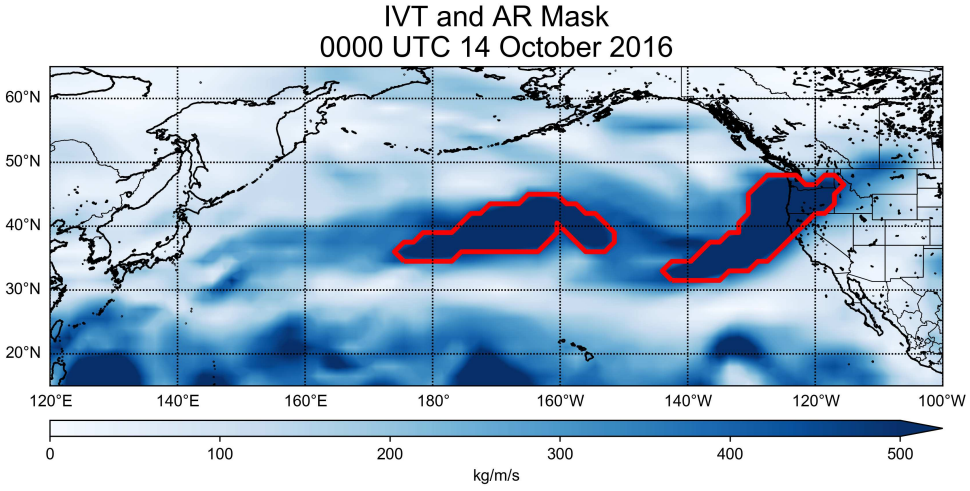

Figure 2.1: Integrated water vapor transport at 0000 UTC on 14 October 2016. Red contours outline two distinct atmospheric river (AR) features detected by the AR detection algorithm.

2.4 Landfall region

Here, we focus our evaluation of AR re-forecasts on those features that make landfall along the West Coast of North America (Figure 2.2). This landfall region is comprised of 18 1.5° x 1.5° grid cells that lie just offshore of North America, from the vicinity of Santa Barbara, CA, northward to near Juneau, AK. The domain is located slightly offshore in order to eliminate the influence of coastal features that may not be resolved at the spatial resolution of the models. From the AR detection algorithm's output, AR "catalogs" (lists of occurrences and non-occurrences) are generated at each of the 18 individual grid cells for the reanalysis data as well as each of the nine models. At each time step, if an AR is detected within a particular grid cell in the landfall region, an AR landfall is recorded in that grid cell's catalog. For reanalysis data, the relevant date of the landfall is recorded. For the re-forecast data, the initialization date and re-forecast lead are recorded. Note that it is possible for a single AR to make landfall at multiple grid cells simultaneously. Moreover, for landfalling ARs that persist for multiple days, each day is considered separately, so the frequency of AR occurrence is quantified in terms of AR landfall days as opposed to unique AR landfall events.

Figure 2.2 shows the NDJF climatology of AR occurrence based on the algorithm run on ERAI data from 1979 through 2016. Climatologically, there are between 1 and 7 AR days per NDJF season along the landfall domain. AR activity is highest over the Pacific Northwest (near about 44°N), while activity gradually decreases to the north and south. This general pattern agrees well with previously published AR climatologies (e.g. Mundhenk et al., 2016a,b).

2.5 Occurrence-based model verification

In terms of AR occurrence at a grid cell, we allow only two outcomes: yes (a landfalling AR is present) or no (a landfalling AR is not present). From the re-forecast and reanalysis-based AR catalogs, these binary values are determined at each time step and grid cell. Occurrence-based model verification is done for each of the 18 grid cells comprising the landfall region by comparing each re-forecast to reanalysis on the valid day. This model verification of a discrete

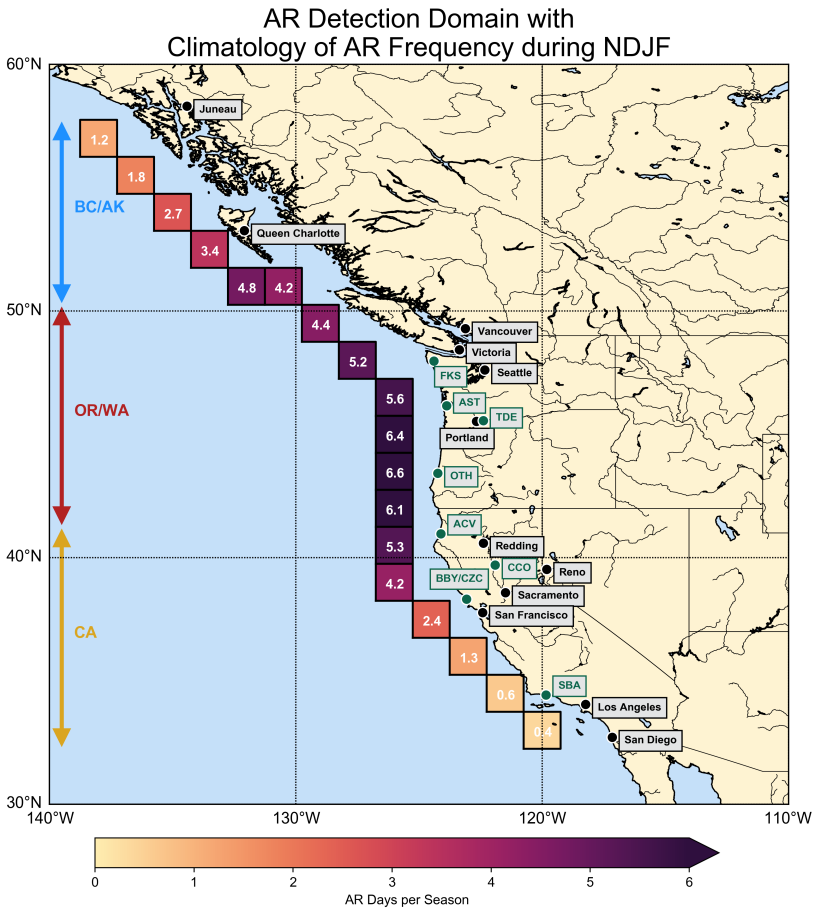

Figure 2.2: The domain used for analysis of re-forecasts of AR occurrence. The colored arrows refer to three sub-regions chosen for the analysis. Shading denotes the climatological number of AR days per NDJF season based on ERA-Interim reanalysis from 1979 through 2016. Black circles highlight the locations of several important population centers along the western coast of North America, while the green circles indicate locations of AR Observatories (AROs).

binary variable allows for the use of a four-outcome (2 x 2) contingency table as described in Wilks (2006). The first outcome, denoted here as "a", corresponds to a situation in which the model correctly re-forecasts an AR for a given valid day. These outcomes are considered "hits". The second outcome, "b", corresponds to a situation in which the model re-forecasts an AR for a given valid day but the reanalysis does not show an AR on that valid day. These outcomes are considered "false alarms". The third outcome, "c", corresponds to a situation in which the model does not re-forecast an AR for a given valid day but the reanalysis shows an AR on that valid day.

These outcomes are considered "misses". The fourth outcome, "d", corresponds to a situation in which the model correctly does not re-forecast an AR for a given valid day. These outcomes are considered "correct rejections". This characterization of AR occurrence-based model verification is similar to that used by Wick et al. (2013). For each model, lead time, and grid cell, counts of the four outcomes are tallied. Contingency-table-based skill metrics (Wilks, 2006) are then calculated based on these tallies. Such skill metrics based on binary occurrence predictands are commonly used in model verification of ARs/extreme precipitation events (e.g. DeFlorio et al., 2018; Nayak et al., 2014; Ralph et al., 2010; Wick et al., 2013).

Though various contingency-table-based skill metrics exist, this study focuses on four particular metrics. First, frequency Bias (B) is the number of AR occurrences re-forecast by the model divided by the number of AR occurrences detected in reanalysis (in the above terminology, hits + false alarms divided by hits + misses):

$$B = \frac{a+b}{a+c} = \frac{\text{hits + false alarms}}{\text{hits + misses}} \tag{2.2}$$

In general, Bias gives a sense of how much a particular model favors re-forecasting ARs compared to reanalysis. Bias can range from 0 to infinity, with Bias greater than 1 implying that the model re-forecasts more ARs than reanalysis and Bias less than 1 implying that the model re-forecasts fewer ARs than reanalysis. Next, Hit Rate (H) is the frequency with which the model re-forecasts an AR for the valid date given that an AR occurs in reanalysis on the valid date (i.e. the frequency of a hit given a hit or a miss), and False Alarm Rate (F, also known as Probability of False Detection) is the frequency with which the model re-forecasts an AR for the valid date given that an AR does not occur in reanalysis on the valid date (i.e. the frequency of a false alarm given a false alarm or correct rejection):

$$\begin{aligned} H &= \frac{a}{a+c} = \frac{\text{hits}}{\text{hits + misses}} \\ F &= \frac{b}{b+d} = \frac{\text{false alarms}}{\text{false alarms + correct rejections}} \end{aligned} \tag{2.3}$$

Both H and F can range from 0 to 1. The denominator of H is the climatological number of AR occurrences, while the denominator of F is the climatological number of AR non-occurrences.

The final skill metric is the Peirce Skill Score (PSS), which combines H and F in order to give a sense of how the model performs compared to a random forecast:

$$PSS = H - F = \frac{\frac{a+d}{n} - \left[\frac{(a+b)(a+c)+(b+d)(c+d)}{n^2}\right]}{1 - \left[\frac{(a+c)^2+(b+d)^2}{n^2}\right]} \tag{2.4}$$

The numerator compares the probability of a correct forecast using the model compared to the probability of getting a correct forecast by random chance. The probability in the denominator is the probability of a correct forecast when forecasting based on the region's climatological AR frequency. By definition, PSS ranges from -1 to 1, with a value of 1 indicating a perfect score and a value of 0 indicating no improvement over a random forecast. Constant forecasts also have a PSS of 0. Negative values imply that the model provides less skill than a random forecast. Since PSS uses a region's climatology as a reference, a forecaster is not heavily penalized for incorrectly predicting a climatologically rare event (such as a landfalling AR) (Wilks, 2006).

PSS is the chosen skill metric here for two main reasons. First, PSS succinctly provides a measure of how much additional re-forecast skill is gained from using a particular model compared to randomly forecasting AR occurrences. Second, since PSS incorporates a region's own climatology, fair comparisons in skill can be made between regions with disparate background climatologies, as is the case with the three sub-regions defined in Figure 2.2. One caveat with using PSS here is that since landfalling ARs, as defined by the detection algorithm, are relatively infrequent events (Figure 2.2), correct rejections are much more likely than the other three outcomes. In this situation, PSS can be artificially improved by simply increasing the number of "yes" re-forecasts (Jolliffe and Stephenson, 2003). Due to this limitation, Nayak et al. (2014) used the SEDI skill metric for AR forecast evaluation in order to eliminate such issues, but this skill metric does not use climatology as a baseline, so it could be problematic to apply SEDI to locations with varying background climatologies. Therefore, we use PSS as a measure of occurrence-based skill.

2.6 Assessment of intensity and location re-forecasts

Errors in AR intensity (as measured by IVT) are examined along and offshore of the West Coast of North America. For each model, IVT is examined at time steps for which an AR occurrence is correctly re-forecast (i.e. hits). In other words, IVT errors are examined for re-forecasts that accurately predict the presence of AR conditions as defined by our algorithm. For each sub-region, all re-forecast hits for the individual grid cells within the sub-region are analyzed. Absolute IVT error is calculated by subtracting the reanalysis IVT from the re-forecast IVT. From this formula, positive absolute IVT errors indicate that the model predicts more IVT than what occurs in reanalysis, and negative absolute IVT errors indicate that the model predicts less IVT than what occurs in reanalysis.

To calculate location errors, we define "landfall location" as the *median* latitude and longitude of the landfall grid cells with which the AR makes contact. After identifying landfalling AR features in reanalysis with landfall locations within a given sub-region, we examine re-forecasts for identified AR days and compare the landfall location from the model to that from reanalysis. If multiple AR features are re-forecast by the model, the feature closest to the reanalysis landfall location is used and compared to reanalysis. Once the re-forecast and reanalysis AR features are identified, the landfall location error is defined as the distance between the re-forecast and reanalysis landfall locations along a great circle.

2.7 Assessment of geopotential height re-forecasts

In order to assess model re-forecasts of the synoptic-scale patterns associated with landfalling ARs, we examine re-forecasts of the 500-hPa geopotential height anomalies during NDJF. Anomalies are calculated with respect to the individual models in order to avoid issues related to systematic model biases. Since the re-forecasts are initialized on the same calendar dates each year, a series of daily climatologies can be generated for a given initialization date and forecast lead (e.g. a 5-day re-forecast initialized on 01 February). Geopotential height anomalies are calculated for each model by subtracting the model's calendar-day average geopotential height from the re-

forecast geopotential height. Geopotential height anomalies in ERAI are calculated by subtracting the reanalysis calendar-day average geopotential height from the reanalyzed geopotential height for the particular day.

For each AR day within a particular sub-region, the sign of the reanalysis geopotential height anomaly is determined at each global grid cell. For each grid cell, if at least 80% of the AR days have a height anomaly of the same sign, the grid cell is considered to have a consistent height anomaly pattern for AR days in that sub-region. Taken together, grid cells with consistent height anomalies for AR days comprise a "favorable" geopotential height pattern (in terms of the sign of the anomaly) for AR landfalls in the sub-region. Figure 2.3 shows these favorable height patterns for each sub-region. Blue grid cells indicate negative height anomalies (an anomalous trough), while red grid cells indicate positive height anomalies (an anomalous ridge). These patterns are associated with anomalous flow patterns that direct AR objects into the particular sub-regions.

Re-forecast skill in predicting geopotential height anomalies is examined for a particular model and sub-region for all days (i.e. independent of whether or not an AR landfall occurred) with height anomaly patterns that resemble the favorable AR landfall patterns. We examine days on which there is a pattern correlation of at least 0.90 between the observed height anomaly pattern and the favorable pattern (in terms of the sign of the anomaly) over the northeast Pacific (defined here as 20-60°N in latitude and 200-250°E in longitude). From visual inspection, we are confident that the dates examined represent days with height anomaly patterns similar to the conducive height anomaly patterns for the different sub-regions.

2.8 S2S empirical prediction model

To explore the predictability of ARs at S2S leads (defined here as 2-5 weeks), we use an empirical prediction scheme introduced by Mundhenk et al. (2018). This model produces forecasts of above or below-normal "AR activity" for a particular domain, where AR activity is calculated as a 5-day running mean. Mundhenk et al. (2018) defines AR activity based on the number of AR features identified by the detection algorithm. However, in order to avoid the need to choose from

various existing detection algorithms, we modify the model and define AR activity in terms of IVT and precipitation, two predictands that are independent of the choice of detection algorithm. In this way, we introduce a more objective method of predicting anomalous water vapor transport that could pose a risk to locations along the West Coast of North America. Using precipitation, we also allow for an assessment of how well models predict the weather-related impacts of ARs.

The model builds an ERAI-based climatological distribution (based on all verification dates during DJFM from 1980-2016) of AR activity and sets the distribution's 50th percentile as a threshold for above or below-normal activity. A conditional distribution of AR activity can be derived from this climatological distribution for a given lead time and MJO/QBO phase combination. For example, a conditional distribution can be derived for AR activity 3 weeks (21 days) following MJO phase 5 during a westerly QBO. A forecast is only made when the MJO phase has persisted for at least 2 days and when at least 1 day has an MJO amplitude greater than 1. The QBO phase is based on the previous month's average 50-hPa wind anomaly. If the conditional distribution's median is greater than the climatological distribution's 50th percentile threshold, a forecast of above-normal AR activity is made. However, if the conditional distribution's median is less than the climatological distribution's 50th percentile threshold, a forecast of below-normal AR activity is made.

