A Summary of GFS Ensemble Integrated Water Vapor Transport Forecasts and Skill along the U.S. West Coast during Water Years 2017–20

JASON M. CORDEIRA and F. MARTIN RALPH

a Meteorology Program, Plymouth State University, Plymouth, New Hampshire
b Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

(Manuscript received 16 July 2020, in final form 12 October 2020)

ABSTRACT: The ability to provide accurate forecasts and improve situational awareness of atmospheric rivers (ARs) is key to impact-based decision support services and applications such as forecast-informed reservoir operations. The purpose of this study is to quantify the cool-season water year skill for 2017–20 of the NCEP Global Ensemble Forecast System forecasts of integrated water vapor transport along the U.S. West Coast commonly observed during landfalling ARs. This skill is summarized for ensemble probability-over-threshold forecasts of integrated water vapor transport magnitudes ≥ 250 kg m⁻¹ s⁻¹ (referred to as P₂₅₀). The P₂₅₀ forecasts near North-Coastal California at 38°N, 123°W were reliable and successful at lead times of 1–8 days with an average success ratio > 0.5 for P₂₅₀ forecasts ≥ 50% at lead times of 8 and Brier skill scores > 0.1 at a lead time of 8–9 days. Skill and accuracy also varied as a function of latitude and event characteristics. The highest (lowest) success ratios and probability of detection values for P₂₅₀ forecasts ≥ 50% occurred on average across Northern California and Oregon (Southern California), whereas the average probability of detection of more intense and longer duration landfalling ARs was 0.1–0.2 higher than weaker and shorter duration events at lead times of 3–9 days. The potential for these forecasts to enhance situational awareness may also be improved, depending on individual applications, by allowing for flexibility in the location and time of verification; the success ratios increased 10%–30% at lead times of 5–10 days allowing for flexibility of ±1.0° latitude and ±6 h in verification.

SIGNIFICANCE STATEMENT: The purpose of this study is to evaluate the efficacy of the Global Ensemble Forecast System model in forecasting enhanced water vapor transport along the U.S. West Coast commonly associated with landfalling atmospheric rivers and heavy precipitation. The model allows us to calculate the probability that enhanced water vapor transport will occur, thereby providing situational awareness for decision-making, such as in water resource management. The results of this study indicate that the model provides useful forecast information relative to climate at lead times of 8–9 days with noticeable variability north and south along the coast and from one year to the next. Future work is aimed at evaluating the skill of additional ensemble model systems.

KEYWORDS: Hydrometeorology; Ensembles; Forecast verification/skill; Numerical weather prediction/forecasting; Probability forecasts/models/distribution

1. Introduction

a. Motivation

Pacific winter storms and landfalling atmospheric rivers (ARs) produce ~30%–50% of annual precipitation (Dettlinger et al. 2011) and >70% of extreme precipitation (i.e., top 5% events annually) along the U.S. West Coast (e.g., Lamjiri et al. 2017, 2020). These landfalling ARs are also known to produce societal impacts in association with extreme winds (e.g., Waliser and Guan 2017), avalanches (e.g., Hatchett et al. 2017), floods (e.g., Ralph et al. 2006, 2020), flash floods and debris flows (e.g., Young et al. 2017; Oakley et al. 2017, 2018), and landslides (e.g., Cordeira et al. 2019). These impacts are quantified by a large majority of National Weather Service–issued watches, warnings, and advisories attributed to landfalling ARs in California (Cordeira et al. 2018) and 84% of all major flood damages over 40 years of FEMA flood claims in the western United States (Corringham et al. 2019). These landfalling ARs also play a primary role in water resources management and water supply across the western United States (e.g., Dettinger et al. 2011; Jasperse et al. 2017; White et al. 2019; Henn et al. 2020). The ability to forecast the occurrence of landfalling ARs and their associated precipitation is therefore critical for hazard mitigation and water resource management (e.g., Guan et al. 2010; Dettinger et al. 2011; Baggett et al. 2017; DeFlorio et al. 2018; Mundhenk et al. 2018; DeFlorio et al. 2019a,b). The primary goal of this study is to quantify the skill of the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) forecasts of enhanced integrated water vapor transport (IVT) along the U.S. West Coast that is commonly observed during landfalling ARs during October–April 2017–20.

