Assessment of Numerical Weather Prediction Model Reforecasts of the Occurrence, Intensity, and Location of Atmospheric Rivers along the West Coast of North America

KYLE M. NARDI AND ELIZABETH A. BARNES

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

F. MARTIN RALPH

Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

(Manuscript received 14 February 2018, in final form 26 July 2018)

ABSTRACT

Atmospheric rivers (ARs)—narrow corridors of high atmospheric water vapor transport—occur globally and are associated with flooding and maintenance of the 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 atmospheric rivers, these studies only cover several years of reforecasts 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 reforecasts from the control runs of 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. 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 overprediction of IVT is common at later lead times. Finally, we show that northward landfall location errors are favored in California, Oregon, and Washington, although southward errors occur more often than expected from climatology. These results highlight the need for model improvements, while helping to identify factors that cause model errors.

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., Ralph et al. 2004; Bao et al. 2006; Ralph et al. 2017b), 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). Because of 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 2005/06 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

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/MWR-D-18-0060.s1>.

Corresponding author: Kyle M. Nardi, knardi@rams.colostate.edu

average, generated about 4 times the daily snow water equivalent accumulation compared to non-AR events over the Sierra Nevada from 2004 to 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., [Ralph et al. 2006](#); [Neiman et al. 2008](#); [Ralph et al. 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., [Rivera et al. 2014](#); [Ralph and Galarneau 2017](#)) 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% and 65% of avalanche fatalities at various locations in the western United States coincided with AR conditions.

Since ARs are high-impact phenomena that affect 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 reforecasts from five dynamical models along the West Coast of the United States during three cool seasons from 2008–09 to 2010–11. These reforecasts came from The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) dataset. The reforecasts 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 reforecasting the occurrence of landfalling ARs in the western United States. Correlations between reforecast 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 reforecast 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. \(2018a\)](#) examined European Centre for Medium-Range Weather Forecasts (ECMWF) AR reforecasts at leads of 1–14 days and found a similar decrease in prediction skill past 10 days over the North Pacific and western United States.

Here, we expand on prior model verification studies by examining reforecasts (also known as hindcasts) of landfalling wintertime ARs from the control runs of nine state-of-the-art weather models at leads of 1–14 days. Reforecasts are retrospective forecasts from a particular numerical model for dates in the past. Reforecasts allow for comparisons to reanalysis in order to assess model performance ([Hamill et al. 2006](#)). Reforecast 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 reforecast (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.

2. Data and methods

a. S2S database

To assess the performance of modern operational weather prediction models in the prediction of ARs, reforecasts provided by the Subseasonal to Seasonal (S2S) International Project database ([Vitart et al. 2017](#)) are used. From the S2S database, control runs from nine different models are analyzed at leads of 1–14 days ([Table 1](#)). For all models, control-run reforecast fields of specific humidity and horizontal wind are analyzed at a horizontal resolution of 1.5° latitude $\times$ 1.5° longitude and at constant pressure levels of 1000, 850, 700, 500, 300, 200, 100, 50, and 10 hPa. For a given initialization, reforecasts are valid at 0000 UTC each day. Reforecast data from four of the models (BOM, ECCO, JMA, and

TABLE 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 center	Initialization years	No. of initializations in NDJF	Missing day 0?
Bureau of Meteorology (BOM)	1981–2013	792	Yes
China Meteorological Administration (CMA)	1994–2014	2520	No
National Centre for Meteorological Research (CNRM)	1994–2006	100	No
Environment and Climate Change Canada (ECCC)	1995–2014	340	Yes
European Centre for Medium-Range Weather Forecasts (ECMWF)	1995–2016	1380	No
Hydrometcentre of Russia (HMCR)	1985–2010	858	No
Japan Meteorological Agency (JMA)	1981–2010	360	Yes
National Centers for Environmental Prediction (NCEP)	1999–2010	1439	Yes
Met Office (UKMO)	1996–2009	224	No

NCEP) do not include output for day 0 (i.e., the initialization time). As seen in Table 1 and Vitart et al. (2017), the temporal range, frequency of initializations, and number of ensembles vary by model. While these differences between models are important to consider, prior model verification studies (e.g., Jie et al. 2017; Schiraldi and Roundy 2017; Vitart 2017) applied this same database to examine forecast skill for multiple models. In fact, Jie et al. (2017) and Vitart (2017) demonstrated that the overall model datasets have enough similarities to make such analysis possible. Therefore, we feel comfortable that the database used here is sufficient for our study. However, given the differences between the models and our analysis of control runs only, as done in Wick et al. (2013), these results should not be interpreted as a comparison of the utility of the nine model forecast systems.