Following Mundhenk et al. (2018), the skill of this empirical model is evaluated using a leave-one-season-out cross validation approach. With this approach, climatological and conditional distributions of DJFM AR activity are generated from all seasons from 1980-2016, sequentially leaving one season out. Forecasts are then made for the DJFM season that was left out. These forecasts are verified using the observed AR responses (i.e. above or below-normal) during that season. For example, forecasts during the year 2000 season are made based on historical AR activity data from all seasons except the 2000 season. These forecasts are compared to observed anomalous AR activity during the 2000 season. Such a procedure is performed for every season in the historical record. Since there are only two possible responses, a 2 x 2 contingency table can be applied to assess model skill. As in Mundhenk et al. (2018), the Heidke Skill Score (HSS) is used:

$$HSS = \frac{\frac{a+d}{n} - \left[\frac{(a+b)(a+c)+(b+d)(c+d)}{n^2}\right]}{1 - \left[\frac{(a+b)(a+c)+(b+d)(c+d)}{n^2}\right]} = \frac{H-E}{T-E} \tag{2.5}$$

Here, H is the number of correct forecasts, T is the total number of forecasts, and E is the number of correct forecasts expected by random chance. Since the threshold for forecasting above or below-normal AR activity is the 50th percentile of the climatological distribution, a random forecast would be expected to be correct 50% of the time. Therefore, E is simply $\frac{T}{2}$. An HSS of 1 implies a perfect model, and an HSS of 0 implies that the model provides no additional skill compared to a random, equal-chances forecast. A negative HSS implies that the model provides less skill than a random forecast. In this framework, an HSS of approximately 0.33 implies the model has a 2-to-1 ratio of correct forecasts to incorrect forecasts. Further information regarding the model and validation procedure can be found in Mundhenk et al. (2018).

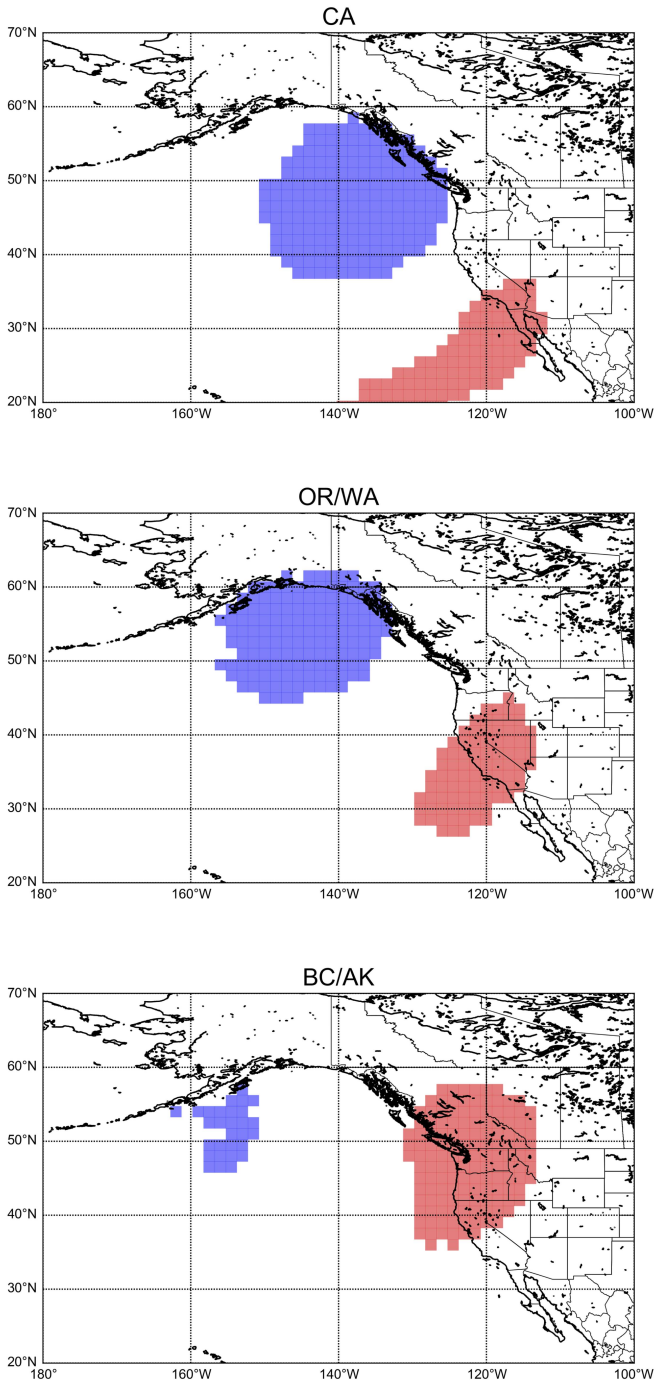

Figure 2.3: Sign of conducive 500-hPa geopotential height anomalies for AR landfalls within the three sub-regions. A blue (red) grid cell implies that at least 80% of AR days within the sub-region featured a negative (positive) geopotential height anomaly at that grid cell.

Chapter 3

AR Re-Forecasts at 1-14 Days

3.1 AR occurrence

Our evaluation of model re-forecast performance begins with our first component of an effective AR forecast: the correct prediction of an AR's presence in a region. Recall that there are four possible outcomes for AR occurrence re-forecasts: hits, false alarms, misses, and correct rejections. Tallies of these outcomes are used to calculate four skill metrics: Bias (B), Hit Rate (H), False Alarm Rate (F), and Peirce Skill Score (PSS).

Figure 3.1 shows Bias for each of the nine models by re-forecast lead time. Here, Bias for each of the 18 grid cells is calculated, and then these 18 values are averaged in order to obtain a single mean Bias for the West Coast of North America per model. Blue colors indicate more AR occurrences in the re-forecasts compared to AR occurrences in reanalysis ("active Bias"), while red colors indicate fewer AR occurrences re-forecast compared to AR occurrences in reanalysis ("quiet Bias"). Overall, Bias tends to be higher at later lead times, with five of the nine models trending toward a pronounced active Bias at later lead times. However, this upward trend is not seen in all of the models. For example, Bias for ECMWF peaks around lead times of 10 to 14 days but actually decreases at later lead times. Additionally, only two models (ECMWF and UKMO) have a quiet Bias for the majority of lead times, with ECMWF exhibiting a quiet Bias for all re-forecast leads. Overall, Figure 3.1 provides a sense of a particular model's propensity to re-forecast landfalling ARs along the West Coast of North America.

Though useful for an overall picture of how often a model re-forecasts an AR, Bias alone does not capture a model's skill at forecasting the occurrence of ARs. Specifically, it is important to understand the frequency of incorrectly re-forecasting an AR when one does not actually occur (i.e. F), as well as the frequency of correctly re-forecasting an AR when one actually does occur (i.e. H). Figure 3.2 shows plots of skill scores with H on the vertical axis and F on the horizontal

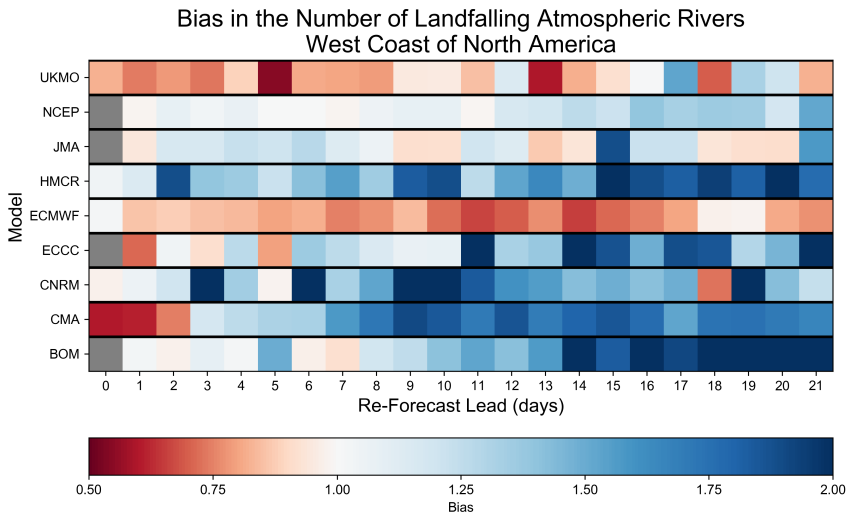

Figure 3.1: Model bias (ratio of AR occurrence re-forecasts to AR observations) plotted as a function of model and re-forecast lead time (in days). Re-forecast leads for which there is no data are shaded gray. Red colors indicate fewer AR occurrences re-forecast than observed in reanalysis data, and blue colors indicate more AR occurrences re-forecast than observed in reanalysis data.

axis. As before, H and F are calculated for each grid cell and then averaged over the West Coast of North America. Figure 3.2 shows re-forecast skill for all nine models out to 14 days, with Day 0 eliminated because data for this lead time are not available for all models. The numerical values of H, F, and PSS for each model are listed in Table B.1 in Appendix B. It is clear that H decreases and F increases as lead time increases, as seen in the movement from upper left to lower right. Overall, there exists a large range of skill values between models. H decreases from between about 0.4 and 0.8 at a lead of 1 day to between about 0.02 and 0.10 at a lead of 14 days. Meanwhile, F increases from between about 0.005 and 0.010 at a lead of 1 day to between about 0.02 and 0.05 at a lead of 14 days.

Figure 3.2 also shows lines of constant PSS, which is derived from H and F. PSS generally decreases as lead time increases. At lead times around 14 days, PSS approaches 0 for all models, indicating that the models provide little additional skill compared to a random forecast. Also, PSS is between approximately 0.4 and 0.8 at a lead of 1 day. Such values imply a sizable drop in skill just 1 day from initialization, which is due to a sharp decrease in H as described above.

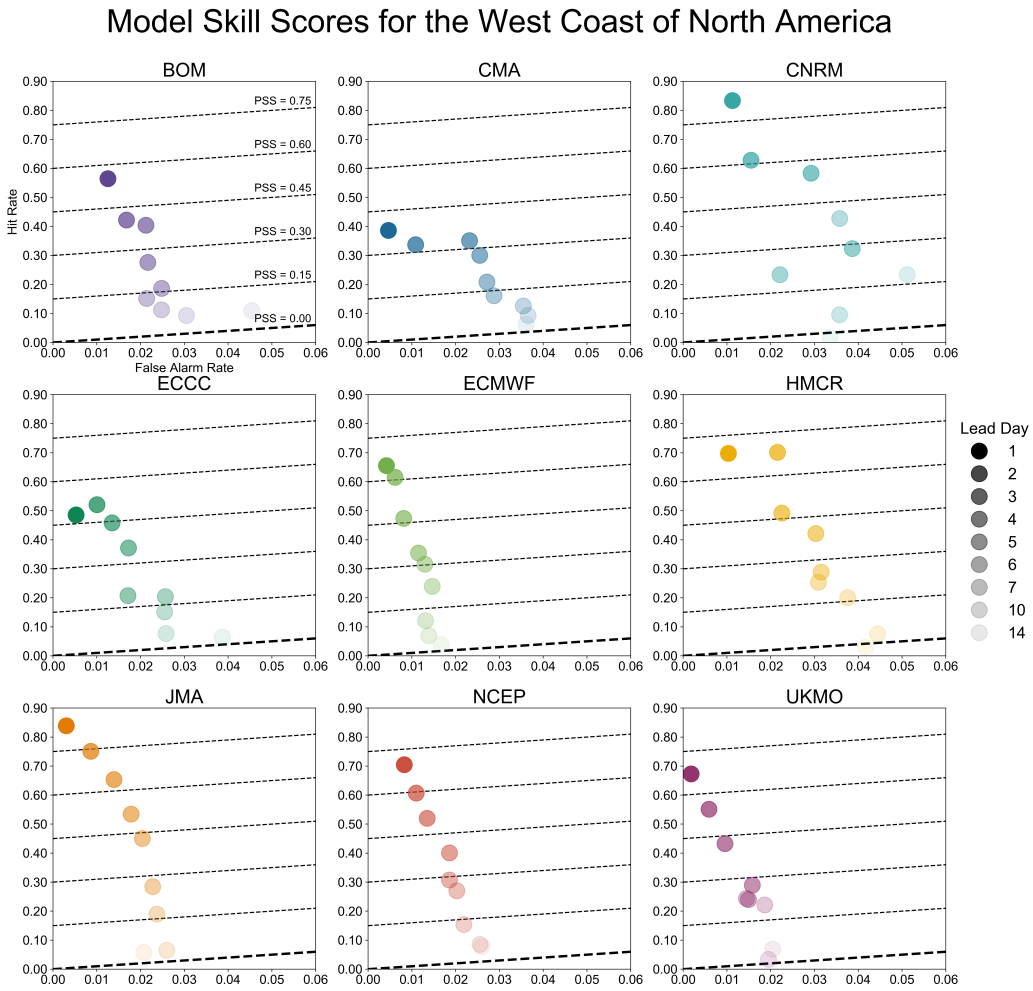

Figure 3.2: Occurrence-based skill scores for the West Coast of North America, with Hit Rate (H) on the y-axis and False Alarm Rate (F) on the x-axis. The black dashed lines denote constant Peirce Skill Score (PSS = H - F). The bold dashed line corresponds to a PSS of 0. Each marker color corresponds to a different model, and each degree of marker transparency corresponds to a different selected re-forecast lead.

In order to examine re-forecast skill at sub-regional scales, Figure 3.3 shows PSS by lead time for three sub-regions (Figure 2.2): CA (Santa Barbara to the California/Oregon border), OR/WA (the California/Oregon border to Vancouver Island), and BC/AK (Vancouver Island to Juneau). The latitude/longitude coordinates for each of these sub-regions are listed in Appendix B. Specifically, for each model and lead time, PSS is calculated for each grid cell within the sub-region, and

then these PSS values are averaged in order to get a single mean PSS for the sub-region. Figure 3.3 shows a steady decrease in model-averaged PSS as lead time increases, with PSS approaching 0 toward 14 days, as already noted, and this trend is consistent across sub-regions. At the same time, Figure 3.3 shows variations in PSS between sub-regions, particularly at leads between 1 and 7 days. At these leads, there is a clear difference in model-averaged re-forecast skill between BC/AK and the other two sub-regions, with BC/AK having lower skill (by approximately 10-20 points) than CA and OR/WA. As defined, a preferred use of PSS is the comparison of locations with different climatologies because the skill metric is adjusted for each sub-region's unique climatology. Thus, the differences in skill between sub-regions cannot solely be attributed to varying background climatologies. These sub-regional differences between 1 and 7 days are statistically-significant (not shown) at 95% using a bootstrapping analysis. In addition, PSS is lowest in BC/AK in all nine individual models for the majority of lead days between 1 and 7. Consistent with our result, Figure 5 of Mundhenk et al. (2018) also showed that occurrence-based AR re-forecast skill (HSS) is about 10-20 points lower in British Columbia than in California at leads less than 7 days.

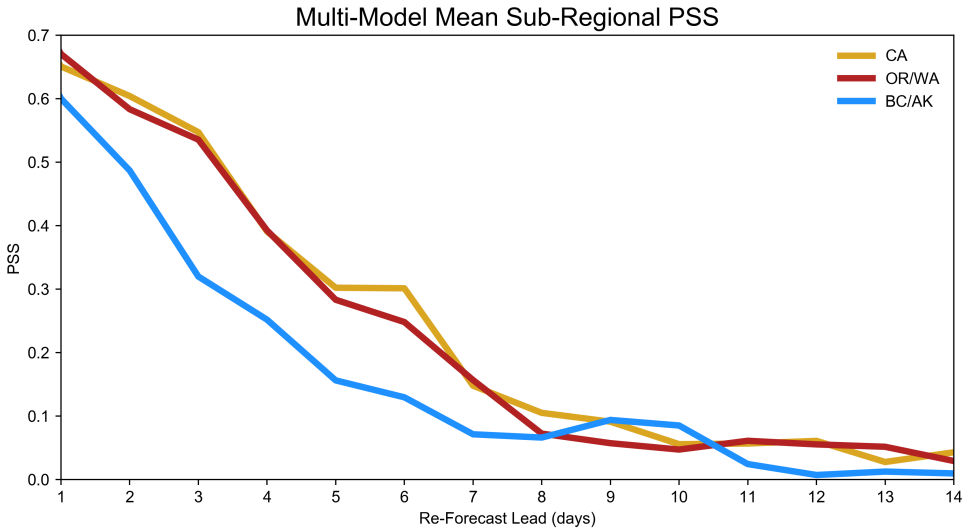

Figure 3.3: Model-averaged Peirce Skill Score (PSS) by re-forecast lead averaged over each of the three sub-regions (see Figure 2.2).