b. Background on AR-related forecast skill

A majority of cool-season precipitation forecasts over the western United States likely inherently account for landfalling ARs through the mesoscale and synoptic characteristics of their parent midlatitude cyclones (Waliser and Cordeira 2020). These characteristics, such as their thermodynamics,
kinematics, and dynamics, including their water vapor distributions, are also generally well simulated by numerical weather prediction (NWP) models (e.g., McMurdie and Mass 2004; McMurdie and Casola 2009; Froude 2010). For example, assessment of NWP forecast skill using a method called “potential predictability” (i.e., the predictability within an ensemble system that is attributable to systematic, rather than stochastic, influences; e.g., Waliser et al. 2003; Luo and Wood 2006; Kumar et al. 2014) identify that forecasts of the occurrence of regional maxima in IVT magnitude often associated with ARs are on average skillful for lead times of 7–9 days over the North Atlantic (Lavers et al. 2014), North Pacific (Lavers et al. 2016), and central United States (Nayak et al. 2014). Global hindcasts of the occurrence of individual AR “objects” also defined by regional maxima in IVT magnitude can be well forecast relative to climatology on 1–2-week time scales by certain NWP models (Nardi et al. 2018) and relative to random chance on 7–10-day time scales considering a subset of hindcasts that are initialized during different states of the climate system [e.g., phases of El Niño–Southern Oscillation (ENSO) or the Pacific–North American pattern; DeFlorio et al. 2018]. Similar studies of AR prediction skill over the eastern North Pacific for three cool seasons illustrate that the regional occurrence of ARs is well forecast at lead times of 9–10 days; however, the landfall occurrence and position are forecast more poorly with decreasing skill as a function of increasing lead time (Wick et al. 2013). For example, the average errors in the landfall position of ARs along the U.S. West Coast exceed 600–900 km at 7–10-day lead times for all five NWP model suites tested during a 3-yr period, including the European Centre for Medium-Range Weather Forecasts (ECMWF) and the NCEP models, among others.

Ralph et al. (2010) identify that evaluations of quantitative precipitation forecast (QPF) errors need to be considered in the context of landfalling ARs, which are often associated with enhanced lower tropospheric water vapor fluxes that generate orographic precipitation. These errors are likely influenced by both water vapor flux magnitude and water vapor flux direction, where small differences in direction can determine whether or not an individual watershed may be susceptible to extreme precipitation and flooding (Neiman et al. 2011). The human-forecaster QPF forecasts by the NCEP Weather Prediction Center (WPC) are often superior to the NWP-derived QPF forecasts [e.g., a higher threat score for QPF > 25.4 mm (24 h)⁻¹; Sukovich et al. (2014)] largely related to enhanced situational awareness afforded by consistent and quality-controlled verification in order to assess forecast trends and bias. Enhanced situational awareness related to knowledge of local topography, climate, and seasonal precipitation regimes along the U.S. West Coast was also deemed key to successful AR-related QPF during the “Atmospheric River Retrospective Forecasting Experiment” (WPC 2012) that sought to identify potential techniques to improve forecasts of AR-induced extreme precipitation events along the U.S. West Coast. A second key outcome related to situational awareness from the aforementioned experiment stated, “Model forecasts of moisture parameters may be helpful in identifying the potential for extreme [precipitation] events, even when the model QPF does not forecast large precipitation amounts” (WPC 2012).

c. The “AR Landfall Tool”

Following the Atmospheric River Retrospective Forecast Experiment in 2012 and in preparation for the CalWater field experiment in 2015 (Ralph et al. 2016), an “AR portal” was developed for various forecast applications in order to situationally assess the intensity, timing, and duration of landfalling ARs (Cordeira et al. 2017). Among the number of deterministic and ensemble-based forecast imagery that comprised the AR portal was a series of well-received diagrams in a forecast lead time–latitude framework spanning the west coast of North America that illustrated IVT data from the NCEP–GEFS ensemble control, ensemble mean, and an ensemble probability-over-threshold for different IVT magnitude values of ≥250, ≥500, and ≥750 kg m⁻¹ s⁻¹ collocatedly known as the AR Landfall Tool (ARLT; Cordeira et al. 2017). The ARLT illustrates forecast IVT data from the NCEP–GESFS version 11.0.0 with 0.5° latitude × 0.5° longitude grid spacing in a pseudo-Hovmöller coastline-spanning framework for lead times out to 16 days. This illustration of IVT data along the coast provides situational awareness of the uncertainty in the likelihood, intensity, duration, and timing of possible landfalling ARs with different IVT magnitude thresholds. The ARLT imagery are generated in an experimental quasi-operational capacity four times daily at the Center for Western Weather and Water Extremes (CW3E) at the Scripps Institution of Oceanography at University of California San Diego following each 6-h initialization of the NCEP–GEFS.¹ These forecasts are also part of cool-season outlooks generated by CW3E that are actively used in situational awareness applications by various partners, public, private, and governmental agencies such as the California Department of Water Resources (J. Jones 2018, personal communication).