b. Reanalysis data

As a means of providing verification for the reforecast models, ERA-Interim (ERA-I) 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., Lavers et al. 2012; Lavers and Villarini 2013; Guan and Waliser 2015; Mundhenk et al. 2016a) and model verification (e.g., Nayak et al. 2014; Baggett et al. 2017; DeFlorio et al. 2018a) studies of AR activity. Furthermore, prior studies (e.g., Lavers et al. 2012; Jackson et al. 2016) found fairly good agreement between different reanalysis products in the depiction of ARs. Reanalysis is chosen over satellite data for several reasons. First, Wick et al. (2013) noted that it is difficult to accurately depict low-level water vapor transport using satellite retrievals. As a result, satellite-based studies such as Wick et al. (2013) required the use of integrated water vapor (IWV), while we wish to analyze integrated water vapor transport (IVT), a combination of water vapor content and transport by the tropospheric winds. Second, there may exist differences in valid time between available

satellite observations and reforecasts. Even though reforecasts are always valid at 0000 UTC, the nearest satellite observation may occur up to several hours before or after 0000 UTC (Wick et al. 2013). Given that an AR lasts about 24 h along the coast on average (Ralph et al. 2013), an offset of several hours is appreciable. Thus, reanalysis is chosen for this study. Instantaneous fields of specific humidity and horizontal wind from reanalysis data are analyzed at a horizontal resolution (viz., $1.5^\circ \times 1.5^\circ$) and vertical resolution that matches the reforecast model data described above. In addition, the reanalysis data have a similar temporal resolution of 24 h and are also valid at 0000 UTC for a 38-yr period spanning from 1 January 1979 to 31 December 2016.

c. AR detection

The aforementioned isobaric fields of specific humidity and horizontal wind are used to calculate 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):

$$\text{IVT} = \sqrt{\left(\frac{1}{g} \int_{1000}^{300} qu \, dp\right)^2 + \left(\frac{1}{g} \int_{1000}^{300} qv \, dp\right)^2}. \quad (1)$$

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 reforecast 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 $500 \text{ kg m}^{-1} \text{ s}^{-1}$, compared to the IVT anomaly threshold of $250 \text{ kg m}^{-1} \text{ s}^{-1}$ used in Mundhenk et al. (2016a). A threshold that incorporates 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. 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 absolute IVT to define ARs in this study. Here $500 \text{ kg m}^{-1} \text{ s}^{-1}$ is chosen because we aim to isolate moderate to strong ARs. This threshold is recognized as a lower limit for moderate to strong ARs, which account for a large portion of hazards associated with ARs (Ralph et al. 2017a). To test the robustness of our conclusions to this threshold, we additionally analyze model skill using detection thresholds of $400 \text{ kg m}^{-1} \text{ s}^{-1}$ and $600 \text{ kg m}^{-1} \text{ s}^{-1}$. The conclusions are robust across these three thresholds, so we focus on the results using $500 \text{ kg m}^{-1} \text{ s}^{-1}$ for the rest of this study. However, results for the other two thresholds are shown in the online supplemental material.

Figure 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, plume-like corridors of high water vapor transport. A detailed description of all modifications to the Mundhenk et al. (2016a) algorithm can be found in the appendix. 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).

d. Landfall region

Here, we focus our evaluation of AR reforecasts on those features that make landfall along the west coast of North America (Fig. 2). This landfall region comprises 18 (total) $1.5^\circ \times 1.5^\circ$ grid cells that lie just offshore of North America, from the vicinity of Santa Barbara, California, northward to near Juneau, Alaska. 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 nonoccurrences) 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

FIG. 1. Integrated water vapor transport at 0000 UTC 14 Oct 2016. Red contours outline two distinct atmospheric river (AR) features detected by the AR detection algorithm.