Figure 3.4 shows PSS at an even-smaller scale by showing the skill metric at each individual grid cell in the landfall domain. For each panel, the skill metrics are calculated after combining and tallying all contingency table counts for the lead days that comprise the given lead window (e.g. lead days 1 through 3). While variation exists across models, Figure 3.4 indicates that the highest PSS across all nine models tends to be seen along the coasts of Northern California and Oregon for leads of 1 through 6 days. At these leads, the lowest PSS generally occurs in the grid cells at the northern extent of the landfall domain (e.g. BC/AK), as already seen in Figure 3.3. However, at leads of 7 through 10 days, no particular sub-region is favored for higher PSS. This also corresponds with Figure 3.3, which shows that PSS is similar between sub-regions at leads greater than 7 days.

3.2 AR intensity

A second important aspect of an effective AR forecast is the correct prediction of the feature's intensity (in terms of IVT). Even though a model may accurately re-forecast a landfalling AR for a particular location, a large error in the re-forecast IVT field can still occur.

Figure 3.5 shows the root mean squared error (RMSE) in landfall IVT (i.e. the IVT at the landfall grid cells) for each model and sub-region. Here, lead time is defined in terms of overlapping 3-day lead windows such that lead times of 1 through 3 days, 2 through 4 days, 3 through 5 days, etc., are grouped. Lead windows are applied in this context as a means of smoothing the results. In Figure 3.5, discontinuities appear in CA and BC/AK due to a lack of re-forecast hits for the particular model during the lead window. Overall, little difference exists in the distribution of IVT RMSE between sub-regions, and for all three sub-regions, the average magnitude of landfall IVT RMSE stays fairly constant between 100 and 250 $\text{kg} \cdot \text{m}^{-1} \cdot \text{s}^{-1}$ as lead time increases, though sample sizes decrease to around 10 for several models (e.g. CNRM, ECCC, and UKMO) at leads of 11 through 14 days.

Figure 3.6 shows the distribution of absolute landfall IVT error for all nine models for leads of 1-3, 4-6, 7-10, and 11-14 days. Due to the sub-region's higher AR climatological frequency

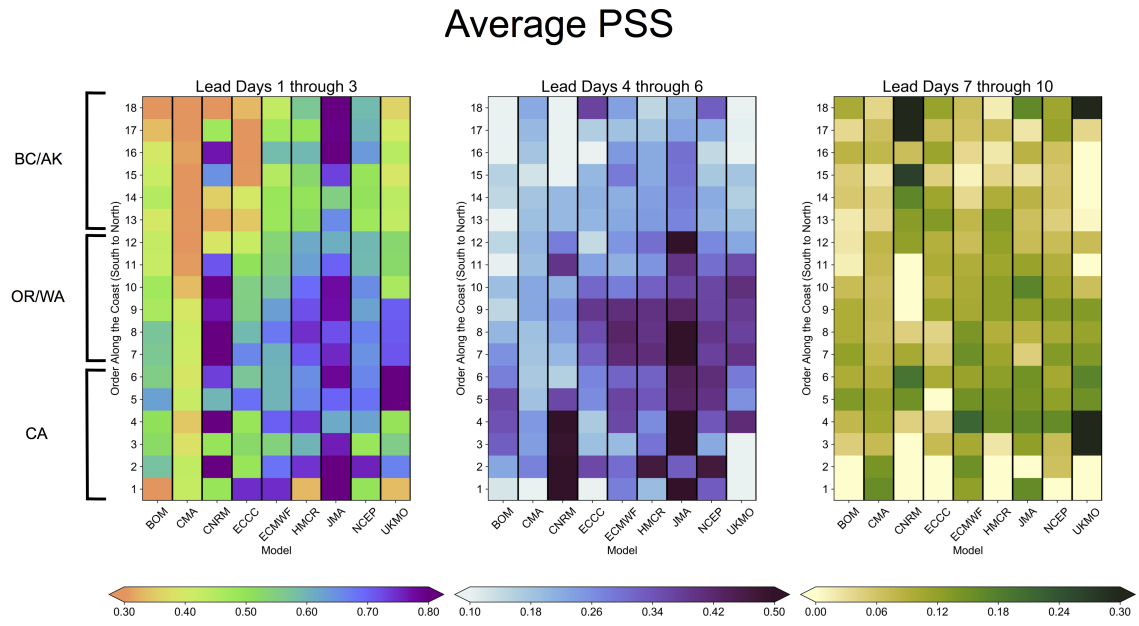

Figure 3.4: Geographical distribution of Peirce Skill Score (PSS) values for the West Coast of North America. Locations along the detection domain (Figure 2.2) are numbered (1 to 18) from south to north along the horizontal axis. Locations 1 through 6 are CA, 7 through 12 are OR/WA, and 13 through 18 are BC/AK. Plotted PSS values are averaged for three different lead windows. Each colorbar denotes a different range of PSS values.

during NDJF, results for OR/WA are shown in Figure 3.6, though the distributions for the other sub-regions are generally similar (see Appendix B). At shorter lead times (i.e. 1-3 days) in OR/WA, distributions of absolute IVT error tend to be centered around 0, while medians of the distributions for most of the models tend to be positive. One notable exception is the median for ECMWF, which is negative at both 1 through 3 days and 4 through 6 days. This indicates that a majority of ECMWF re-forecasts at these lead times have a propensity for features that are less intense than reanalysis. At leads greater than 7 days, all models appear to favor positive absolute IVT error. It is important to remember, however, that the re-forecasts that make up these distributions are hits,

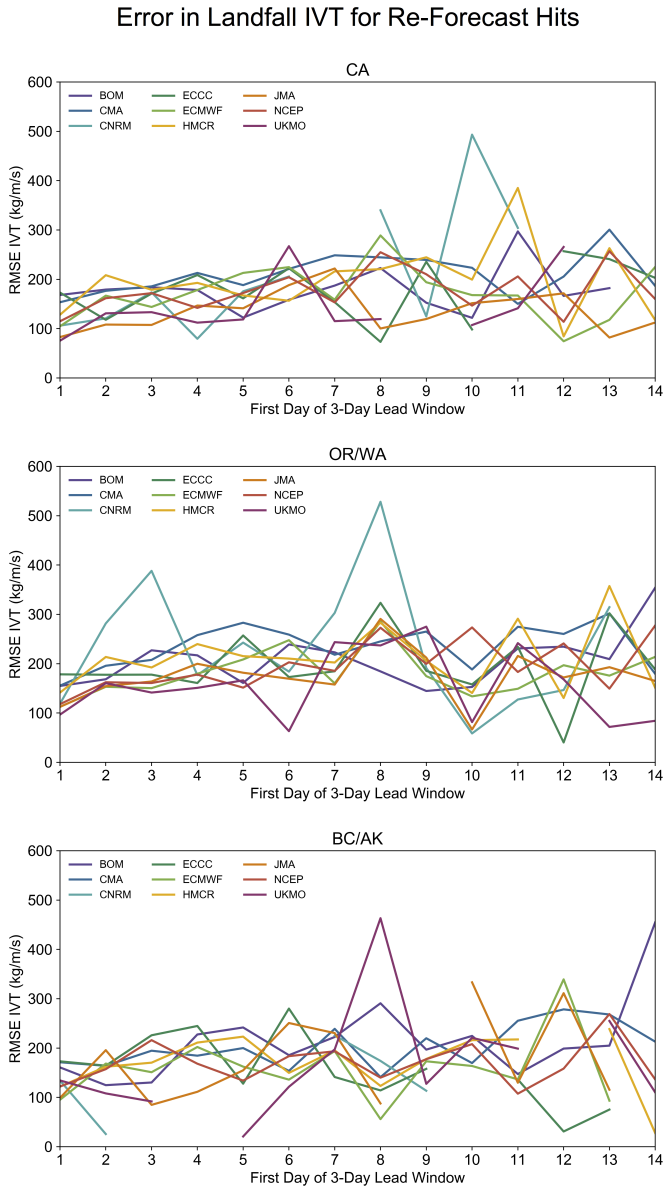

Figure 3.5: Root mean squared error (RMSE) in landfall IVT magnitude for the model and ERA-Interim for re-forecast “hits" in the sub-region during 3-day lead windows. Discontinuities appear for lead windows without any re-forecast hits in the catalog.

so the a priori assumption is that the re-forecasts are correctly predicting the presence of an AR feature but may not have the correct IVT.

Though this paper restricts its study to landfalling ARs, it can be illuminating to understand errors in AR intensity by looking upstream of the West Coast of North America. Re-forecast and

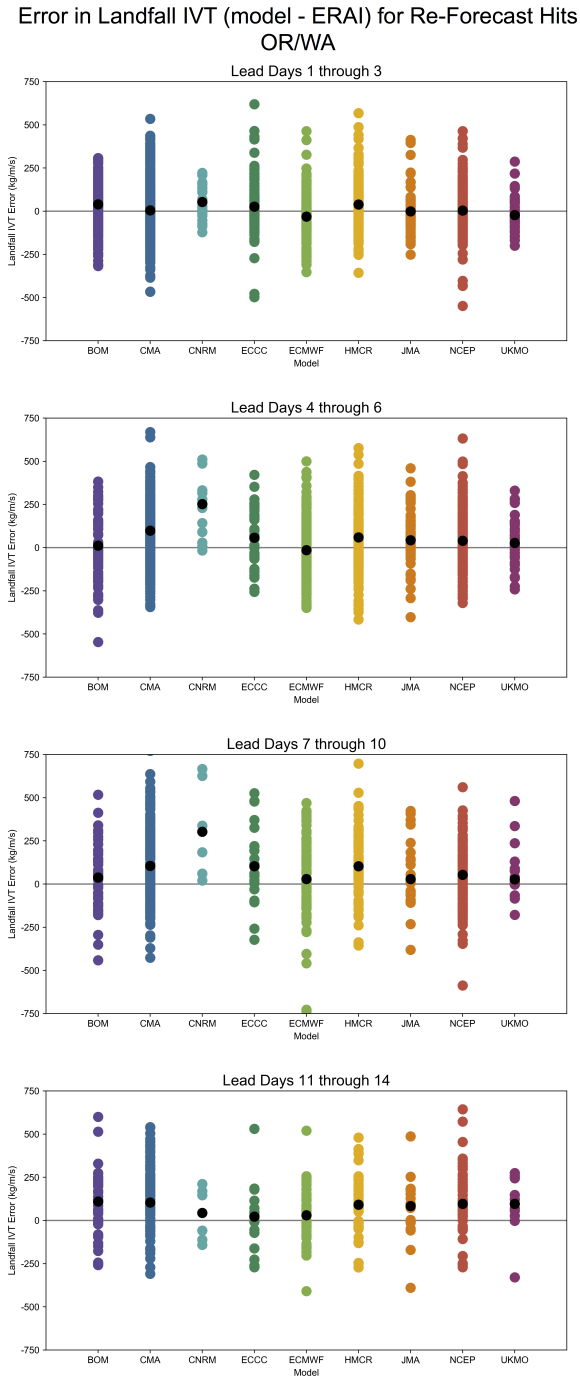

Figure 3.6: Distribution of absolute error in landfall IVT for the model and ERA-Interim (model - ERAI) for re-forecast hits in OR/WA during four different lead windows (1-3 days, 4-6 days, 7-10 days, and 11-14 days). Black dots denote the median of the distribution.

reanalysis fields of IVT for each model, lead time, and sub-region are composited and compared. For such analysis, re-forecast hits are used once again. Results are shown for ECMWF and NCEP, two of the models with the most initializations in our dataset. Figure 3.7 shows the percent difference between the composite IVT fields for ECMWF and reanalysis at lead times of 1 through 3 days for hits, and Figure 3.8 shows the percent difference for NCEP. Black dots denote grid cells at which the absolute difference in means is statistically significant based on a two-sided t-test at 95% confidence. ECMWF IVT re-forecasts show extensive low-IVT biases upstream of the landfall sub-regions. Since an AR feature (i.e. a corridor of high IVT) must be present in the sub-region, absolute low-IVT biases for ECMWF (not shown) are even more prominent. Together, Figures 3.6 and 3.7 point to a low-IVT bias in ECMWF, at short lead times, that exists offshore. By contrast, NCEP (Figure 3.8) shows a less pronounced low-IVT bias offshore, while pronounced high-IVT biases appear farther west over the Pacific.

The intensity errors seen in ECMWF are likely connected to the occurrence-based Bias metric shown in Figure 3.1. Recall that Figure 3.1 showed that ECMWF, at all lead times, re-forecasts fewer ARs along the West Coast of North America compared to reanalysis. Since the AR detection algorithm uses a static IVT cutoff, the dry bias in ECMWF's offshore IVT likely resulted in fewer detected ARs.

3.3 AR landfall location

A third important component of an effective AR forecast is the correct prediction of the location of the AR feature. Recalling the occurrence-based outcomes, a model may not re-forecast an AR for a particular grid cell when one actually occurs (a miss). The model may still re-forecast a landfalling AR for that valid time, but the feature may be located elsewhere along the coast. In another scenario, a model may correctly re-forecast an AR landfall in a sub-region but may not predict the correct landfall location. Therefore, counts of occurrence-based outcomes such as hits and misses cannot solely describe a model's landfall location error. Thus, to better quantify landfall

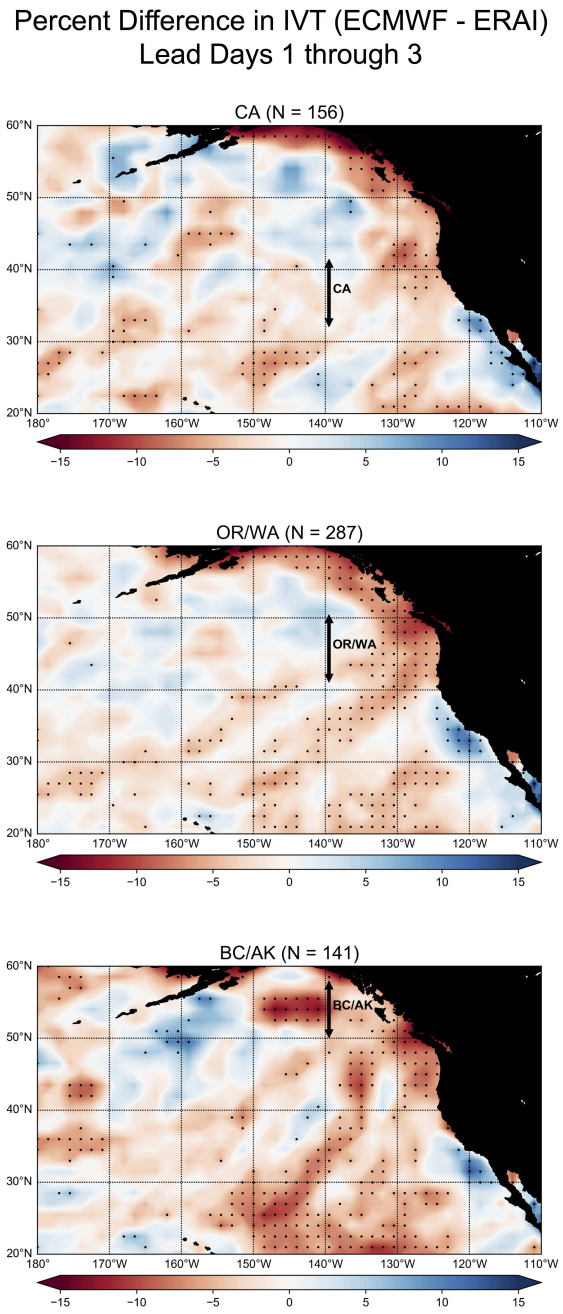

Figure 3.7: Percent difference between composite IVT for the ECMWF re-forecasts and ERA-Interim [(ECMWF - ERAI)/ERAI] for re-forecast hits in the sub-region between lead times of 1 and 3 days. Black dots denote absolute differences that are statistically significant at the 95% confidence level. The sample size is given in the plot titles.

location error, we ask: If an AR makes landfall in a particular location in reanalysis, where did the model tend to place the feature in its re-forecast?