Examples of three ARLT images are shown for forecasts initialized at 0000 UTC 31 January 2017 and 0000 UTC 7 February 2019 of the ensemble-mean IVT magnitude (Figs. 1a and 2a), and the probability-over-threshold for IVT magnitudes ≥ 250 kg m⁻¹ s⁻¹ (hereafter written as F₂₅₀; Figs. 1b and 2b). These two forecast initializations are chosen as they occurred prior to landfalling ARs that produced rawinsonde-observed IVT magnitudes > 750–1000 kg m⁻¹ s⁻¹ at Bodega Bay (BBY), California, at 1200 UTC 7 February 2017 and 1200 UTC 13 February 2019, respectively. These landfalling ARs also produced challenges to water resources infrastructure and management at Lake Oroville in California in 2017 (e.g., France et al. 2018; Vano et al. 2019; White et al. 2019) and statewide flooding from north-coastal regions to the southern-inland deserts in 2019, respectively (e.g., Ralph et al. 2020). The 2017 forecast highlights approximately six different periods of enhanced F₂₅₀ > 50% associated with a series of landfalling ARs along the western North American coastline (labeled in Fig. 1b) occurring between 2 and 11 February 2017.

¹The ARLT imagery may be accessed via CW3E’s website at http://cw3e.ucsd.edu.
The $P_{250}$ forecasts were complemented by ensemble forecasts of ensemble-mean IVT magnitudes $> 300–400 \text{ kg m}^{-1} \text{s}^{-1}$ (Fig. 1b). The 2019 forecast highlights two periods of enhanced $P_{250} > 50\%$ along the western North America coastline (labeled in Fig. 2b) occurring between 12 and 18 February. These $P_{250}$ forecasts were associated with forecasts ensemble-mean IVT magnitudes $> 500 \text{ kg m}^{-1} \text{s}^{-1}$ (Figs. 2a,b). In both examples, the forecasts of ensemble mean IVT magnitude and $P_{250}$ in the coastline-spanning framework provides situational awareness and illustration if the timing, duration, and propagation characteristics of the landfalling ARs.

The ARLT imagery is a framework for illustrating deterministic and ensemble IVT-related parameters in an effort to improve situational awareness along the U.S. West Coast. Herein we focus on ensemble forecasts of $P_{250}$ as a function of forecast lead time and coastal latitude given the relative lack of verification statistics for probability-over-threshold forecasts as compared to ensemble forecasts of IVT magnitude (e.g.,
DeFlorio et al. 2018). Section 2 provides additional information on the data and methods, and section 3 summarizes the skill of the $P_{250}$ forecasts for North-Coastal California for the NCEP–GEFS grid point closest to BBY (38°N, 123°W). This grid point is chosen for closer analysis given its central coastal location in the motivating example from early February 2017 (Fig. 1). Section 4 summarizes the skill of the $P_{250}$ forecasts for the west coast of North America for the NCEP–GEFS grid points spanning the coast as illustrated in Figs. 1 and 2, whereas sections 5 and 6 provide additional considerations and the conclusions, respectively.

2. Data and methods

The ARLT skill is assessed using the operational $0.5°$ latitude $\times$ $0.5°$ longitude NCEP–GEFS version 11.0.0 gridded forecast data (hereafter referred to as the GEFS) verifying during the first seven months (October–April) of water years (WY) ending in 2017, 2018, 2019, and 2020. Assessing skill relative to verification for October through April necessitated obtaining forecast initializations beginning on 14 September prior to the start of each WY. These data were obtained from The Interactive Grand Global Ensemble (TIGGE; Bougeault et al. 2010) data portal at ECMWF for which $0.5°$ data were available for 1762 of the 1832 (96.2%) forecast initializations needed to verify forecasts during WY17 and WY18. The missing data for 70 initializations verifying in WY17 and WY18 were calculated from $1.0°$ latitude $\times$ $1.0°$ longitude GEFS data obtained from the National Centers for Environmental Information (NCEI) Climate Data Archive and were simply linearly interpolated to $0.5°$ grid spacing. The data were obtained in quasi-real time at 0000 UTC 10 Feb 2019.
0.5° grid spacing for all 1836 initializations verifying in WY19 and WY20.

The GEFS data are used to calculate ensemble-member forecasts of IVT magnitude following the methodology of Neiman et al. (2008) and Moore et al. (2012) using the 1000-, 925-, 850-, 700-, 500-, and 300-hPa isobaric levels. The ensemble member forecasts of IVT magnitude are calculated at 62 latitude and longitude locations along or just offshore of the west coast of North America every 0.5° latitude between 25° and 55°N (see right panels of Figs. 1 or 2). The resulting forecasts are organized in a forecast lead time–coastal latitude Hovmöller framework for lead times from 0 to 16 days every 6 h. As described in section 1, the probability of AR landfall in this study is defined as the probability or fraction of the 20 ensemble members that contain an IVT magnitude $\geq 250 \text{ kg m}^{-1} \text{s}^{-1}$ (i.e., $P_{250}$). This IVT magnitude threshold is a common threshold used for identifying ARs from gridded reanalysis and forecast data over the northeast Pacific (Rutz et al. 2014; Cordeira et al. 2017; Ralph et al. 2019). Different IVT magnitude thresholds and thresholds of other meteorological parameters that are used in the global detection of ARs from gridded reanalysis and forecast data are reviewed in Shields et al. (2018). The $P_{250}$ is computed for all forecasts initialized between 0000 UTC 14 September and 1800 UTC 30 April in each WY at all 62 latitude locations for the 65 6-h lead times. These forecasts are compared against verification computed as the GEFS ensemble-mean 0-h forecast IVT magnitude $\geq 250 \text{ kg m}^{-1} \text{s}^{-1}$ every 6 h from 0000 UTC 1 October to 1800 UTC 30 April in each WY.