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 reforecast data, the initialization date and reforecast 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 shows the NDJF climatology of AR occurrence based on the algorithm run on ERAI data from 1979 to 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).

e. 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 reforecast 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 composing the landfall region by comparing each reforecast to reanalysis on the valid day. This model verification of a discrete binary variable allows for the use of a four-outcome (2×2) contingency table as described in Wilks (2006). The first outcome, denoted here as "a," corresponds to a situation in which the model correctly reforecasts an AR for a given valid day. These outcomes are considered "hits." The second outcome, "b," corresponds to a situation in which the model reforecasts 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,"

FIG. 2. The domain used for analysis of reforecasts of AR occurrence. The colored arrows refer to three subregions chosen for the analysis. Shading denotes the climatological number of AR days per NDJF season based on ERA-Interim reanalysis from 1979 to 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).

corresponds to a situation in which the model does not reforecast 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 reforecast 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., Ralph et al. 2010; Wick et al. 2013; Nayak et al. 2014; DeFlorio et al. 2018a).

Though various contingency-table-based skill metrics exist, this study focuses on four particular metrics. First, frequency bias (referenced here as “bias” or “*B*”) is the number of AR occurrences reforecast by the model divided by the number of AR occurrences detected in reanalysis (in the above terminology, hits + false alarms divided by hits + misses):

$$B = \frac{a + b}{a + c} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}. \quad (2)$$

In general, bias gives a sense of how much a particular model favors reforecasting ARs compared to reanalysis. Bias can range from 0 to infinity, with bias greater than 1 implying that the model reforecasts more ARs than reanalysis and bias less than 1 implying that the model reforecasts fewer ARs than reanalysis. Next, hit rate (*H*)

is the frequency with which the model reforecasts 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 reforecasts 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):

$$H = \frac{a}{a+c} = \frac{\text{hits}}{\text{hits} + \text{misses}},$$

$$F = \frac{b}{b+d} = \frac{\text{false alarms}}{\text{false alarms} + \text{correct rejections}}. \quad (3)$$

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 nonoccurrences.

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:

$$\text{PSS} = H - F$$

$$= \frac{\frac{a+d}{n} - \left[\frac{(a+b)(a+c) + (b+d)(c+d)}{n^2} \right]}{1 - \left[\frac{(a+c)^2 + (b+d)^2}{n^2} \right]}. \quad (4)$$

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 reforecast 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 subregions defined in Fig. 2. One caveat with using PSS here is that since landfalling ARs, as defined by the detection algorithm,

are relatively infrequent events (Fig. 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" reforecasts (Jolliffe and Stephenson 2003). Because of 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.

f. Assessment of intensity and location reforecasts

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 reforecast (i.e., hits). In other words, IVT errors are examined for reforecasts that accurately predict the presence of AR conditions as defined by our algorithm. For each subregion, all reforecast hits for the individual grid cells within the subregion are analyzed. Absolute IVT error is calculated by subtracting the reanalysis IVT from the reforecast 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 subregion, we examine reforecasts for identified AR days and compare the landfall location from the model to that from reanalysis. If multiple AR features are reforecast by the model, the feature closest to the reanalysis landfall location is used and compared to reanalysis. Once the reforecast and reanalysis AR features are identified, the landfall location error is defined as the distance between the reforecast and reanalysis landfall locations along a great circle.

3. Results

a. AR occurrence

Our evaluation of model reforecast 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 reforecasts: hits, false alarms, misses, and correct rejections. Tallies of these outcomes are used to calculate four skill metrics: bias

FIG. 3. Model bias (ratio of AR occurrence reforecasts to AR observations) plotted as a function of model and reforecast lead time (in days). Reforecast leads for which there are no data are shaded gray. Red colors indicate fewer AR occurrences reforecast than observed in reanalysis data, and blue colors indicate more AR occurrences reforecast than observed in reanalysis data.

(B), hit rate (H), false alarm rate (F), and Peirce skill score (PSS).

Figure 3 shows bias for each of the nine models by reforecast 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. Recall that this analysis is only done for the control runs of each model. Blue colors indicate more AR occurrences in the reforecasts compared to AR occurrences in reanalysis (“active bias”), while red colors indicate fewer AR occurrences reforecast 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 10–14 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 reforecast leads. Overall, Fig. 3 provides a sense of a particular model’s propensity to reforecast landfalling ARs along the west coast of North America.