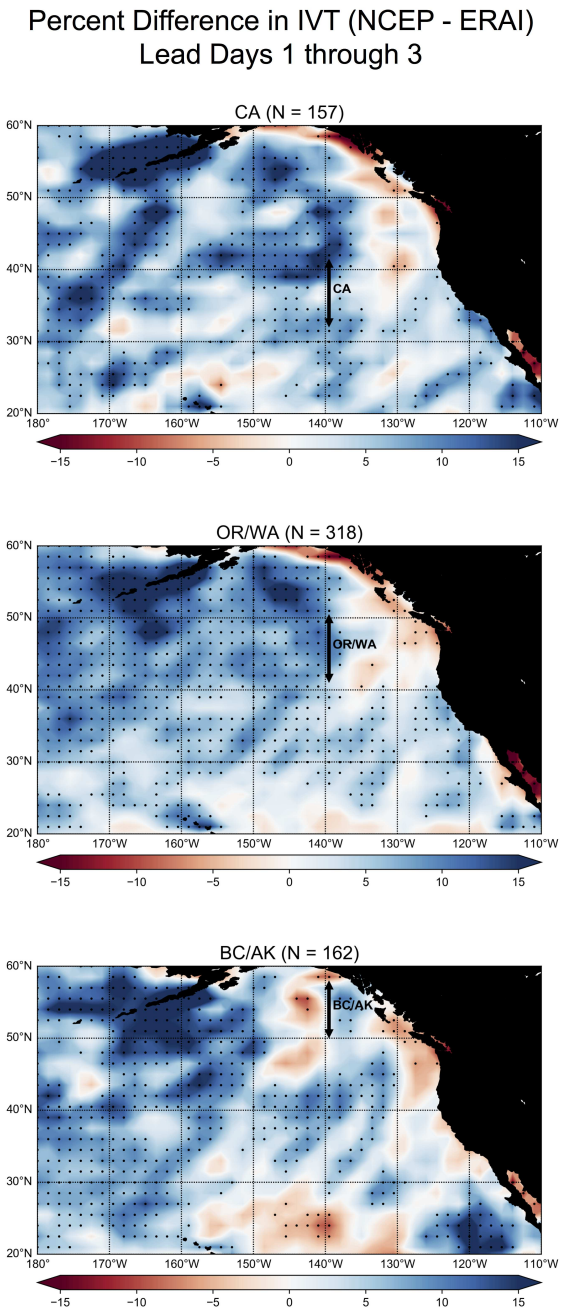

Figure 3.8: Percent difference between composite IVT for the NCEP re-forecasts and ERA-Interim [(NCEP - ERAI)/ERAI] for re-forecast hits in the sub-region between lead times of 1 and 3 days. Black dots denote absolute differences that are statistically significant at the 95% confidence level. The sample size is given in the plot titles.

Figure 3.9 shows, for each sub-region, the distribution of root mean squared error (RMSE) in landfall location error for all nine models as a function of lead time (3-day lead windows). These

errors give a sense of how far away a model places an AR given that one occurs in reanalysis. Landfall location RMSE errors for lead times of 1 through 3 days are generally between 100 and 600 km. All three sub-regions show an increase in landfall RMSE error as lead time increases, with errors exceeding 1000 km as lead times approach 14 (or more) days. At leads of 14 days, it is difficult to determine whether or not the AR feature in reanalysis is the same feature that was forecast 14 days before. However, errors exceeding 1000 km could potentially be explained by errors in predicting synoptic-scale patterns, which have been shown to modulate AR activity between southern Alaska and California (Mundhenk et al., 2016b). Nonetheless, the increase in landfall location RMSE error corresponds with the aforementioned decrease in occurrence-based skill at lead times approaching 14 days. Figure 3.9 does not show statistical significance of the RMSE errors for individual models, so comparisons between models should be made with caution.

Even though RMSE provides a measure of the magnitude of landfall location error, it is also of interest to understand whether the feature is more likely to be re-forecast to the north or south of where it actually makes landfall in reanalysis. Thus, errors in landfall latitude are calculated as the reanalysis landfall latitude subtracted from the re-forecast landfall latitude. For instance, a positive landfall latitude error implies that the model re-forecasts an AR landfall location too far to the north (i.e. a “northward" error). The frequency of northward location error is calculated for each sub-region and plotted in Figure 3.10. The frequencies are calculated as the number of northward errors divided by the number of non-zero (i.e. northward or southward) errors. Perfect landfall location re-forecasts are not included in these calculations. Thus, a frequency greater than 0.5 implies that given an incorrect landfall location re-forecast, the model is more likely to re-forecast the feature farther to the north compared to reanalysis. By contrast, a frequency less than 0.5 implies that the model is more likely to re-forecast the feature farther to the south.

As seen in Figure 3.10, non-zero landfall location errors in CA tend to be northward at all lead times. An exception occurs at the 1-3 day lead window, when BOM, ECMWF, and NCEP slightly favor southward location errors. BC/AK shows the opposite tendency, with models favoring southward non-zero location errors at all lead times. Meanwhile, in OR/WA, a majority of the

models favor southward non-zero landfall location errors at lead windows of 1 through 3 days and 2 through 4 days. However, at later lead times, none of the modeled landfall locations in OR/WA show a consistent propensity to be too far north or south.

What is the likelihood of getting these frequencies by random chance? Since only two non-zero landfall location error outcomes (northward vs. southward) are possible, statistical significance can be tested using the binomial distribution. It could be assumed that, by random chance, a model is equally likely to place a landfalling feature too far to the north vs. too far to the south. Based on this assumption, an incorrect landfall location re-forecast in central CA means that the chances of the model placing the AR feature over northern CA, OR/WA, or BC/AK is equal to the chances of the model placing the AR feature over southern CA (or points farther to the south). However, Figure 2.2 shows that, climatologically, some locations (e.g. central CA) are more likely to see ARs making landfall to the north than to the south. Given that the dynamical models generally reproduce climatology well (not shown), the nine dynamical models likely have a propensity to re-forecast landfalling ARs in climatologically-favored locations. Therefore, the assumption of equal probabilities of northward and southward error may not be the best null hypothesis to test.

Instead, an alternative null hypothesis is that, by random chance alone, reanalysis-based climatology governs the frequency of northward landfall location error for the models. For a given sub-region, statistical significance of the frequency of northward landfall location error (compared to the frequency from climatology) for each grid cell is tested. If greater than 50% of the grid cells within the sub-region have northward frequencies that are statistically significant at 95% confidence, then the frequency for the entire sub-region is considered statistically significant compared to climatology (denoted by dots in Figure 3.10). Gray shading denotes the range of northward landfall location error frequencies expected from climatology.

In CA, it is expected that, by random chance (climatology), about 90-100% of the incorrect landfall location re-forecasts will be too far to the north. However, at lead times less than 10 days, most of the models have significantly lower northward frequencies (approximately 40-80%). In other words, there are more southward errors than anticipated from climatology. Results after

10 days are similar but not statistically significant due to lower sample sizes. In OR/WA, it is expected that about 50-85% of the incorrect landfall location re-forecasts will be too far to the north. However, most of the models in OR/WA have more southward errors than expected from climatology alone, as was seen in CA. In addition, the propensity toward more southward errors in OR/WA than expected from climatology is statistically significant out to 14-16 days for several models: CMA, HMCR, and NCEP. In BC/AK, it is expected that only about 5-35% of the incorrect landfall location re-forecasts will be too far to the north. At early lead times (e.g. 1-3 days, 2-4 days) several models (CMA, ECMWF, and NCEP) have significantly more northward landfall location errors than expected from climatology alone. Otherwise, the results in BC/AK are not statistically significant, implying that the models follow climatology in terms of the errors in their placement of landfalling ARs in this sub-region.

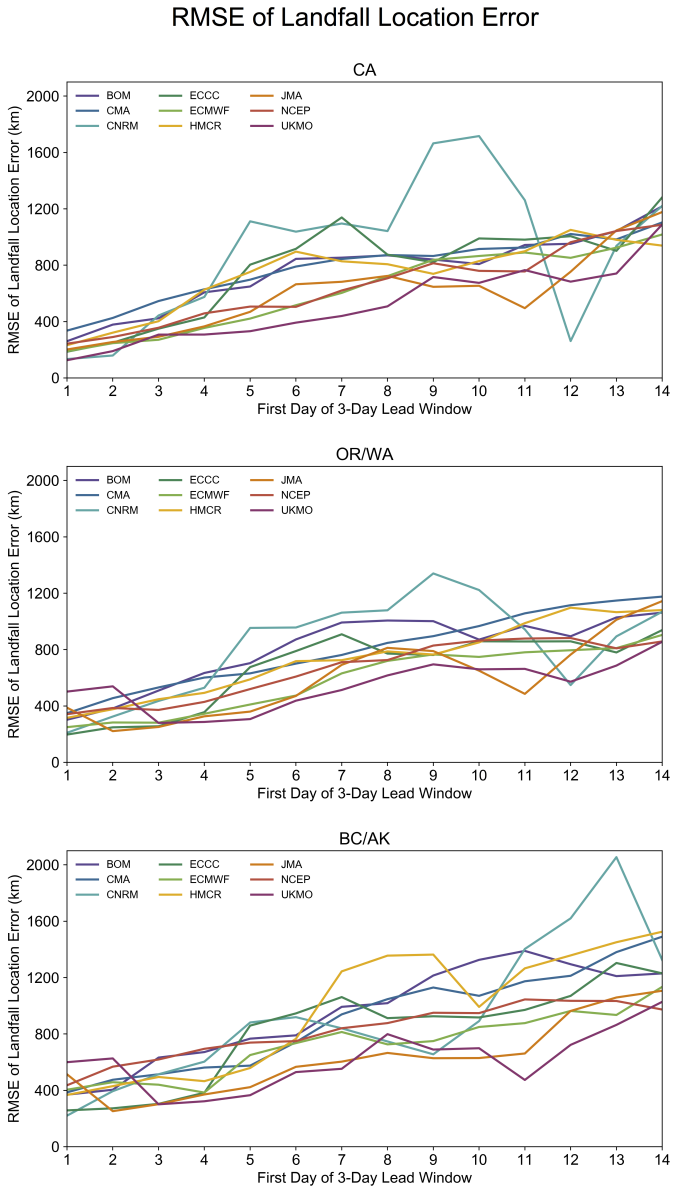

Figure 3.9: Root mean squared error (RMSE) in landfall location error (in km) for ARs observed in the sub-region during 3-day lead windows. An AR must be re-forecast somewhere along the West Coast of North America for the same day as the AR observation.

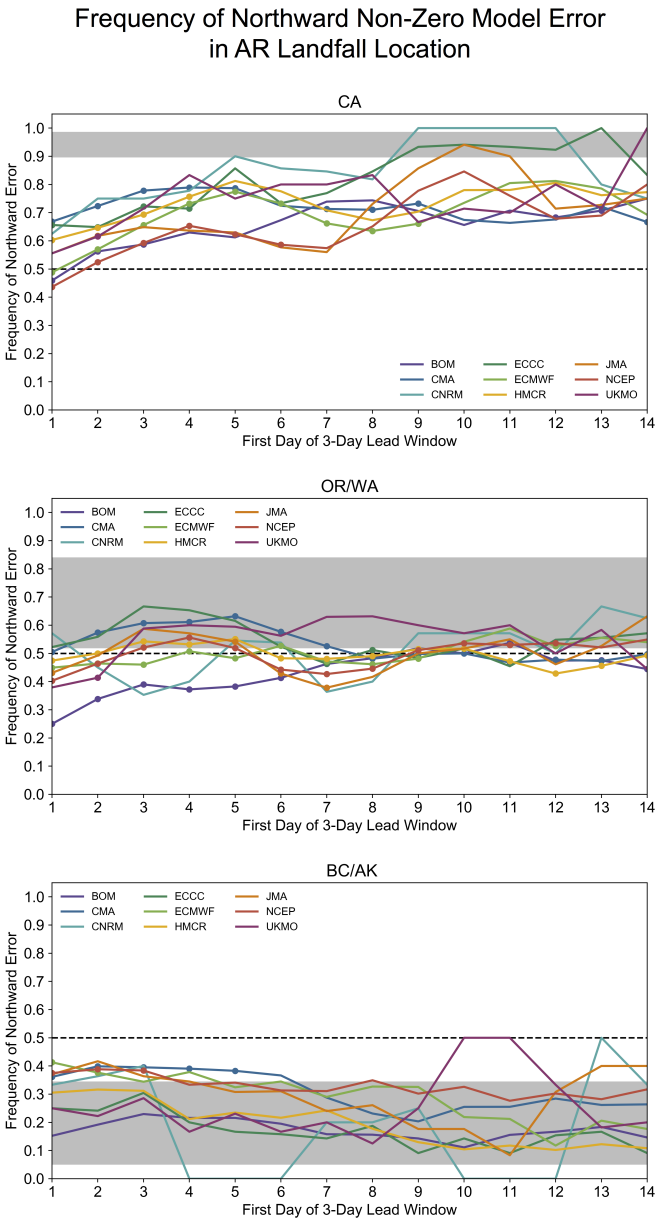

Figure 3.10: Fraction of positive latitudinal landfall location errors for ARs observed in the sub-region during 3-day lead windows. An AR must be re-forecast somewhere along the West Coast of North America for the same day as the AR observation. A positive error indicates that the model's median landfall location is too far to the north. The frequency is calculated as the number of positive non-zero latitude errors divided by the total number of non-zero latitude errors. Dots denote frequencies that are statistically different from what is expected from climatology. Gray shading represents the range of likelihood of having an AR feature to the north (for each grid cell comprising the sub-region) based on ERA-Interim climatology.

Chapter 4

Geopotential Height Re-Forecasts at 1-7 Days

Figure 3.3 highlights a statistically-significant difference in PSS between BC/AK and the other two sub-regions at leads of 1-7 days. We explore this difference in skill between sub-regions by assessing model skill in re-forecasting the 500-hPa geopotential height anomaly pattern over the northeast Pacific. Though there are other possible factors that may influence a model's ability to accurately predict landfalling ARs, Mundhenk et al. (2016b) demonstrated a modulation in AR activity between Alaska and California based on anomalous 500-hPa geopotential height patterns over the northeast Pacific. Specifically, high AR activity was favored in Alaska when an anomalous 500-hPa ridge was located over the Gulf of Alaska. By contrast, high AR activity was favored in California when an anomalous 500-hPa trough was located over the Gulf of Alaska. Figure 2.3 provides further evidence of the sub-regional difference in conducive geopotential height anomaly patterns associated with AR landfalls. With this in mind, we attempt to answer the following question: Do models have less skill in predicting the conducive geopotential height anomaly patterns for AR landfalls in BC/AK compared to CA and OR/WA?

We evaluate model skill using the Spearman rank correlation between the re-forecast and reanalysis 2D spatial fields of 500-hPa geopotential height anomalies. This analysis is done for all days within each sub-region when height anomaly patterns resemble the conducive patterns (in terms of the sign of the anomaly). Once the relevant dates are chosen for analysis, rank correlations are calculated using the actual anomaly fields (as opposed to the signs). In this context, Spearman rank correlation is preferred over RMSE because we are interested in re-forecasts of the overall spatial pattern of geopotential height anomalies, as opposed to the actual values at each grid point. Figure 4.1 shows the average correlation of 500-hPa geopotential height anomaly re-forecasts between ECMWF and ERAI for days during NDJF with height anomaly patterns that resemble the conducive landfall patterns. Leads of 1-7 days are highlighted here because these are the lead days that have the largest PSS differences between sub-regions (Figure 3.3).

At a lead of 1 day, spatial correlations between ECMWF and ERAI are about 0.98 for each of the three sub-regions. For all three sub-regions, correlation decreases from 1 to 7 days, with correlations between about 0.7 and 0.8 at a lead of 7 days. BC/AK has correlations that are consistently lower than those of the other two sub-regions, though differences between BC/AK and OR/WA are not as large as seen in Figure 3.3. Nonetheless, on average, ECMWF is less skillful in predicting the conducive 500-hPa height anomaly pattern for BC/AK compared to other locations along the West Coast of North America. This is especially evident at leads greater than 3 days. Figure 4.2 shows average correlations between NCEP and ERAI. Overall, correlations in all three sub-regions are slightly lower for NCEP than for ECMWF. Correlations range from 0.95 at a lead of 1 day to between about 0.65 and 0.75 at a lead of 7 days. As is the case with ECMWF, model skill is consistently lower for BC/AK. For both models at leads of 4-6 days, we find the differences between CA and BC/AK to be statistically-significant at 95% confidence using bootstrapping. However, differences between OR/WA and BC/AK are not statistically significant at leads of 4-6 days.