Forecast skill is primarily assessed using occurrence-based metrics that are formulated from a four-outcome ($2 \times 2$) contingency table (Wilks 2006). The four-outcome contingency table describes whether or not populations of threshold forecasts of a non-AR landfall (i.e., $P_{250}$ under a given percentage threshold) verify at different lead times. The metrics used in this study derived from the contingency table include the success ratio, false alarm ratio (i.e., one minus the success ratio), probability of detection (POD; i.e., hit rate), and the probability of false detection (POFD; i.e., false alarm rate). The success ratio and false alarm ratio identify what fraction of forecasted events verify or fail to verify, whereas the POD and POFD identify what fraction of observed events were or were not forecasted, respectively.

Additional metrics derived from the contingency table include the equitable threat score (ETS) and relative operating characteristic (ROC) curves (Hanley and McNeil 1982). The ETS measures the fraction of observed events that were correctly forecast and considers skill relative to random chance, whereas the ROC curves express how well observed events were forecast using the phase space created by combinations of the POD and POFD. The ROC score is also calculated from the area under the ROC curve, ranges from 1 to 0, and is proportional to the critical success index. Finally, the Brier skill score (BSS) is calculated in order to assess the relative skill of the $P_{250}$ forecasts as compared to a reference probabilistic forecast. In this study, the reference probabilistic forecast and skill of random chance is calculated as the number of verifying times with IVT magnitude $\geq 250 \text{ kg m}^{-1} \text{s}^{-1}$ for all four years or in each WY. The skill metrics are calculated for all data in the four-WY period, herein referred to as the aggregate skill, and also for each individual WY, given at least 10 valid forecasts.

3. Results: North-Coastal California

a. Verification and forecast overviews

The cool-season portions of the four WYs contained 514 6-h times (of 3396; 15.1%) with GEFS 0-h ensemble-mean IVT

![Fig. 3. Time series of 6-hourly GEFS 0-h ensemble-mean IVT magnitudes (kg m⁻¹ s⁻¹; black lines) at 38°N, 123°W and 24-h precipitation (mm; red lines) ending at 1200 UTC from the Northern Sierra 8-station index each day between 1 Oct and 30 Apr for (a) WY17, (b) WY18, (c) WY19, and (d) WY20.](image-url)
The statistical consistency of the GEFS IVT magnitude forecasts is assessed via a dispersion diagram (following Talagrand et al. 1997) and a binned spread–skill plot (e.g., van den Dool 1989; Wang and Bishop 2003). The dispersion diagram (Fig. 4a) shows a comparison of the average root-mean-square error (RMSE) of the GEFS ensemble mean IVT magnitude forecasts with the GEFS ensemble spread of the IVT magnitude as a function of lead time. The binned spread–skill plot (Fig. 4b) further analyzes this important attribute of probabilistic predictions by assessing the quality of the match between the ensemble spread and the ensemble mean for each equally populated bin of the spread for a given lead time.

Figure 4 demonstrates that the RMSE is on average larger than the associated ensemble spread representing underdispersion of the GEFS IVT forecasts. The ensemble spread is therefore on average not large enough to represent the standard error of the ensemble mean and implies, among possible other deficiencies, potential limitations of the GEFS uncalibrated ensemble-derived probability-over-threshold forecasts. As a consequence, ensemble member forecasts of IVT magnitude could cluster too closely above (or below) 250 kg m\(^{-1}\) s\(^{-1}\). In other words, they are excessively sharp at the expense of statistical consistency, resulting in overconfident \(P_{250}\) forecasts, which in turn could mislead a decision-maker basing their decision on these forecasts. This result also suggests that calibrating the GEFS ensemble forecasts of IVT magnitude, which is beyond the scope of this study, will likely improve the value of probability-over-threshold forecasts of landfalling ARs for the decision-making process (e.g., Johnson and Bowler 2009). Additional limitations of the study methodology are discussed in section 5.