Though useful for an overall picture of how often a model reforecasts 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 reforecasting an AR when one does not actually occur (i.e., F), as well as the frequency of

correctly reforecasting an AR when one actually does occur (i.e., H). Figure 4 shows a plot of skill scores with H on the vertical axis and F on the horizontal axis. As before, H and F are calculated for each grid cell and then averaged over the west coast of North America. Figure 4 shows reforecast 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 S1 in the supplemental material.) It is clear that H decreases and F increases as lead time increases, as seen in the movement from top left to bottom right. Overall, there exists a large range of skill values between models. Here 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 4 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, 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 sizeable drop in skill just 1 day from initialization, which is due to a sharp decrease in H as described above.

To examine reforecast skill at subregional scales, Fig. 5 shows PSS by lead time for three subregions (Fig. 2): Santa Barbara, California, to the California–Oregon

Model Skill Scores for the West Coast of North America

FIG. 4. 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 reforecast lead.

border (CA), the California–Oregon border to Vancouver Island (OR/WA), and Vancouver Island to Juneau, Alaska (BC/AK). (The latitude–longitude coordinates for each of these subregions are listed in the supplemental material.) Specifically, for each model and lead time, PSS is calculated for each grid cell within the subregion, and then these PSS values are averaged in order to get a single mean PSS for the subregion. The multimodel average PSS for each subregion is denoted by colored lines. Figure 5 shows a steady decrease in PSS as lead time increases, with PSS approaching 0 toward 14 days, as already noted, and this trend

is consistent across subregions. At the same time, Fig. 5 shows variations in PSS between subregions, particularly at leads between 1 and 7 days. At these leads, there is a clear difference in model-averaged reforecast skill between BC/AK and the other two subregions, 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 subregion's unique climatology. Thus, the differences in skill between subregions cannot solely be attributed to varying

FIG. 5. Model-averaged PSS by reforecast lead averaged over each of the three subregions (see Fig. 2).

background climatologies. These subregional 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, Fig. 5 of Mundhenk et al. (2018) also showed that the occurrence-based AR reforecast skill (Heidke skill score) is about 10–20 points lower in British Columbia than in California at leads shorter than 7 days.

Figure 6 shows PSS at an even smaller scale by showing the skill metric at each individual grid cell in the landfall region. For each panel, the skill metrics are

calculated after combining and tallying all contingency table counts for the lead days that compose the given lead window (e.g., lead days 1–3). While variation exists across models, Fig. 6 indicates that the highest PSS across all nine models tends to be seen along the coasts of Northern California and Oregon for leads of 1–6 days. At these leads, the lowest PSS generally occurs in the grid cells at the northern extent of the landfall domain (i.e., BC/AK), as already seen in Fig. 5. However, at leads of 7–10 days, no particular subregion is favored for higher PSS. This also corresponds with Fig. 5, which shows that PSS is similar between subregions at leads greater than 7 days.

b. AR intensity

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

Figure 7 shows the root-mean-square error (RMSE) in landfall IVT (i.e., the IVT at the landfall grid cells) for each model and subregion. Here, lead time is defined in terms of overlapping 3-day lead windows such that lead times of 1–3 days, 2–4 days, 3–5 days, etc., are grouped. Lead windows are applied in this context as a means of smoothing the results. In Fig. 7, discontinuities appear in

FIG. 6. Geographical distribution of PSS values for the west coast of North America. Locations along the detection domain (Fig. 2) are numbered (1–18) from south to north along the horizontal axis. Locations 1–6 are CA, 7–12 are OR/WA, and 13–18 are BC/AK. Plotted PSS values are averaged for three different lead windows. Each color bar denotes a different range of PSS values.

Error in Landfall IVT for Re-Forecast Hits

FIG. 7. RMSE in landfall IVT magnitude for the model and ERA-Interim for reforecast “hits” in the subregion during 3-day lead windows. Discontinuities appear for lead windows without any reforecast hits in the catalog.

CA and BC/AK because of a lack of reforecast hits for the particular model during the lead window. Overall, little difference exists in the distribution of IVT RMSE between subregions, and for all three subregions, the average magnitude of landfall IVT RMSE stays fairly constant between 100 and $250 \text{ kg m}^{-1} \text{ s}^{-1}$ as lead time increases, though sample sizes decrease to around 10 for several models (e.g., CNRM, ECCC, and UKMO) at leads of 11–14 days.