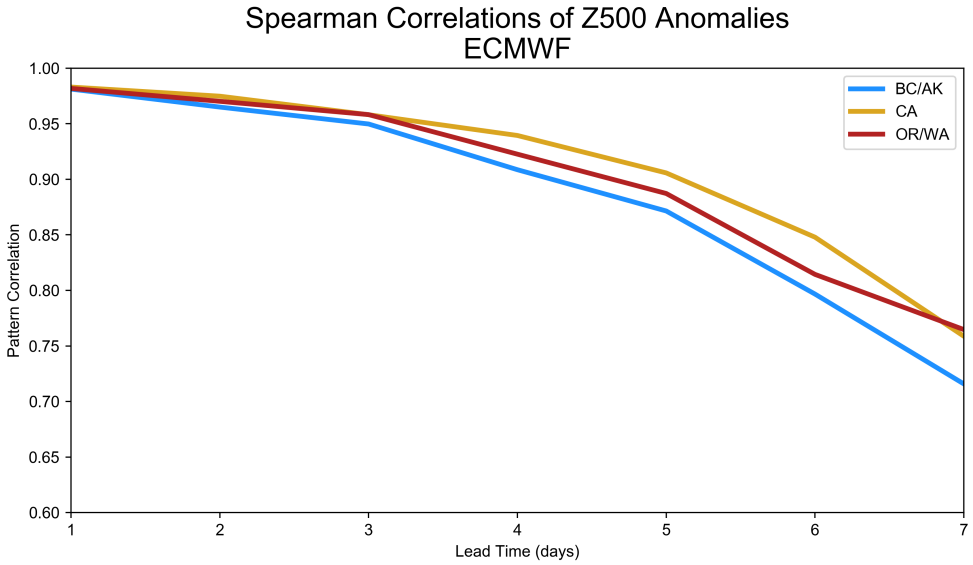

Figure 4.1: Average Spearman rank correlation of geopotential height anomaly re-forecasts between ECMWF and ERAI for NDJF days with geopotential height anomaly patterns that resemble the sub-region's conducive pattern. Correlations are calculated over a swath of the northeast Pacific from 20-60°N latitude and 200-250°E longitude.

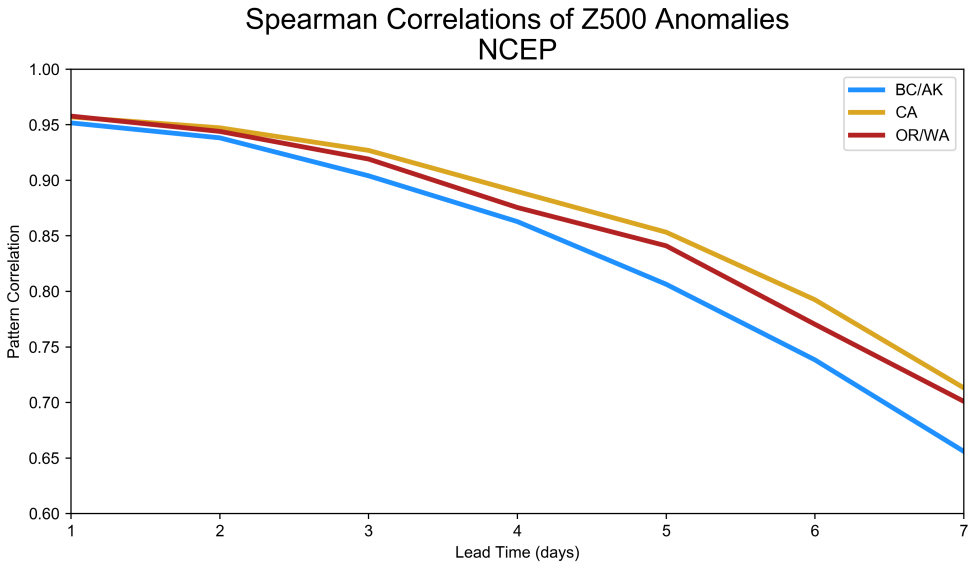

Figure 4.2: As in Figure 4.1, but for NCEP.

Figure 2.3 indicates that more grid cells are in agreement on the sign of the geopotential height anomaly for AR landfall days in CA and OR/WA compared to BC/AK. In a broad sense, this indicates that the conducive height anomaly patterns for CA and OR/WA are more consistent between AR days than for BC/AK. Analysis (not shown) of conducive patterns for each grid cell along the West Coast of North America indicates that the spatial extent of height anomaly agreement is highest for landfalls in northern California and southern Oregon. Farther to the north, only about half as many grid cells are in agreement on AR landfall days. Future studies could examine the relationship between the consistency of upstream geopotential height patterns and the predictability of AR landfalls.

Chapter 5

S2S Forecasting with an Empirical Model

Figures 3.2 and 3.3 indicate that models provide little additional skill at leads greater than 14 days. Thus, we examine the ability of an empirical model from Mundhenk et al. (2018) to predict anomalous weekly AR activity at S2S leads (2-5 weeks). In their study, Mundhenk et al. (2018) measured AR activity in terms of the number of detected AR landfalls within a particular domain. However, counting AR landfalls requires the choice of an AR detection algorithm. Since various detection algorithms exist (e.g. Guan and Waliser, 2015; Lavers et al., 2012; Ralph et al., 2004), output by the model and interpretation by an operational forecaster could vary based on the choice of AR detection algorithm. Instead, here we modify the empirical model to instead predict IVT, a continuous variable with little variation between methods of calculation.

Figure 5.1 shows the empirical model's skill (in terms of HSS multiplied by 100) in predicting IVT along the coast of California. Note that the California domain comes from Figure 1a of Mundhenk et al. (2018) and differs slightly from the CA sub-region defined previously for re-forecasts at 1-14 days. We also examine other domains along the West Coast of North America (Appendix B), but we choose to focus on California here due to its concentration of social and economic interests. Skill is plotted by MJO phase on the vertical axis and forecast lead (center of the 5-day running mean) on the horizontal axis. The panels depict skill for easterly, westerly, and all QBO phases. MJO phase/forecast lead combinations with colored (red or blue) cells have positive skill. White cells have no skill or negative skill. Red colors denote a forecast of below-normal activity, while blue colors denote a forecast of above-normal activity. As in Mundhenk et al. (2018), statistical significance of these positive HSS values is assessed using a bootstrapping procedure (1000 iterations) that randomizes the conditional distributions from which forecasts are made. Consistent with Mundhenk et al. (2018), statistical significance is measured at 80% and 90% confidence levels.

Figure 5.1 should be interpreted as follows: For MJO phase 6 during a westerly QBO, the model forecasts above-normal IVT along the coast of California valid 14-18 days later. Darker shades indicate higher positive skill, while lighter shades indicate lower positive skill. Figure 5.1 implies that the model is only skillful for certain phase/lead combinations. Therefore, the model provides skillful "forecasts of opportunity", meaning that a forecaster can benefit from the model's additional skill when appropriate conditions or "opportunities" arise. Evident in Figure 5.1 is that some phase/lead combinations have statistically-significant predictive skill beyond 3 weeks. Also, the westerly QBO phase exhibits skill in predicting below-normal activity about 1 week after phase 5 and about 4 weeks after phase 1. Such a response to the phase of the MJO is consistent with the time scale of the MJO (approximately 30-90 days). However, such a clear MJO signal is not as apparent for the easterly QBO. This result differs from previous work (e.g. Hendon and Abhik, 2018; Marshall et al., 2017; Son et al., 2017; Yoo and Son, 2016; Zhang and Zhang, 2018), which found that the MJO is stronger during the easterly QBO phase.

We further modify the model to predict anomalous daily total precipitation. Due to the higher resolution of the precipitation dataset, new domains (Figure 5.2) are defined along the West Coast of North America that differ from those used for IVT forecasts. In addition, since precipitation impacts can be felt well inland of the coast (e.g. over the Sierra Nevadas in California), coastal and interior domains are examined separately for California (Appendix B). Figure 5.3 shows the model's skill in predicting anomalous precipitation for the new Coastal California domain. The patterns in skill and dominant response are similar to those seen with IVT forecasts. However, overall model skill appears to be higher for precipitation forecasts. The pattern of skill and dominant response is similar for the Interior California domain (Appendix B). In addition, these results resemble those shown by Mundhenk et al. (2018) for forecasts of AR counts in California.

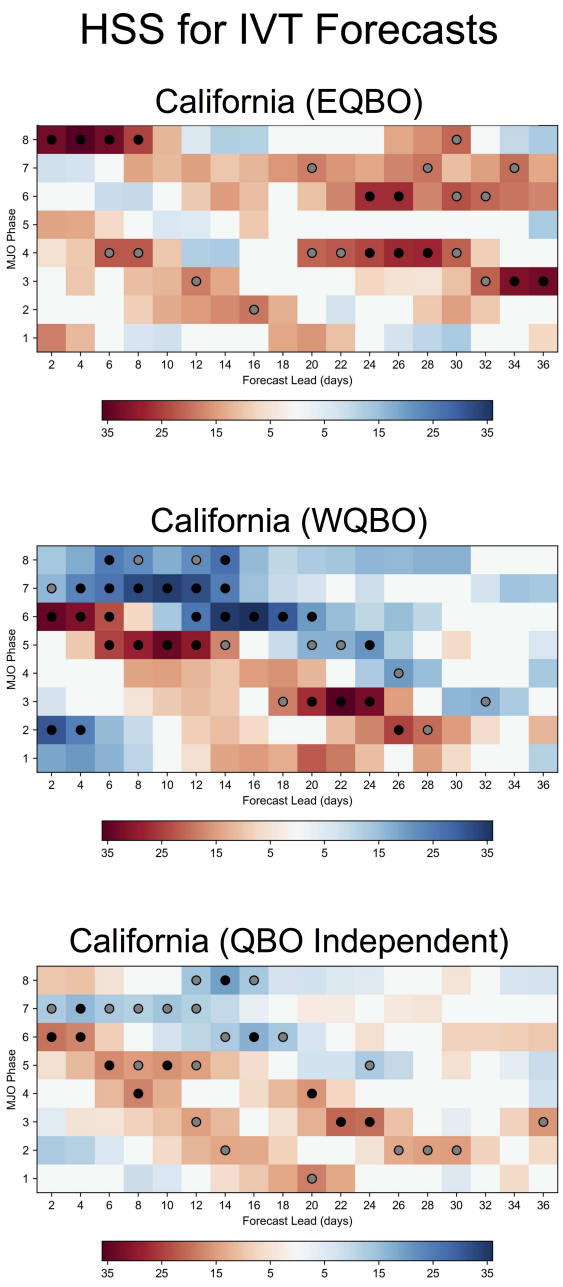

Figure 5.1: Heidke Skill Score (HSS) for IVT forecasts for the original California landfall domain (Mundhenk et al., 2018) during easterly QBO, westerly QBO, and all QBO conditions. Shaded cells indicate positive HSS, while blue (red) shading indicates a forecast of above-normal (below-normal) IVT. Note that the California landfall domain as defined here is different from the domain defined for 1-14 day re-forecasts. For each phase/lead combination, gray dots indicate statistical significance at 80% confidence, while black dots indicate statistical significance at 90% confidence (Mundhenk et al., 2018).

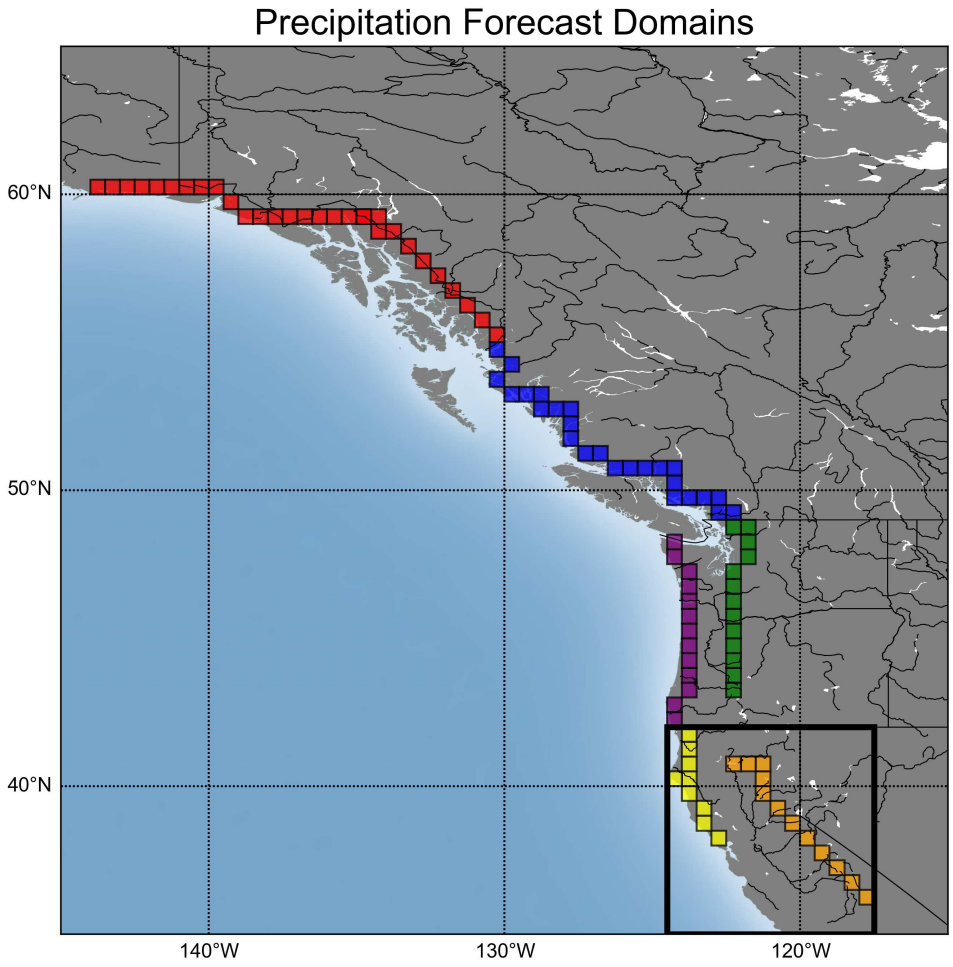

Figure 5.2: Domains for the prediction of anomalous precipitation using the empirical model. The domains of interest here are Coastal California (yellow) and Interior California (orange). We also examine (Appendix B) the Alaska domain (red); the British Columbia domain (blue); the Coastal Pacific Northwest domain (purple); and the Interior Pacific Northwest domain (green).