An analysis of lead time versus verification time of \(P_{250}\) at 38\(^\circ\)N, 123\(^\circ\)W for each WY (Fig. 5; i.e., “\(d_{\text{Prog/dt}}\)”) illustrates intraseasonal variability in lead-time prediction of consecutive 6-h times with IVT magnitudes \(\geq 250\) kg m\(^{-1}\) s\(^{-1}\) likely associated with landfalling ARs over North-Coastal California. Visual inspection of individual verification times or consecutive verification times with IVT magnitudes \(\geq 250\) kg m\(^{-1}\) s\(^{-1}\) (cf. Figs. 3 and 5) suggests that some of these likely landfalling ARs are forecast with longer-lead predictability (e.g., \(P_{250}\) increasing to values above 50\% at lead times of 8–11 days during February 2017 labeled with a white arrow in Fig. 5a) while some are forecast with shorter-lead predictability (e.g., \(P_{250}\) increasing to values above 50\% at lead times of just 4–6 days during February 2019 labeled with a white arrow in Fig. 5c).

Further illustration of this lead-time variability is illustrated for all 65 forecasts verifying at the times of maximum IVT magnitude at 38\(^\circ\)N, 123\(^\circ\)W associated with the landfalling ARs in section 5.
magnitudes \( \gtrsim 250 \text{ kg m}^{-1} \text{s}^{-1} \) at lead times up to 14 days prior to verification, but is discontinuous for lead times of 9–10.5 days, an example of both forecast jumpiness (Zsoter et al. 2009) and caution in using trends in \( d \text{Prog}/dt \) as an operational forecast tool (e.g., Hamill 2003). The forecasts also feature latitudinal bias with aforementioned enhanced \( P_{250} \) and ensemble mean IVT magnitudes \( \sim 2^\circ \) too far north at lead times of 5–8 days and \( \sim 2^\circ \) too far south at lead times of 2–3 days (Fig. 6a). The 13 February 2019 example features \( P_{250} \) values increasing to above 50% at lead times of 5 days and \( P_{250} \) values increasing to above 75% at a lead time of 4 days (Fig. 6b).

The average \( P_{250} \) values as a function of lead time for all forecasts prior to periods with IVT magnitudes \( \gtrsim 250 \text{ kg m}^{-1} \text{s}^{-1} \) (i.e., events) are \( \sim 15\%–20\% \) at lead times > 10 days and are \( \sim 75\% \) at lead times of 2–3 days (Fig. 7). Alternatively, the average \( P_{250} \) values as a function of lead time for all forecasts prior to periods with IVT magnitudes < 250 kg m\(^{-1}\)s\(^{-1}\) (i.e., nonevents) are \( \sim 10\%–16\% \) at all lead times > 6 days and decrease below 5% as lead time decreases on average to

---

**FIG. 5.** A forecast-lead illustration of the GEFS probability of IVT magnitude \( \gtrsim 250 \text{ kg m}^{-1} \text{s}^{-1} \) (shaded according to scale) for verification times from 0000 UTC 1 Oct to 1800 UTC 30 Apr for (a) WY17, (b) WY18, (c) WY19, and (d) WY20 at 38°N, 123°W. These types of images are sometimes referred to as a “\( d \text{Prog}/dt \)” analyses and illustrate how forecast values change over time leading up to verification. The white arrows in (a) and (b) denote 1200 UTC 7 Feb 2017 and 1200 UTC 13 Feb 2019, respectively, as discussed in the text.

**FIG. 6.** A forecast-lead illustration of the GEFS ensemble-mean IVT magnitude (contoured every 100 kg m\(^{-1}\)s\(^{-1}\) starting at 100 kg m\(^{-1}\)s\(^{-1}\)) and probability of IVT magnitude \( \gtrsim 250 \text{ kg m}^{-1} \text{s}^{-1} \) (shaded according to scale) for all latitudes prior to verifying at (a) 1200 UTC 7 Feb 2017 and (b) 1200 UTC 13 Feb 2019. Note that these times represent the time of maximum IVT magnitude at 38°N, 123°W for two high-impact landfalling ARs discussed in the text.
The difference between average \( P_{250} \) forecast values prior to events and nonevents begins to grow rapidly at lead times of 8–10 days. On an annual basis, WY17 contained average \( P_{250} \) values prior to events that were 5–7 percentage points higher than the 4-yr aggregate and 10–15 percentage points higher than WY18, WY19, and WY20 at lead times > 2 days (i.e., note the vertical difference between solid lines in Fig. 7). With respect to a given average \( P_{250} \) value, WY17 featured higher average \( P_{250} \) values at longer lead times with values that increased above different percentage values 0.5–1.0 days ahead of WY18, 1.0–1.5 days ahead of WY19 and WY20 (i.e., note the horizontal difference between solid lines in Fig. 7). The two aforementioned case studies with events centered on 1200 UTC 7 February 2017 and 1200 UTC 13 February 2019 are shown as reference on Fig. 7 with the former illustrating \( P_{250} \) values increasing above 50% at a lead time of 10 days and the latter illustrating \( P_{250} \) values increasing above 50% at a lead time of ~4.5 days (cf. Figs. 6 and 7).