Figure 8 shows the distribution of absolute landfall IVT error for all nine models for lead windows of 1–3, 4–6, 7–10, and 11–14 days. Because of the subregion's higher AR climatological frequency during NDJF, results for OR/WA are shown in Fig. 8, though the distributions for the other subregions are generally similar (see the supplemental material). 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 1–3 days and 4–6 days. This indicates that a majority of ECMWF reforecasts 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 reforecasts that make up these distributions are hits, so the a priori assumption is that the reforecasts 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. Reforecast and reanalysis fields of IVT for each model, lead time, and subregion are composited and compared. For such analysis, reforecast hits are used once again. Results are shown for ECMWF and reanalysis at lead times of 1–3 days for hits, and Fig. 10 shows the percent difference for NCEP. To account for correlation between consecutive days of a multiday AR event, sample sizes for statistical significance testing are adjusted to the number of AR “events.” Here, we consider AR occurrences separated by at least 2 days to be separate events. 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 reforecasts show extensive low-IVT biases upstream of the landfall domains (Fig. 9). Since an AR feature (i.e., a corridor of high IVT) must be present in the subregion, absolute low-IVT biases for ECMWF

Error in Landfall IVT (model - ERAI) for Re-Forecast Hits OR/WA

FIG. 8. Distribution of absolute error in landfall IVT for the model and ERA-Interim (model - ERAI) for reforecast hits in OR/WA during four different lead windows (1–3, 4–6, 7–10, and 11–14 days). Black dots denote the median of the distribution.

Percent Difference in IVT (ECMWF - ERAI)
Lead Days 1 through 3

FIG. 9. Percent difference between composite IVT for the ECMWF reforecasts and ERA-Interim $[(\text{ECMWF} - \text{ERA-Interim})/\text{ERA-Interim}]$ for reforecast hits in the subregion between lead times of 1 and 3 days. Black dots denote absolute differences that are statistically significant at the 95% confidence level. The number (N) of individual AR “events” is given in the plot titles.

Percent Difference in IVT (NCEP - ERAI)
Lead Days 1 through 3

FIG. 10. As in Fig. 9, but for NCEP.

(not shown) are even more prominent. Together, Figs. 8 and 9 point to a low-IVT bias in ECMWF, at short lead times, which exists offshore. By contrast, NCEP (Fig. 10) 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 Fig. 3. Recall that Fig. 3 showed that ECMWF, at all lead times, reforecasts fewer ARs along the west coast of North America compared to reanalysis. Since the AR detection algorithm uses a static IVT cutoff, the low-IVT bias in ECMWF's offshore IVT likely resulted in fewer detected ARs.

c. 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 reforecast an AR for a particular grid cell when one actually occurs (a miss). The model may still reforecast a landfalling AR for that valid time, but the feature may be located elsewhere along the coast. In another scenario, a model may correctly reforecast an AR landfall in a subregion 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 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 reforecast?

Figure 11 shows, for each subregion, the distribution of 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 RMSEs for lead times of 1–3 days are generally between 100 and 600 km. All three subregions show an increase in landfall RMSE 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 corresponds with the aforementioned decrease in occurrence-based skill at lead times approaching 14 days. Figure 11 does not show statistical significance of the RMSEs 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 reforecast 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 reforecast landfall latitude. For instance, a positive landfall latitude error implies that the model reforecasts an AR landfall location too far to the north (i.e., a “northward” error). The frequency of northward location error is calculated for each subregion and plotted in Fig. 12. The frequencies are calculated as the number of northward errors divided by the number of nonzero (i.e., northward or southward) errors. Perfect landfall location reforecasts are not included in these calculations. Thus, a frequency greater than 0.5 implies that given an incorrect landfall location reforecast, the model is more likely to reforecast 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 reforecast the feature farther to the south.

As seen in Fig. 12, nonzero 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 nonzero location errors at all lead times. Meanwhile, in OR/WA, a majority of the models favor southward nonzero landfall location errors at lead windows of 1–3 days and 2–4 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 nonzero 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 versus too far to the south. Based on this assumption, an incorrect landfall location reforecast 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, Fig. 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 reforecast 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 subregion, statistical significance of the frequency of northward landfall location error

RMSE of Landfall Location Error

FIG. 11. RMSE in landfall location error (in km) for ARs observed in the subregion during 3-day lead windows. An AR must be reforecast somewhere along the west coast of North America for the same day as the AR observation.