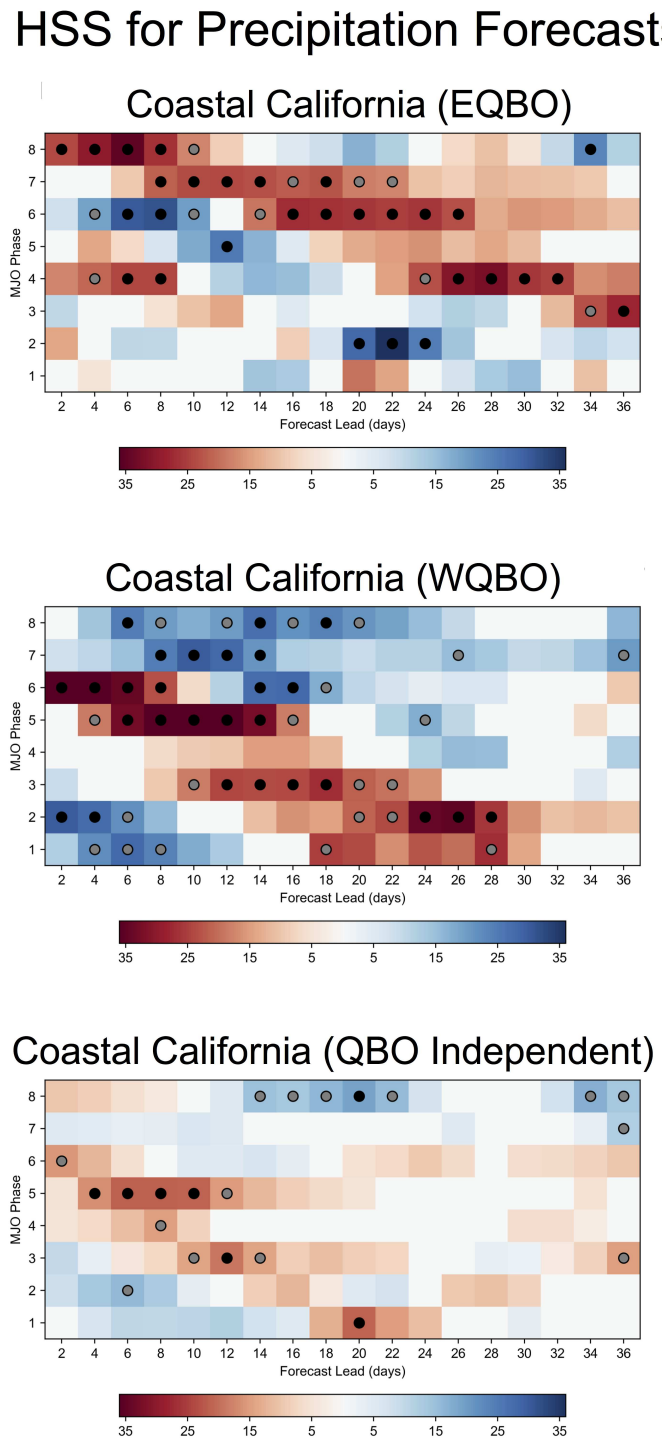

Figure 5.3: As in Figure 5.1, but for precipitation forecasts in the Coastal California domain.

Chapter 6

Discussion and Conclusions

Our study examines re-forecasts of landfalling atmospheric rivers (ARs) along the West Coast of North America from nine numerical weather prediction (NWP) models, each covering about 10-30 years of re-forecasts. In total, our study examines over 8000 re-forecast initializations. For the purposes of model verification, re-forecasts are compared to atmospheric reanalysis data (ERAI), and models are assessed with respect to three components of an effective AR re-forecast: AR occurrence, AR intensity, and AR landfall location.

Occurrence-based re-forecast skill is examined for all nine models in order to determine how often each model correctly re-forecasts the presence (or lack thereof) of a landfalling AR. As lead time increases, occurrence-based skill decreases for all nine models, as seen in prior model verification studies (e.g. DeFlorio et al., 2018; Nayak et al., 2014; Wick et al., 2013). By a lead of 14 days, models generally provide little additional skill (compared to a random forecast), as similarly demonstrated by Nayak et al. (2014) for the central United States and DeFlorio et al. (2018) for the North Pacific and western United States. A novel finding of our study is that re-forecast skill varies by sub-region, with BC/AK having less occurrence-based skill than CA and OR/WA. Our study also shows that ECMWF, on average, predicts fewer AR occurrences than reanalysis at all lead times for the entire West Coast of North America.

Errors in AR intensity re-forecasts are also assessed. For all models, the AR landfall IVT RMSE error stays fairly constant with lead time. These IVT RMSE errors typically range from approximately 100-250 $\mathrm{kg} \cdot \mathrm{m}^{-1} \cdot \mathrm{s}^{-1}$. At leads of 1 through 3 days, positive and negative absolute IVT errors appear to be equally likely, though positive absolute IVT errors appear to be slightly favored at lead times beyond 3 days. In addition, relative (percent) IVT errors across the North Pacific are examined for ECMWF and NCEP. In particular, ECMWF shows a statistically significant low-IVT bias offshore of the West Coast of North America. This low-IVT bias for ECMWF may

explain why, on average, the model consistently predicts fewer AR landfall occurrences compared to reanalysis.

Errors in AR landfall location are also examined. For all nine models, there is a pronounced increase in landfall location RMSE error as lead time increases. At leads of 1 through 3 days, landfall location RMSE errors generally range from about 100 km to 600 km, with errors exceeding 1000 km at leads greater than 14 days. This tendency corresponds well with findings from other model verification studies across the West Coast of North America (Wick et al., 2013) and the central United States (Nayak et al., 2014). In addition, our study quantifies how often each re-forecast is too far south vs. too far north. We find a frequency of southward model landfall location error in CA and OR/WA that is statistically higher than expected from climatology.

Errors in 500-hPa geopotential height re-forecasts are also assessed for ECMWF and NCEP at leads of 1 to 7 days. All three sub-regions have distinct height anomaly patterns that are consistently observed on AR landfall days. We find that the conducive geopotential height anomaly pattern for BC/AK AR landfalls features an anomalous ridge over the Pacific Northwest. This result corresponds with the findings of Mundhenk et al. (2016b), who found that increased AR activity is favored in Alaska when an anomalous 500-hPa ridge is located over the Gulf of Alaska. For both models, skill decreases from Day 1 re-forecasts to Day 7 re-forecasts, and model skill in re-forecasting the conducive height anomaly pattern associated with AR days in BC/AK is consistently lower than for the other two sub-regions.

Finally, we explore the ability of an empirical model (Mundhenk et al., 2018) to predict anomalous weekly AR activity at S2S leads (2-5 weeks) based on the current phase of the MJO and QBO. Unlike in Mundhenk et al. (2018), AR activity is defined in terms of both IVT and daily total precipitation. In this way, we demonstrate an objective method of forecasting the risks and impacts associated with ARs. For the California coast, we show that this model has the potential to provide skillful "forecasts of opportunity" at leads greater than 14 days, the point beyond which dynamical models provide little additional skill. During the westerly phase of the QBO, the response of forecast skill to MJO phase implies a propagation of an MJO signal with a period between 30-90 days.

This pattern of skill and dominant response for coastal California is similar to what is demonstrated for AR count forecasts in Mundhenk et al. (2018).

Our study focuses solely on the control runs of each of the nine models. Since the incorporation of ensemble runs would likely provide additional forecast skill, the results shown in this study likely represent a lower limit in model forecast skill. Future work could incorporate re-forecast output from the available ensemble simulations. Also, the models used in this study vary in terms of time period and initialization frequency. Once available, datasets with models initialized on the same days should be used to best compare and contrast performance among models. In addition, our study does not use multi-day windows for AR landfall occurrence, as described in Wick et al. (2013). They found an improvement in occurrence-based skill when applying 2-day windows for AR landfall occurrence. It is also important to remember that the temporal resolution of the data used here is daily. Therefore, short-duration AR landfalls that occur between daily 0000 UTC time steps are not captured in this study. Also not captured in this analysis are weaker plumes of water vapor transport that fail to reach the chosen IVT threshold of 500 kg $\cdot$ m^{-1} $\cdot$ s^{-1}. Different AR detection algorithms may decrease or increase the total number of landfalling features examined. The spatial resolution of 1.5° x 1.5° also adds uncertainty to the exact locations of AR landfalls.

Nevertheless, our study agrees with the findings of prior AR model verification studies and shows the need for additional improvement of NWP models in the forecasting of landfalling ARs. Our results imply that there exists a link between accurately predicting geopotential height anomalies and accurately predicting AR occurrence. We also demonstrate the potential for predictability of AR activity at subseasonal leads using the MJO and QBO. These results highlight potential avenues of forecast improvement and the extension of forecast skill past the current limits.

Chapter 7

Future Work

Overall, this work demonstrates the continued pitfalls of NWP models at leads of 1-14 days in predicting AR landfalls and suggests possible sources of predictability (e.g. the synoptic-scale flow pattern) in order to improve such forecasts in the future. However, additional work needs to be done in order to further understand the current model limitations and improve performance. As this work builds on the assessments performed by Wick et al. (2013) and Nayak et al. (2014), future studies will be required to continually evaluate the current state of model performance. The analysis of AR re-forecasts outlined here demonstrates the importance of evaluating AR occurrence, intensity, and landfall location. However, future work should examine other geometric characteristics such as AR length, width, and orientation, as these characteristics may influence the degree of human-felt impacts (e.g. flooding).

The relationship between AR characteristics and the synoptic-scale setup (e.g. Hecht and Cordeira, 2017; Mundhenk et al., 2016b) should also be explored further, as an increased understanding of the link between AR activity and synoptic-scale conditions has the potential to lead to increased predictability at short and medium-range leads. The link between ARs and synoptic conditions is especially interesting in the context of differences in AR occurrence-based skill between sub-regions. In Chapter 4, we examine re-forecasts of geopotential height anomalies in order to explain such differences in skill, and we find that geopotential height re-forecasts exhibit a similar difference between sub-regions. Future work should explore this connection further while also considering other potential explanations for the demonstrated sub-regional disparities in skill. For instance, it is hypothesized that IVT errors offshore also adversely impact the occurrence-based skill of AR landfall re-forecasts.

Additional studies should continue to seek sources of predictability of ARs at S2S leads. Our analysis builds on that from others (e.g. Baggett et al., 2017; DeFlorio et al., 2018; Mundhenk et al., 2018) and demonstrates the ability to use climate teleconnection patterns like the MJO and QBO

to better predict AR activity at leads of 2 weeks and beyond. Future studies should specifically continue to explore the link between the MJO and QBO and how that relationship impacts AR predictability. For example, given prior work (e.g. Hendon and Abhik, 2018; Marshall et al., 2017; Son et al., 2017; Yoo and Son, 2016; Zhang and Zhang, 2018) that would suggest the opposite, why does the empirical model applied in this study exhibit a clearer MJO teleconnection pattern during westerly QBO events? Such work will allow for the continued improvement of the empirical model, which has the potential to be incorporated into a suite of S2S forecast tools that can be used by operational forecasters.

Finally, though we focus on landfalling ARs along the West Coast of North America, this analysis can be extended to other regions. Prior work has demonstrated that AR activity and associated extreme precipitation occur across the United States, and the world. Analysis performed here, especially related to forecasts at S2S leads, opens the door for future work to improve AR predictability outside of the western United States. For example, the empirical model could be applied to the Great Plains of the United States, where AR features associated with the low-level jet are capable to producing extreme summertime precipitation (e.g. Nayak et al., 2014). Agricultural and emergency management interests in these locations would likely benefit from additional time in preparation for such extreme precipitation and its impacts.

Our desire is for this work to serve as a bridge to future endeavors that explore ways of increasing predictability of ARs and their impacts. The avenues of future study suggested above represent just a handful of questions that need to be asked in order to improve forecast accuracy and reliability for those who rely on such information to make crucial economic and public-safety decisions.

Bibliography

Baggett, C., E. Barnes, E. Maloney, and B. Mundhenk, 2017: Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. *Geophys. Res. Lett.*, **44**, 7528–7536, doi: 10.1002/2017GL074434.

Baldwin, M., and Coauthors, 2001: The Quasi-Biennial Oscillation. *Rev. Geophys.*, **39**, 179–229.

Bao, J., S. Michelson, P. Neiman, F. Ralph, and J. Wilczak, 2006: Interpretation of enhanced integrated water vapor bands associated with extratropical cyclones: Their formation and connection to tropical moisture. *Mon. Wea. Rev.*, **134**, 1063–1080.

Chen, M., W. Shi, P. Xie, V. Silva, V. Kousky, R. Higgins, and J. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.*, **113**, D04 110, doi:10.1029/2007JD009132.

Dacre, H., P. Clark, O. Martinez-Alvarado, M. Stringer, and D. Lavers, 2015: How do atmospheric rivers form? *Bull. Amer. Meteor. Soc.*, **96**, 1243–1255, doi:10.1175/BAMS-D-14-00031.1.

D'Andrea, F., and Coauthors, 1998: Northern Hemisphere atmospheric blocking as simulated by 15 atmospheric general circulation models in the period 1979-1988. *Climate Dyn.*, **14**, 385–407.

Davini, P., and F. D'Andrea, 2016: Northern hemisphere atmospheric blocking representation in global climate models: Twenty years of improvements? *J. Climate*, **29**, 8823–8840, doi: 10.1175/JCLI-D-16-0242.1.

DeFlorio, M., D. Waliser, B. Guan, D. Lavers, F. Ralph, and F. Vitart, 2018: Global assessment of atmospheric river prediction skill. *J. Hydrometeor.*, doi:10.1175/JHM-D-17-0135.1, in press.

Dettinger, M., F. Ralph, T. Das, P. Neiman, and D. Cayan, 2011: Atmospheric rivers, floods and the water resources of California. *Water*, **3**, 445–478, doi:10.3390/w3020445.

Guan, B., N. Molotch, D. Waliser, E. Fetzer, and P. Neiman, 2010: Extreme snowfall events linked to atmospheric rivers and surface air temperature via satellite measurements. *Geophys. Res. Lett.*, **37**, L20 401, doi:10.1029/2010GL044696.

Guan, B., and D. Waliser, 2015: Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *J. Geophys. Res.*, **120**, 12 514–12 535, doi:10.1002/2015JD024257.

Hamill, T., J. Whitaker, and S. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, 33–46, doi:10.1175/BAMS-87-1-33.

Hatchett, B., S. Burak, J. Rutz, N. Oakley, E. Bair, and M. Kaplan, 2017: Avalanche fatalities during atmospheric river events in the western United States. *J. Hydrometeor.*, **18**, 1359–1374, doi:10.1175/JHM-D-16-0219.1.

Hecht, C., and J. Cordeira, 2017: Characterizing the influence of atmospheric river orientation and intensity on precipitation distributions over north coastal California. *Geophys. Res. Lett.*, **44**, doi:10.1002/2017GL074179.

Henderson, S., E. Maloney, and E. Barnes, 2016: The influence of the Madden-Julian oscillation on Northern Hemisphere winter blocking. *J. Climate*, **29**, 4597–4616, doi:10.1175/JCLI-D-15-0502.1.

Hendon, H., and S. Abhik, 2018: Differences in vertical structure of the Madden-Julian oscillation associated with the quasi-biennial oscillation. *Geophys. Res. Lett.*, **45**, 4419–4428, doi:10.1029/2018GL077207.

Jackson, D., M. Hughes, and G. Wick, 2016: Evaluation of landfalling atmospheric rivers along the U.S. West Coast in reanalysis data sets. *J. Geophys. Res.*, **121**, 2705–2718, doi:10.1002/2015JD024412.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 1st ed., John Wiley and Sons, 254 pp.

Lavers, D., F. Ralph, P. Neiman, G. Wick, C. Scott, D. McCollor, and T. White, 2014: Atmospheric rivers in southeast Alaska and British Columbia: The Bella Coola event of 2010 and Alaska events of 2012. *AGU Fall Meeting Abstracts*.

Lavers, D., and G. Villarini, 2013: The nexus between atmospheric rivers and extreme precipitation across Europe. *Geophys. Res. Lett.*, **40**, 3259–3264, doi:10.1002/grl.50636.

Lavers, D., G. Villarini, R. Allan, E. Wood, and A. Wade, 2012: The detection of atmospheric rivers in atmospheric reanalyses and their links to British winter floods and the large-scale climatic circulation. *J. Geophys. Res.*, **117**, D20 106, doi:10.1029/2012JD018027.

Marshall, A., H. Hendon, S. Son, and Y. Lim, 2017: Impact of the quasi-biennial oscillation on predictability of the Madden-Julian oscillation. *Climate Dyn.*, **49**, 1365–1377, doi:10.1007/s00382-016-3392-0.

Matsueda, M., M. Kyouda, Z. Toth, H. Tanaka, and T. Tsuyuki, 2011: Predictability of an atmospheric blocking event that occurred on 15 December 2005. *Mon. Wea. Rev.*, **139**, 2455–2470, doi:10.1175/2010MWR3551.1.

Mundhenk, B., E. Barnes, and E. Maloney, 2016a: All-season climatology and variability of atmospheric river frequencies over the North Pacific. *J. Climate*, **29**, 4885–4903, doi:10.1175/JCLI-D-15-0655.1.

Mundhenk, B., E. Barnes, E. Maloney, and C. Baggett, 2018: Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden-Julian oscillation and quasi-biennial oscillation. *Nature Clim. and Atmos. Sci.*, **1**, doi:10.1038/s41612-017-0008-2.