b. Forecast skill

The \( P_{250} \) forecasts at 38°N, 123°W during all four WYs are generally reliable with \( P_{250} \) forecasts of 50% at lead times of 7, 8, and 9 days verifying an average of 46% of the time (Fig. 8a). Higher probability \( P_{250} \) forecasts at longer lead times occur less frequently (recall that analyses are not performed for frequencies less than 10) and are also generally less reliable. For example, \( P_{250} \) forecasts of 50% at lead times of 10, 11, and 12 days verify an average of 33% of the time. Similarly, lower probability \( P_{250} \) forecasts at longer lead times are also less reliable and lack resolution with \( P_{250} \) forecasts of 0% at lead times of 13, 14, and 15 days verifying an average of 9% of the time. The success ratio for all aggregated \( P_{250} \) forecasts ≥ 50% decreases from ~0.8 at lead times of 1–3 days to ~0.5 at lead times of ~8–9 days (Fig. 8b). Note that several \( P_{250} \) forecast thresholds were investigated (e.g., ≥25% and ≥75%) and we chose to illustrate ≥50% for simplicity (see also section 5). The success ratio of \( P_{250} \) forecasts ≥ 50% during WY17 is >0.50 through a lead time of ~10 days and is approximately double the success ratio value of \( P_{250} \) forecasts ≥ 50% during WY19 thereafter. Overall, the GEFS \( P_{250} \) forecasts are reliable on average for most \( P_{250} \) forecasts between 30% and 70% at lead times through ~9 days (Fig. 8b). These forecasts are also successful on average more than half of the time at lead times through ~9 days for \( P_{250} \) forecasts using a threshold ≥ 50% (Fig. 8b). There is considerable WY-to-WY variability in the success of \( P_{250} \) forecasts at lead times ≥ 7 days (Fig. 8b).

Comparisons of the POD and POFD for different threshold \( P_{250} \) forecasts at 38°N, 123°W via a ROC curve analysis illustrate that events with IVT magnitudes ≥ 250 kg m\(^{-1}\) s\(^{-1}\) are predicted on average with high POD and low POFD through lead times of 7, 8, and 9 days (Fig. 8c). The 1–6-day forecasts with the highest POD and lowest POFD (i.e., nearest the upper left corner of the diagram) contained threshold \( P_{250} \) forecasts ≥ 20%–30%. These thresholds are consistent with the average \( P_{250} \) forecast values prior to events in Fig. 7 that begin to increase above 20%–30% at a lead time of ~8 days. These thresholds may seem low, but note they include all \( P_{250} \) forecasts over that threshold and that as lead time decreases...
relative to an event, the $P_{250}$ forecast should increase quickly to 100% for a perfect or sharp ensemble forecast. For $P_{250}$ forecasts of events with lead times $\leq 10$ days, the ROC curves project more linearly along the diagonal line that describes POD = POFD. The points along these ROC curves are still left of this diagonal, which is summarized by the ROC score that decreases from 1.0 to $\approx 0.5$ as lead time increases to 16 days (Fig. 8d). The ROC scores for each WY are relatively similar to one another for lead times through $\approx 10$ days. Whereas the success ratio of $P_{250}$ forecasts $\approx 50\%$ during WY17 at leads times $> 7$ days is higher than the success ratio in WY18 and WY19, the ROC score is highest for WY18 at lead times $> 6$ days (Figs. 8b,d). This result suggests that forecasts during WY17 contained higher $P_{250}$ values and were on average more successful at longer lead times, but they did not as accurately forecast events as well as forecasts during WY18. The WY19 had the least successful and least accurate $P_{250}$ forecasts of the four WYs at lead times $> 7$ days and WY20 fell somewhere in between.

Both the ETS and BSS skill metrics allow for a measure of $P_{250}$ forecast skill adjusted for successful forecasts due to random chance or as compared to a benchmark value (Fig. 9). The ETS for $P_{250}$ forecasts $\geq 50\%$ at 38°N, 123°W during all four WYs decreases on average from $\approx 0.7$ (subjectively “good”) at lead times of 0–1 day to $\approx 0.2$ (i.e., subjectively “not that good”) at lead times of $> 7$ days and asymptotes toward zero (i.e., “no skill over random chance”) at lead times of 10 days (Fig. 9a). The ETS for $P_{250}$ forecasts $\geq 50\%$ for lead times of 2–9 days in WY17 and WY18 are similar to the 4-yr aggregate value and $\approx 0.10$ higher than the ETS for $P_{250}$ forecasts $\geq 50\%$ in WY19. Given the ETS value similarities among WYs, the apparent success of $P_{250}$ forecasts suggested in Fig. 8b at lead times $> 7$ days during WY17 was likely driven by random chance and the large frequency of events. The BSS for all $P_{250}$ forecasts is also scored relative to random chance and considers forecasts of both events and nonevents (e.g., high $P_{250}$ forecasts prior to events with IVT magnitudes $\geq 250$ kg m$^{-1}$ s$^{-1}$ or low $P_{250}$ forecast prior to nonevents with