(cf. the frequency from climatology) for each grid cell is tested. If greater than 50% of the grid cells within the subregion have northward frequencies that are statistically significant at 95% confidence, then the frequency for the entire subregion is considered statistically significant compared to climatology (denoted by dots in Fig. 12). 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 reforecasts 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 because of lower sample sizes. In OR/WA, it is expected that about 50%–85% of the incorrect landfall location reforecasts 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, HMCN, and NCEP. In BC/AK, it is expected that only about 5%–35% of the incorrect landfall location reforecasts 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 subregion.

4. Discussion and conclusions

Our study examines reforecasts of landfalling atmospheric rivers (ARs) along the west coast of North America from the control runs of nine numerical weather prediction (NWP) models, each covering about 10–30 years of reforecasts. In total, our study examines over 8000 reforecast initializations. For the purposes of model verification, reforecasts are compared to atmospheric reanalysis data (ERA-Interim), and models are assessed with respect to three components of an effective AR reforecast: AR occurrence, AR intensity, and AR landfall location.

Occurrence-based reforecast skill is examined for all nine models in order to determine how often each model correctly reforecasts 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., Wick et al. 2013; Nayak et al. 2014; DeFlorio et al. 2018a). By a lead of 14 days, models generally provide little additional skill (cf. a random forecast), as similarly demonstrated by Nayak et al. (2014) for the central United States and DeFlorio et al. (2018a) for the North Pacific and western United States. A novel finding of our study is that reforecast skill varies by subregion, 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 reforecasts are also assessed. For all models, the AR landfall IVT RMSE stays fairly constant with lead time. These IVT RMSEs typically range from approximately 100–250 kg m^{−1} s^{−1}. At leads of 1–3 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.

Finally, errors in AR landfall location are examined. For all nine models, there is a pronounced increase in landfall location RMSE as lead time increases. At leads of 1–3 days, landfall location RMSEs generally range from about 100 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 reforecast is too far south versus 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.

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 reforecast output from the available ensemble simulations. Furthermore, the models used in this study vary by time period, initialization frequency, and ensemble size. Future studies should take advantage of new datasets that have more uniformity (e.g., in terms of time period and initialization frequency) between models. In addition, our study does not use multiday windows for AR

Frequency of Northward Non-Zero Model Error in AR Landfall Location

FIG. 12. Fraction of positive latitudinal landfall location errors for ARs observed in the subregion during 3-day lead windows. An AR must be reforecast 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 nonzero latitude errors divided by the total number of nonzero latitude

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 $500 \text{ kg m}^{-1} \text{ s}^{-1}$. Different AR detection algorithms may decrease or increase the total number of landfalling features examined. The spatial resolution of $1.5^\circ \times 1.5^\circ$ also adds uncertainty to the exact locations of AR landfalls. As a final note, because of the differences in model datasets and our use of control runs only, this study does not aim to draw conclusions about which model(s) should be preferred when making AR-related forecasts. Rather, the results presented here provide an overview of the uncertainty and skill associated with nine different models over various leads and geographical regions. Additional study is necessary in order to understand why certain errors (such as a large drop in PSS from initialization to day 1) occur in some models as opposed to others.

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. Results also show spatial variations in reforecast skill, with models consistently more skillful in some locations (i.e., CA and OR/WA) compared to others (BC/AK). Geographical differences in reforecast skill may be related to the synoptic-scale setup across the North Pacific Ocean, which is important for the characteristics of landfalling ARs along the west coast of North America (e.g., [Hecht and Cordeira 2017](#)). In terms of AR activity, [Mundhenk et al. \(2016b\)](#) found a clear modulation between Alaska and California based on the geopotential height field across the North Pacific. Specifically, they found that the presence of a blocking high across the northeast Pacific is favorable for increased AR activity in Alaska. Prior work has shown that dynamical models, though improving, struggle in predicting the location and duration of atmospheric blocking events (e.g., [D'Andrea et al. 1998](#), [Palmer et al. 2008](#); [Matsueda et al. 2011](#); [Davini and D'Andrea](#)

[2016](#)). Therefore, errors in modeling blocking events across the northeast Pacific may explain errors in forecasting landfalling ARs for BC/AK.