Mundhenk, B., E. Barnes, E. Maloney, and K. Nardi, 2016b: Modulation of atmospheric rivers near Alaska and the U.S. West Coast by Northeast Pacific height anomalies. *J. Geophys. Res.*, **121**, 12 751–12 765, doi:10.1002/2016JD025350.

Nardi, K., E. Barnes, and F. Ralph, 2018: Assessment of numerical weather prediction model reforecasts of the occurrence, intensity, and location of atmospheric rivers along the West Coast of North America. *Mon. Wea. Rev.*, doi:10.1175/MWR-D-18-0060.1.

Nayak, M., G. Villarini, and D. Lavers, 2014: On the skill of numerical weather prediction models to forecast atmospheric rivers over the central United States. *Geophys. Res. Lett.*, **41**, 4354–4362, doi:10.1002/2014GL060299.

Neiman, P., F. Ralph, A. White, D. Kingsmill, and P. Persson, 2002: The statistical relationship between upslope flow and rainfall in California's coastal mountains: Observations during CALJET. *Mon. Wea. Rev.*, **130**, 1468–1492.

Neiman, P., F. Ralph, and G. Wick, 2008: Meteorological characteristics and overland precipitation impacts of atmospheric rivers affecting the West Coast of North America based on eight years of SSM/I satellite observations. *J. Hydrometeor.*, **9**, 22–47, doi:10.1175/2007JHM855.1.

Neiman, P., L. Schick, F. Ralph, M. Hughes, and G. Wick, 2011: Flooding in western Washington: The connection to atmospheric rivers. *J. Hydrometeor.*, **12**, 1337–1358, doi:10.1175/2011JHM1358.1.

Palmer, T., F. Doblas-Reyes, A. Weisheimer, and M. Rodwell, 2008: Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bull. Amer. Meteor. Soc.*, **89**, 459–470, doi:10.1175/BAMS-89-4-459.

Ralph, F., and M. Dettinger, 2012: Historical and national perspectives on extreme West Coast precipitation associated with atmospheric rivers during December 2010. *Bull. Amer. Meteor. Soc.*, **93**, 783–790, doi:10.1175/BAMS-D-11-00188.1.

Ralph, F., and T. Galarneau, 2017: The Chiricahua Gap and the role of easterly water vapor transport in southeastern Arizona monsoon precipitation. *J. Hydrometeor.*, **18**, 2511–2520, doi:10.1175/JHM-D-17-0031.1.

Ralph, F., P. Neiman, and G. Wick, 2004: Satellite and CALJET aircraft observations of atmospheric rivers over the eastern North Pacific Ocean during the winter of 1997/1998. *Mon. Wea. Rev.*, **132**, 1721–1745.

Ralph, F., P. Neiman, G. Wick, S. Gutman, M. Dettinger, D. Cayan, and A. White, 2006: Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, **33**, L13 801, doi:10.1029/2006GL026689.

Ralph, F., E. Sukovich, D. Reynolds, M. Dettinger, S. Weagle, W. Clark, and P. Neiman, 2010: Assessment of extreme quantitative precipitation forecasts and development of regional extreme event thresholds using data from HMT-2006 and COOP observers. *J. Hydrometeor.*, **11**, 1286–1304, doi:10.1175/2010JHM1232.1.

Ralph, F., and Coauthors, 2017: Dropsonde observations of total integrated water vapor transport within North Pacific atmospheric rivers. *J. Hydrometeor.*, **18**, 2577–2596, doi:10.1175/JHM-D-17-0036.1.

Rivera, E., F. Dominguez, and C. Castro, 2014: Atmospheric rivers and cool season extreme precipitation events in the Verde River Basin of Arizona. *J. Hydrometeor.*, **15**, 813–829, doi:10.1175/JHM-D-12-0189.1.

Rutz, J., and W. Steenburgh, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921, doi:10.1175/MWR-D-13-00168.1.

Son, S., Y. Lim, C. Yoo, H. Hendon, and J. Kim, 2017: Stratospheric control of the Madden-Julian oscillation. *J. Climate*, **30**, 1909–1922, doi:10.1175/JCLI-D-16-0620.1.

Trenberth, K., 1997: The definition of El Niño. *Bull. Amer. Meteor. Soc.*, **78**, 2771–2777.

Tseng, K., E. Barnes, and E. Maloney, 2017: Prediction of the midlatitude response to strong Madden-Julian oscillation events on s2s time scales. *Geophys. Res. Lett.*, **45**, 463–470, doi:10.1002/2017GL075734.

Waliser, D., and B. Guan, 2017: Extreme winds and precipitation during landfall of atmospheric rivers. *Nature*, **10**, 179–183, doi:10.1038/ngeo2894.

Wick, G., P. Neiman, F. Ralph, and T. Hamill, 2013: Evaluation of forecasts of the water vapor signature of atmospheric rivers in operational numerical weather prediction models. *Wea. Forecasting*, **28**, 1337–1352, doi:10.1175/WAF-D-13-00025.1.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed., Elsevier Academic Press, 648 pp.

Xie, P., A. Yatagai, M. Chen, T. Hayasaka, Y. Fukushima, C. Liu, and S. Yang, 2007: A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeor.*, **8**, 607–626, doi:10.1175/JHM583.1.

Yoo, C., and S. Son, 2016: Modulation of the boreal wintertime Madden-Julian oscillation by the stratospheric quasi-biennial oscillation. *Geophys. Res. Lett.*, **43**, 1392–1398, doi:10.1002/2016GL067762.

Zhang, C., 2005: Madden-Julian Oscillation. *Rev. Geophys.*, **43**, RG2003, doi:10.1029/2004RG000158.

Zhang, C., 2013: Madden-Julian Oscillation: Bridging Weather and Climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, doi:10.1175/BAMS-D-12-00026.1.

Zhang, C., and B. Zhang, 2018: QBO-MJO connection. *J. Geophys. Res.*, **123**, 2957–2967, doi:10.1002/2017JD028171.

Zhou, Y., and H. Kim, 2017: Prediction of atmospheric rivers over the North Pacific and its connection to ENSO in the North American multi-model ensemble (NMME). *Climate Dyn.*, doi:10.1007/s00382-017-3973-6, [Available online at https://doi.org/10.1007/s00382-017-3973-6].

Zhu, Y., and R. Newell, 1994: Atmospheric rivers and bombs. *Geophys. Res. Lett.*, **21**, 1999–2002, doi:10.1029/94GL01710.

Zhu, Y., and R. Newell, 1998: A proposed algorithm for moisture fluxes from atmospheric rivers. *Mon. Wea. Rev.*, **126**, 725–735.

Appendix A

Atmospheric River Detection Algorithm

The atmospheric river (AR) detection algorithm used in this study is an updated, generalized version of the algorithm introduced in Mundhenk et al. (2016a). A brief summary of the updated algorithm, as applied to our study, follows below. For further details about the original algorithm, please reference Mundhenk et al. (2016a). The algorithm is generalized to run on data of different spatial and temporal resolutions (for our study, 1.5° by 1.5° spatial resolution and daily temporal resolution). The algorithm scans fields of Integrated Water Vapor Transport (IVT) and detects grid cells (candidate objects) at which the intensity threshold is met or exceeded. The algorithm uses a constant AR intensity threshold of 500 kg $\cdot$ m^{-1} $\cdot$ s^{-1} of full IVT, and there is no criteria for mean intensity across the candidate object. Candidate objects are then put through various geometric tests in order to isolate narrow corridors of high water vapor transport. Geometric criteria consider feature area, length, aspect ratio, eccentricity, and origin. The candidate object must have an area greater than 300,000 km^2, and the object's length must be greater than 1400 km. The candidate object's aspect ratio (i.e. the ratio between length and width) must be greater than 1.4. Candidate objects with centroids below 16° N in latitude that have a mean intensity greater than 750 kg $\cdot$ m^{-1} $\cdot$ s^{-1} are considered to be of tropical origin. These tropical candidate objects are then tested for such criteria as large spread in wind direction and the presence of a tropical cyclone eye-hole, for example. If such criteria are met, then the tropical candidate objects are excluded. In addition, this version of the algorithm does not segment candidate objects with multiple intensity peaks, as done in Mundhenk et al. (2016a). Candidate objects that pass through the intensity and geometry criteria are considered AR features.

Appendix B

Supplemental Materials

B.1 Coordinates of landfall sub-regions

This supplementary materials section lists the latitude/longitude locations of the grid cells within each sub-region of the landfall domain. There are three sub-regions (CA, OR/WA, and BC/AK) that contain six grid cells each. In total, there are 18 grid cells within the landfall domain. The grid cell locations (latitude, longitude) for each sub-region are listed below. All latitudes are in degrees north, and all longitudes are in degrees east:

$$\begin{aligned} CA &= [(33.0, 240.0), (34.5, 238.5), (36.0, 237.0), (37.5, 235.5), (39.0, 234.0), (40.5, 234.0)] \\ OR/WA &= [(42.0, 234.0), (43.5, 234.0), (45.0, 234.0), (46.5, 234.0), (48.0, 232.5), (49.5, 231.0)] \\ BC/AK &= [(51.0, 229.5), (51.0, 228.0), (52.5, 226.5), (54.0, 225.0), (55.5, 223.5), (57.0, 222.0)] \end{aligned} \tag{B.1}$$

B.2 Skill metrics by model at selected leads

This supplementary materials section lists numerical values of Hit Rate (H), False Alarm Rate (F), and Peirce Skill Score (PSS) for each model at selected leads (see Table B.1). These values correspond to the points plotted in Figure 3.2

B.3 Landfall IVT error for other sub-regions

This supplementary materials section contains plots showing the distribution of landfall Integrated Water Vapor Transport (IVT) error for the CA and BC/AK sub-regions. As seen in OR/WA, there are roughly equal probabilities of positive and negative landfall IVT error at leads of 1 through 3 days. Positive IVT error is generally favored at later lead times.

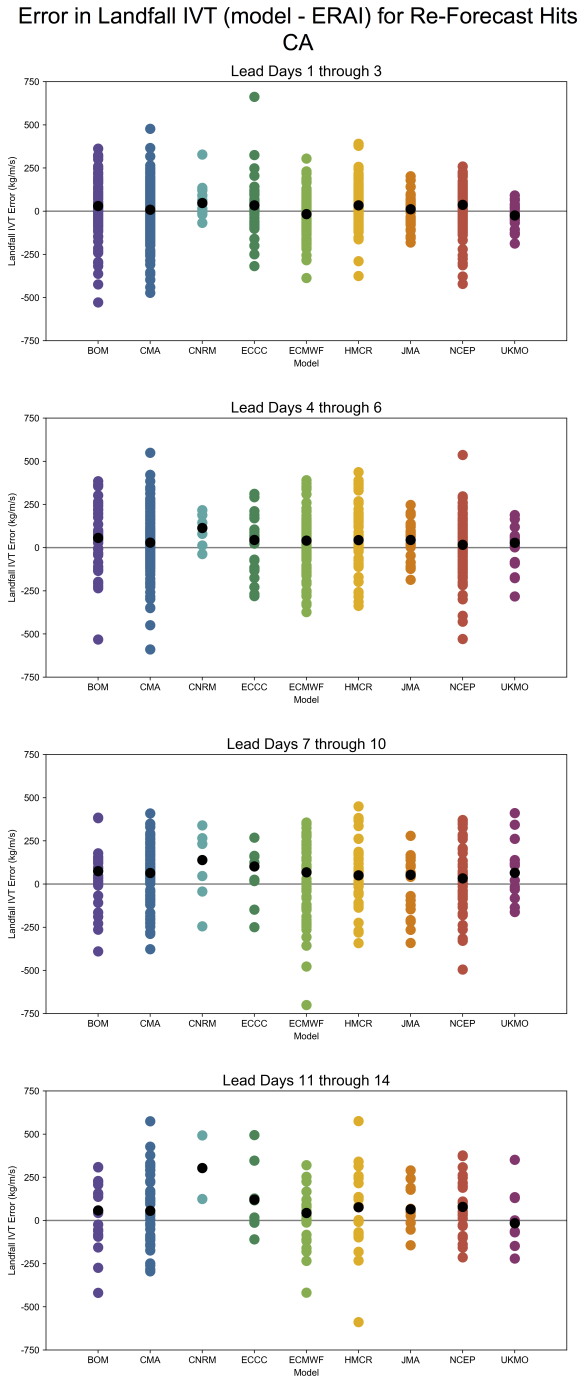

Figure B.1: Distribution of error in landfall IVT for the model and ERA-Interim (model - ERAI) for re-forecast hits in CA during four different ranges of lead time (1-3 days, 4-6 days, 7-10 days, and 11-14 days). Black dots denote the median of the distribution.

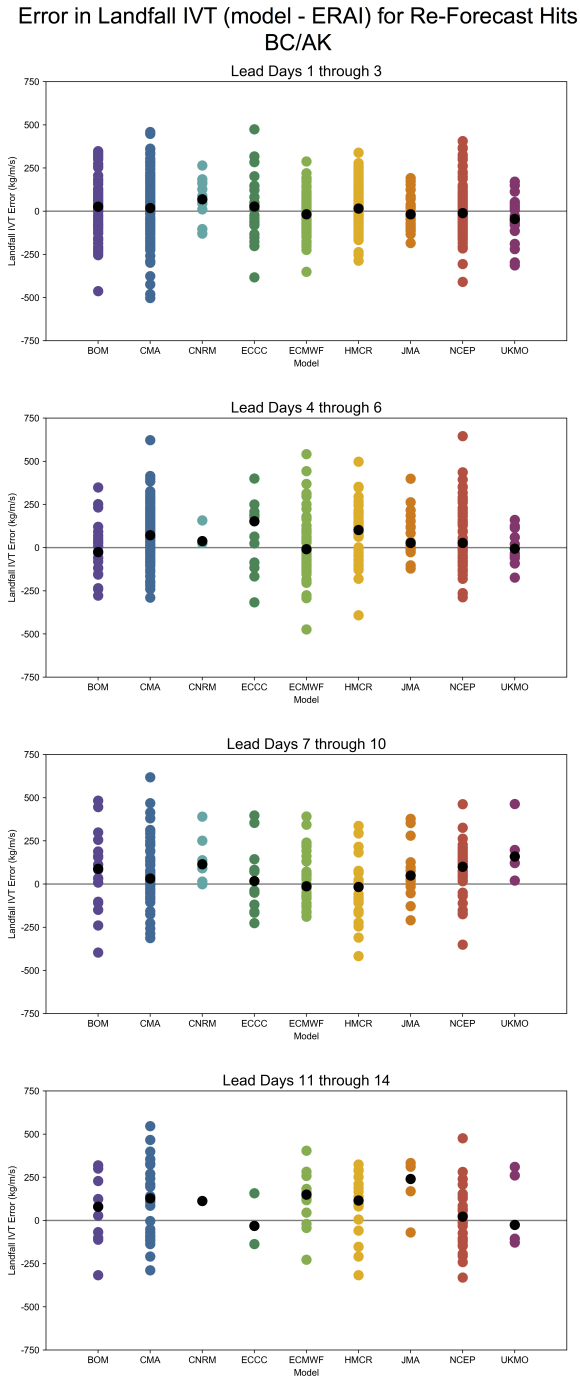

Figure B.2: As in B.1, but for BC/AK.

B.4 S2S empirical model results for other sub-regions

This supplementary materials section shows the empirical model's skill in predicting anomalous precipitation and IVT for other selected domains along the West Coast of North America. For reference, see Figure 5.2 for the precipitation domain locations and Figure 1a of Mundhenk et al. (2018) for the IVT domain locations. Overall, there is less skill for the other domains compared to California for both precipitation and IVT (see Chapter 5). However, as seen in Chapter 5, the empirical model exhibits more skill in predicting anomalous precipitation compared to IVT. Figure B.3 shows similar patterns in precipitation forecast skill between Interior California and Coastal California, while the Coastal Pacific Northwest domain has a similar pattern to the California domains, especially during westerly QBO (Figure B.5). In addition, Alaska (Figure B.4) exhibits a pattern in precipitation response and skill that appears to be roughly opposite that of California during westerly QBO. Such a result is interesting given the modulation in AR activity demonstrated by Mundhenk et al. (2016b) between Alaska and California.