---

**Fig. 8.** Skill analyses for $P_{250}$ forecasts at 38°N, 123°W: (a) reliability for $P_{250}$ forecasts for grouped forecast lead days colored according to the legend with reference lines for 1:1 reliability, no skill, and no resolution (dashed black line as annotated), (b) success ratio for threshold $P_{250}$ forecasts using a threshold of $\geq 50\%$ for all forecasts (black) and each WY (colored according to legend), (c) ROC curves illustrating the probability of false detection vs probability of detection for threshold $P_{250}$ forecasts for grouped forecast lead days colored according to the legend with reference for no skill (black solid line), and (d) ROC score for threshold $P_{250}$ forecasts for all lead times for all forecasts (black) and each WY (colored according to the legend). Recall that analyses in (a) and (b) are only performed for populations of forecasts $\geq 10$. Individual lead times in (a) and (c) are grouped and averaged according to the legend.
IVT magnitudes < 250 kg m\(^{-1}\) s\(^{-1}\), respectively). The BSS during all four WYs decreases on average to a score of zero (i.e., “no skill over random chance”) at leads times of ~13 days (Fig. 9b). The BSS for WY18 and WY20 are generally higher than the 4-yr aggregate value and decrease to a score near zero at a lead time of 14–16 days. The BSS for WY19 is on average lower than the 4-yr aggregate value and decreases to zero at a lead time of 10.5 days. In summary, the ETS suggests that threshold \(P_{250}\) forecasts for IVT magnitudes $\geq 250$ kg m\(^{-1}\) s\(^{-1}\) can be skillful relative to random chance at lead times of ~7–9 days and the BSS suggests that \(P_{250}\) forecasts for events and nonevents can be skillful relative to random chance at lead times of 9–12 days. Note that nonevents with IVT magnitudes < 250 kg m\(^{-1}\) s\(^{-1}\) are a more common (~4:1) occurrence during WY17–20 than events with IVT magnitudes $\geq 250$ kg m\(^{-1}\) s\(^{-1}\), which suggests the potential usefulness of low-probability \(P_{250}\) forecasts in also forecasting the nonoccurrence of enhanced IVT magnitudes along the U.S. West Coast at lead times of 9–12 days.

4. Results: West Coast summary

The cool-season portions of the four WYs contained a total of ~250 6-h event times in WY17 and ~150 6-h event times in WY18–20 between 42.5\(^\circ\) and 46.5\(^\circ\) N along the Oregon and Washington coastlines (Fig. 10a). The number of event times decreases to ~100 per WY at latitudes > 48\(^\circ\) N in southern Canada and to ~50 per WY at latitudes < 30\(^\circ\) N in northern Mexico. Similar to results previously shown at 38\(^\circ\) N, the average \(P_{250}\) forecast values by latitude as a function of lead
time for all forecasts of events are generally <20% at lead times > 10 days (Fig. 10b). These values generally increase to >50% as lead time decreases to ~5.5 days along the Oregon coast, ~5 days along the Washington and California coasts, and <4 days along the northern Mexico coast (Fig. 10b).

The forecast skill for the West Coast is summarized by lead time and coastal latitude using a $P_{250}$ forecast threshold of ≥50% with the success ratio (Fig. 11a), the POD (Fig. 11b), and the ETS (Fig. 11c). The success ratio of $P_{250}$ forecasts ≥ 50% is >0.5 at lead times > 10 days along the Oregon and Washington coasts (~40°–48°N) where the frequencies of events with IVT magnitudes ≥ 250 kg m⁻¹ s⁻¹ are also highest across all WYs (Fig. 10a). Otherwise, the success ratio of $P_{250}$ forecasts ≥ 50% is >0.5 at lead times of ~7–9 days along the central California, Northern California, Oregon, and Washington coasts (~34°–48°N). The success ratio of $P_{250}$ forecasts ≥ 50% at lead times of 2–9 days is higher along the north coast of California and Oregon (~38°–44°N) as compared to values along the coasts of Southern California (~32°–34°N) and southwest Canada (>48°N). Similarly, the POD of events using $P_{250}$ forecasts ≥ 50% is relatively higher along the Oregon and Washington coasts (~40°–48°N) with POD > 0.5 at lead times > 5 days and POD > 0.3 at lead times > 9 days (Fig. 11b), and is relatively lower along the coasts of Southern California (~32°–34°N) and southwest Canada (>48°N) with POD > 0.5 at lead times of ~4–5 days and POD > 0.3 at lead times of ~6–7 days. In other words, less than half of events were correctly forecast by $P_{250}$ forecasts ≥ 50% at lead times > 5 days in Northern California and at lead times > 4 days in Southern California. When accounting for success due to random chance, the ETS of $P_{250}$ forecasts ≥ 50% is <0.5 at lead times of >3–4 days across all coastal locations (Fig. 11c). The ETS remains relatively higher along the north coast of California and Oregon (~38°–44°N) as compared to the coasts of Southern California (~32°–34°N) and southwest Canada (>48°N). The success ratio, POD, and ETS skill metrics suggest that the most successful and accurate $P_{250}$ forecasts of events for any given lead time during WY17–20 occurred on average for coastal locations in Northern California and Oregon. The least successful and least accurate $P_{250}$ forecasts at lead times of 2–7 days occurred on average for coastal locations in Southern California.