While our study finds that models provide little additional forecast skill at leads greater than 14 days, recent studies of the impacts of low-frequency climate variability on ARs have demonstrated the potential to extend occurrence-based model skill to longer lead times. For example, [Zhou and Kim \(2018\)](#) examined AR reforecasts over the North Pacific [from the North American Multimodel Ensemble (NMME) dataset from 1981 to 2012] through the lens of El Niño–Southern Oscillation (ENSO). They found improvements in predictions of seasonal AR frequency over the northeast Pacific during ENSO winters. 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. \(2018a\)](#) compared global ECMWF ensemble reforecasts to ERAI at leads of 1–14 days from 1996 to 2013. Using an AR detection algorithm that allows for various location error thresholds, 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. [Baggett et al. \(2017\)](#) demonstrated the potential to use the Madden–Julian oscillation (MJO) and quasi-biennial oscillation (QBO) for subseasonal to seasonal 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 reforecasting anomalous weekly AR activity along the west coast of North America. The empirical model was compared to reforecasts 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. Still other efforts (e.g., [DeFlorio et al. 2018b](#); [Vitart and Robertson 2018](#)) are continuing to build on these works and further advance the prediction of extreme events at S2S leads. Therefore, recent efforts have shown the potential to improve upon model AR forecast skill beyond what we have demonstrated is currently available in deterministic forecasts by NWP models.

←

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 composing the subregion) based on ERA-Interim climatology.

Acknowledgments. This research was supported by a subaward under the Forecast Informed Reservoir Operations (FIRO) project at the Center for Western Weather and Water Extremes. The FIRO project is in partnership with local and federal agencies such as the Sonoma County Water Agency and the U.S. Army Corps of Engineers. We thank Cory Baggett, Kai-Chih Tseng, and Alexander Goodman for their assistance in obtaining the data and calculating IVT, and we thank Russ Schumacher for his insights. We also thank Bryan Mundhenk for his contributions in the development of the AR detection algorithm. 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.

APPENDIX

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 $500 \text{ kg m}^{-1} \text{ s}^{-1}$ 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 $300\,000 \text{ km}^2$, 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 $750 \text{ kg m}^{-1} \text{ s}^{-1}$ 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 eyehole, 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.

REFERENCES

- 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, <https://doi.org/10.1002/2017GL074434>.
- 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, <https://doi.org/10.1175/MWR3123.1>.
- 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, <https://doi.org/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, <https://doi.org/10.1007/s003820050230>.
- 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, <https://doi.org/10.1175/JCLI-D-16-0242.1>.
- DeFlorio, M., D. Waliser, B. Guan, D. Lavers, F. Ralph, and F. Vitart, 2018a: Global assessment of atmospheric river prediction skill. *J. Hydrometeorol.*, **19**, 409–426, <https://doi.org/10.1175/JHM-D-17-0135.1>.
- , —, —, F. Ralph, and F. Vitart, 2018b: Global evaluation of atmospheric river subseasonal prediction skill. *Climate Dyn.*, <https://doi.org/10.1007/s00382-018-4309-x>, 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, <https://doi.org/10.3390/w3020445>.
- Guan, B., and D. Waliser, 2015: Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *J. Geophys. Res. Atmos.*, **120**, 12 514–12 535, <https://doi.org/10.1002/2015JD024257>.
- , 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**, L20401, <https://doi.org/10.1029/2010GL044696>.
- Hamill, T., J. Whitaker, and S. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, <https://doi.org/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. Hydrometeorol.*, **18**, 1359–1374, <https://doi.org/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**, 9048–9058, <https://doi.org/10.1002/2017GL074179>.
- Jackson, D., M. Hughes, and G. Wick, 2016: Evaluation of land-falling atmospheric rivers along the U.S. West Coast in reanalysis data sets. *J. Geophys. Res. Atmos.*, **121**, 2705–2718, <https://doi.org/10.1002/2015JD024412>.