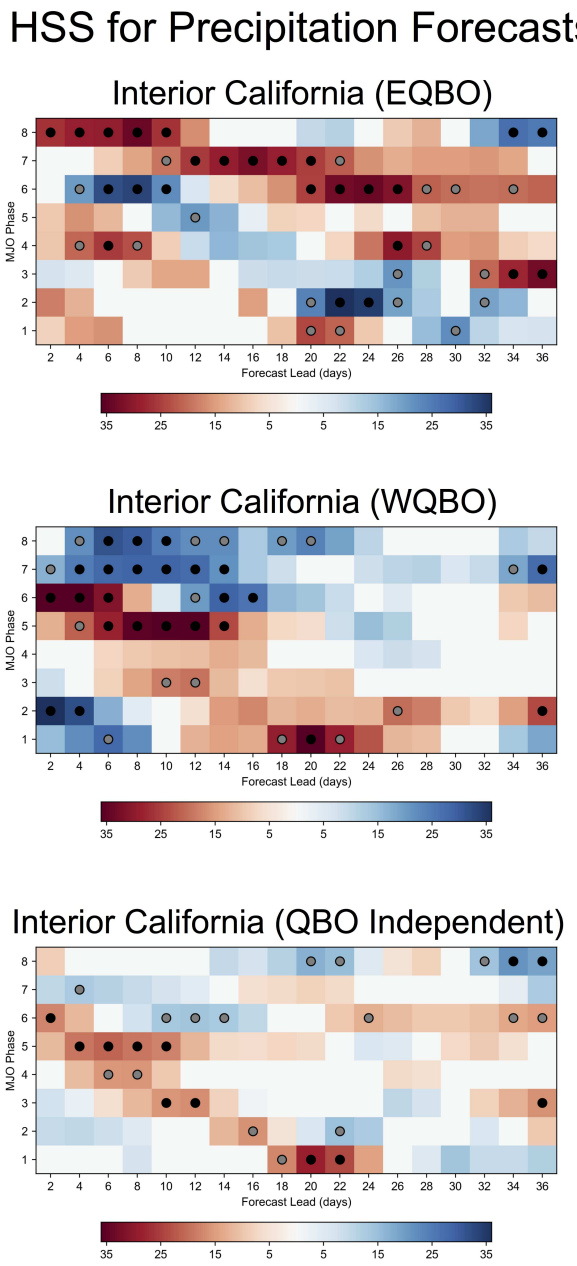

Figure B.3: Heidke Skill Score (HSS) for precipitation forecasts for the Interior California landfall domain (Mundhenk et al., 2018) during easterly QBO, westerly QBO, and all QBO conditions. Shaded cells indicate positive HSS, while blue (red) shading indicates a forecast of above-normal (below-normal) IVT. For each phase/lead combination, gray dots indicate statistical significance at 80% confidence, while black dots indicate statistical significance at 90% confidence (Mundhenk et al., 2018).

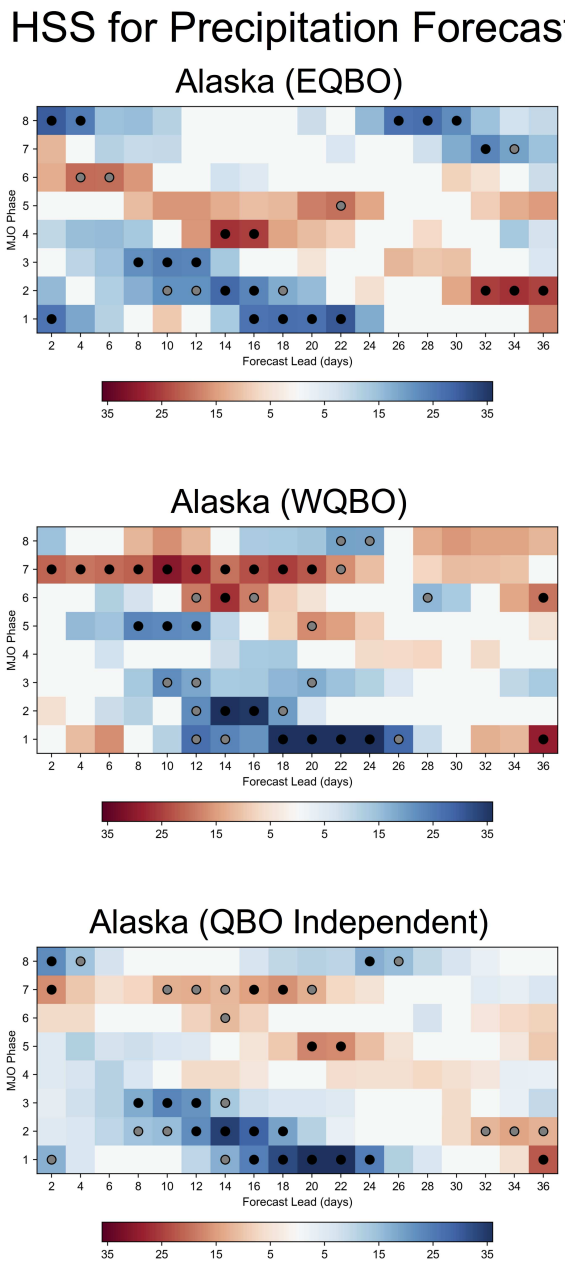

Figure B.4: As in B.3, but for Alaska.

Table B.1: Skill metrics for all nine models at selected lead times (in days). Skill metrics are rounded to 3 decimal places. These skill metrics correspond to the data points plotted in Figure 3.2.

Modeling Center	Re-Forecast Lead	Hit Rate (H)	False Alarm Rate (F)	Peirce Skill Score (PSS)
BOM	1	0.564	0.013	0.552
	2	0.422	0.017	0.405
	3	0.404	0.021	0.383
	4	0.276	0.022	0.254
	5	0.187	0.025	0.162
	6	0.152	0.021	0.130
	7	0.113	0.025	0.088
	10	0.093	0.030	0.063
	14	0.110	0.045	0.065
CMA	1	0.386	0.005	0.382
	2	0.337	0.011	0.326
	3	0.351	0.023	0.328
	4	0.300	0.026	0.275
	5	0.208	0.027	0.181
	6	0.161	0.029	0.133
	7	0.126	0.035	0.091
	10	0.093	0.037	0.057
	14	0.063	0.036	0.027
CNRM	1	0.834	0.011	0.823
	2	0.628	0.015	0.613
	3	0.583	0.029	0.554
	4	0.324	0.039	0.285
	5	0.234	0.022	0.212
	6	0.428	0.036	0.392
	7	0.096	0.036	0.060
	10	0.235	0.051	0.183
	14	0.017	0.034	-0.017
ECCC	1	0.486	0.005	0.481
	2	0.521	0.010	0.511
	3	0.458	0.013	0.445
	4	0.372	0.017	0.355
	5	0.208	0.017	0.190
	6	0.204	0.026	0.178
	7	0.151	0.025	0.126
	10	0.077	0.026	0.051
	14	0.064	0.039	0.026

Modeling Center	Re-Forecast Lead	Hit Rate (H)	False Alarm Rate (F)	Peirce Skill Score (PSS)
ECMWF	1	0.656	0.004	0.651
	2	0.615	0.006	0.609
	3	0.474	0.008	0.466
	4	0.355	0.012	0.343
	5	0.316	0.013	0.303
	6	0.239	0.015	0.225
	7	0.121	0.013	0.108
	10	0.070	0.014	0.056
	14	0.038	0.017	0.022
HMCR	1	0.698	0.010	0.688
	2	0.702	0.022	0.680
	3	0.492	0.023	0.470
	4	0.422	0.030	0.392
	5	0.288	0.032	0.257
	6	0.254	0.031	0.223
	7	0.201	0.038	0.163
	10	0.076	0.044	0.032
	14	0.030	0.042	-0.012
JMA	1	0.839	0.003	0.836
	2	0.751	0.009	0.742
	3	0.654	0.014	0.640
	4	0.534	0.018	0.517
	5	0.451	0.020	0.430
	6	0.284	0.023	0.262
	7	0.190	0.024	0.166
	10	0.065	0.026	0.039
	14	0.057	0.021	0.037
NCEP	1	0.705	0.008	0.697
	2	0.607	0.011	0.596
	3	0.520	0.013	0.507
	4	0.401	0.019	0.382
	5	0.307	0.019	0.289
	6	0.270	0.020	0.250
	7	0.154	0.022	0.132
	10	0.085	0.026	0.060
	14	0.078	0.026	0.052
UKMO	1	0.673	0.002	0.671
	2	0.551	0.006	0.546
	3	0.433	0.010	0.423
	4	0.289	0.016	0.274
	5	0.240	0.015	0.225
	6	0.244	0.014	0.230
	7	0.222	0.019	0.203
	10	0.034	0.019	0.014
	14	0.069	0.020	0.049

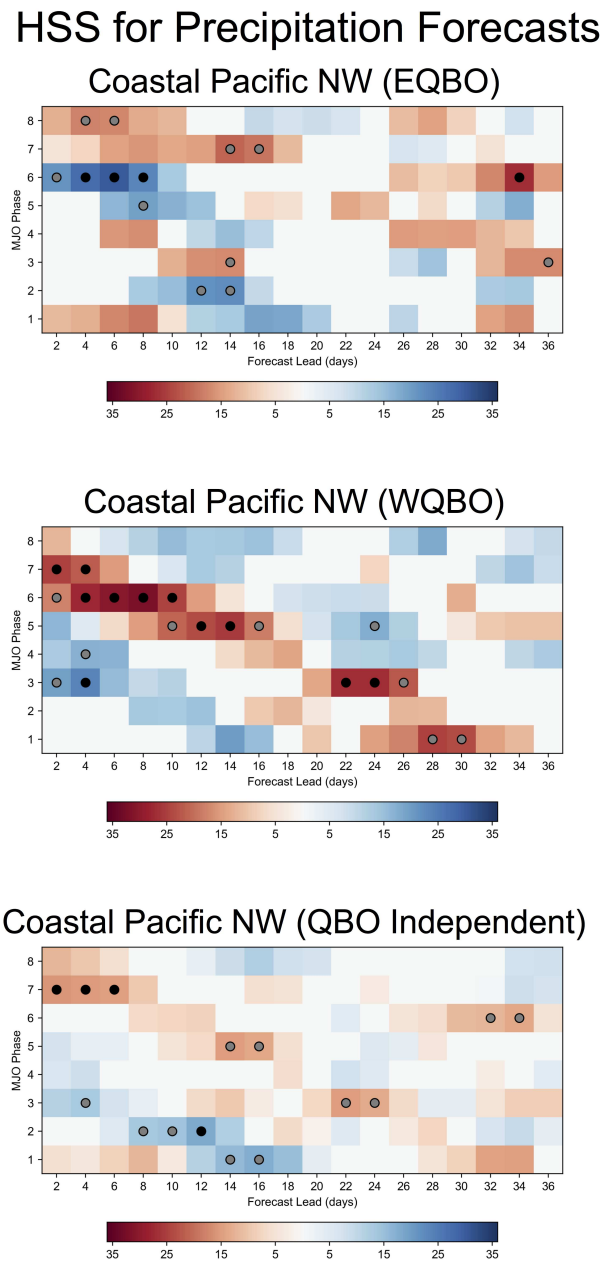

Figure B.5: As in B.3, but for the Coastal Pacific Northwest.

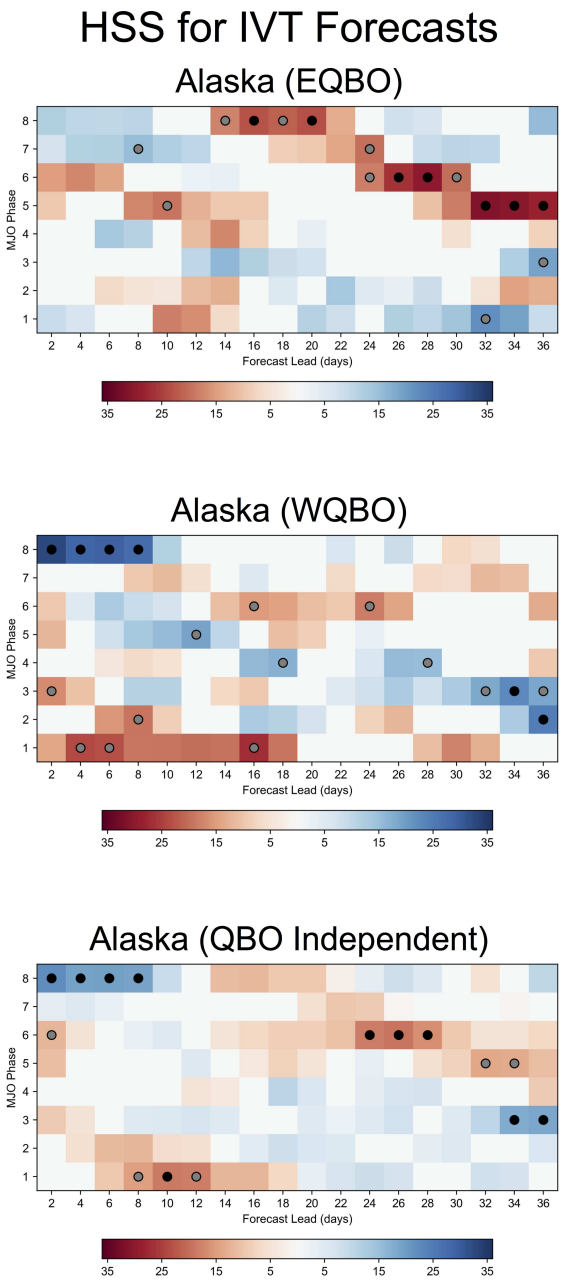

Figure B.6: As in B.3, but for IVT forecasts for Alaska.

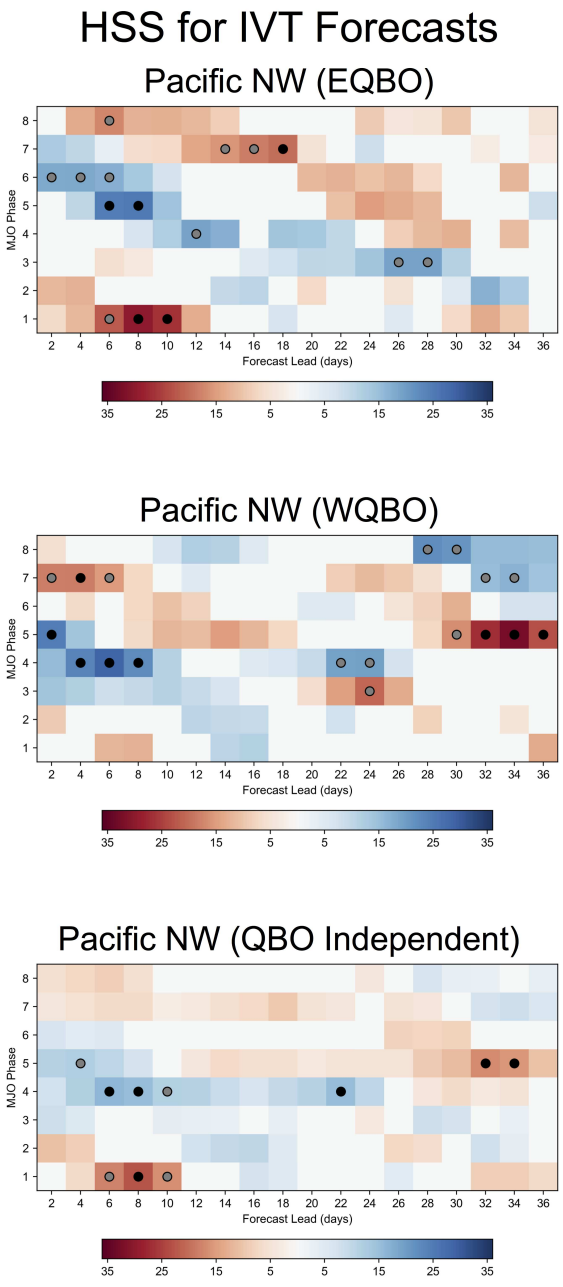

Figure B.7: As in B.3, but for IVT forecasts for the Pacific Northwest