- Jie, W., F. Vitart, T. Wu, and X. Liu, 2017: Simulations of the Asian summer monsoon in the sub-seasonal to seasonal prediction (S2S) database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2282–2295, <https://doi.org/10.1002/qj.3085>.
- 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., and G. Villarini, 2013: The nexus between atmospheric rivers and extreme precipitation across Europe. *Geophys. Res. Lett.*, **40**, 3259–3264, <https://doi.org/10.1002/grl.50636>.
- , —, 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**, D20106, <https://doi.org/10.1029/2012JD018027>.
- , 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. *2014 Fall Meeting*, San Francisco, CA, Amer. Geophys. Union, Abstract A53K-3353.
- 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, <https://doi.org/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, <https://doi.org/10.1175/JCLI-D-15-0655.1>.
- , —, —, 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, <https://doi.org/10.1002/2016JD025350>.
- , —, —, and C. Baggett, 2018: Skillful empirical sub-seasonal prediction of landfalling atmospheric river activity using the Madden-Julian oscillation and quasi-biennial oscillation. *Nature Climate Atmos. Sci.*, **1**, 20177, <https://doi.org/10.1038/s41612-017-0008-2>.
- 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, <https://doi.org/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, [https://doi.org/10.1175/1520-0493\(2002\)130<1468:TSRBUF>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1468:TSRBUF>2.0.CO;2).
- , —, 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, <https://doi.org/10.1175/2007JHM855.1>.
- , L. Schick, F. Ralph, M. Hughes, and G. Wick, 2011: Flooding in western Washington: The connection to atmospheric rivers. *J. Hydrometeor.*, **12**, 1337–1358, <https://doi.org/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, <https://doi.org/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, <https://doi.org/10.1175/BAMS-D-11-00188.1>.
- , 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, <https://doi.org/10.1175/JHM-D-17-0031.1>.
- , P. Neiman, and G. Wick, 2004: Satellite and CALJET aircraft observations of atmospheric rivers over the eastern North Pacific Ocean during the winter of 1997/98. *Mon. Wea. Rev.*, **132**, 1721–1745, [https://doi.org/10.1175/1520-0493\(2004\)132<1721:SACAOO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2004)132<1721:SACAOO>2.0.CO;2).
- , —, —, S. Gutman, M. Dettinger, D. Cayan, and A. White, 2006: Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, **33**, L13801, <https://doi.org/10.1029/2006GL026689>.
- , 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, <https://doi.org/10.1175/2010JHM1232.1>.
- , T. Coleman, P. Neiman, R. Zamora, and M. Dettinger, 2013: Observed impacts of duration and seasonality of atmospheric-river landfalls on soil moisture and runoff in coastal northern California. *J. Hydrometeor.*, **14**, 443–459, <https://doi.org/10.1175/JHM-D-12-076.1>.
- , J. Rutz, J. Cordeira, M. Dettinger, M. Anderson, L. Schick, C. Smallcomb, and D. Reynolds, 2017a: A scale to characterize the strength and impacts of atmospheric rivers. *2017 Fall Meeting*, New Orleans, LA, Amer. Geophys. Union, Abstract H23F-1749.
- , and Coauthors, 2017b: Dropsonde observations of total integrated water vapor transport within North Pacific atmospheric rivers. *J. Hydrometeor.*, **18**, 2577–2596, <https://doi.org/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, <https://doi.org/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, <https://doi.org/10.1175/MWR-D-13-00168.1>.
- Schiraldi, N., and P. Roundy, 2017: Seasonal-to-subseasonal model forecast performance during agricultural drought transition periods in the U.S. Corn Belt. *Mon. Wea. Rev.*, **145**, 3687–3708, <https://doi.org/10.1175/MWR-D-17-0026.1>.
- Vitart, F., 2017: Madden-Julian Oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2210–2220, <https://doi.org/10.1002/qj.3079>.
- , and A. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *Nature Climate Atmos. Sci.*, **3**, <https://doi.org/10.1038/s41612-018-0013-0>.
- , and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–176, <https://doi.org/10.1175/BAMS-D-16-0017.1>.
- Waliser, D., and B. Guan, 2017: Extreme winds and precipitation during landfall of atmospheric rivers. *Nature*, **10**, 179–183, <https://doi.org/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, <https://doi.org/10.1175/NAF-D-13-00025.1>.
- Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Elsevier Academic Press, 648 pp.
- Zhou, Y., and H. Kim, 2018: Prediction of atmospheric rivers over the North Pacific and its connection to ENSO in the North American multi-model ensemble (NMME). *Climate Dyn.*, **51**, 1623–1637, <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, <https://doi.org/10.1029/94GL01710>.
- , and —, 1998: A proposed algorithm for moisture fluxes from atmospheric rivers. *Mon. Wea. Rev.*, **126**, 725–735, [https://doi.org/10.1175/1520-0493\(1998\)126<0725:APAFMF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1998)126<0725:APAFMF>2.0.CO;2).