For threshold $P_{250}$ forecasts of both events and nonevents (IVT magnitudes of ≥250 and <250 kg m⁻¹ s⁻¹, respectively), the BSS illustrates that the most successful $P_{250}$ forecasts during WY17–20 occurred on average for coastal locations in California (~34°–42°N) at lead times > 7 days and Northern California (~38°–42°N) at lead times of 2–7 days (Fig. 11d). The least successful $P_{250}$ forecasts at lead times of 2–7 days occurred on average for coastal locations in Southern California.
California (~32°–34°N). Each WY featured relatively higher or lower BSS depending on latitude at lead times of three, five, and seven days (Fig. 12). For example, WY19 (WY17) contained the highest (lowest) 3- and 5-day BSS for latitudes > 44°N, whereas WY20 contained the lowest BSS for latitudes between 33° and 37°N (Figs. 12a,b). The 7-day analysis illustrates negative BSS values during WY20 along the Southern California coast near 34°N and during WY17 along the southwest Canada coast near 51°N (Fig. 12c). These negative BSS values, and other latitudes with near-zero BSS values (e.g., <0.1) demonstrate very little skill of P250 forecasts over random chance at a lead time of 7 days.

The forecast graphics and skill metrics in Figs. 10–12 suggest that the most successful and accurate P250 forecasts of events for any given lead time during WY17–20 occurred on average for coastal locations in Northern California and Oregon between ~36° and 44°N (Figs. 11c,d and 12). The least successful and least accurate P250 forecasts of events for any given lead time during WY17–20 occurred on average for coastal locations in Southern California (Figs. 11c,d and 12). This coastal variability is further illustrated via a summary schematic of the approximate lead times at which the success ratio degrades to a value of 0.5, the POD degrades to a value of 0.5, and the BSS degrades to a value of 0.2 (Fig. 13).

5. Additional considerations

This study considered all P250 forecasts irrespective of the characteristics of the verifying events, such as whether a P250 forecast verified with single or consecutive 6-h periods with IVT magnitudes ≥ 250 kg m⁻¹ s⁻¹ (i.e., event duration) or whether a P250 forecast verified during an event with maximum IVT magnitudes of 250 or 750 kg m⁻¹ s⁻¹ (i.e., event intensity). Assessing the ensemble forecasts and the verification data using these characteristics may better describe the prediction skill associated with landfalling ARs, as opposed to shorter duration and less intense periods of coastal IVT magnitude maxima that may not necessarily be associated with landfalling ARs. For example, Rutz et al. (2014) demonstrate that landfalling ARs typically have an average duration of ~20 h in North-Coastal California and Ralph et al. (2019) identify that even weak events require a minimum duration of 12 h to be included in their AR intensity scale. Changing criteria of the event duration and intensity influences the results shown with respect to the average P250 forecast values as a function of lead time. For example, the average P250 forecast values as a function of lead time for all forecasts for an isolated 6-h event (i.e.,

![Fig. 12. Brier skill score for all P250 forecasts (black) and in each WY (colored according to legend) by latitude at a lead time of (a) 3, (b) 5, and (c) 7 days.](image)

### Average lead time for skill to degrade to value

<table>
<thead>
<tr>
<th>Metric</th>
<th>Success ≥50%</th>
<th>POD ≥50%</th>
<th>BSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P250 threshold degradation value</td>
<td>6.5 d</td>
<td>4.5 d</td>
<td>5.5 d</td>
</tr>
<tr>
<td>55°W</td>
<td>&gt;10 d</td>
<td>5.0 d</td>
<td>6.0 d</td>
</tr>
<tr>
<td>50°W</td>
<td>&gt;10 d</td>
<td>5.5 d</td>
<td>7.0 d</td>
</tr>
<tr>
<td>45°W</td>
<td>9.0 d</td>
<td>5.0 d</td>
<td>7.0 d</td>
</tr>
<tr>
<td>40°W</td>
<td>7.0 d</td>
<td>4.5 d</td>
<td>6.5 d</td>
</tr>
<tr>
<td>35°W</td>
<td>7.0 d</td>
<td>4.0 d</td>
<td>6.5 d</td>
</tr>
</tbody>
</table>

![Fig. 13. A summary of the approximate lead times for P250 forecast skill based on the success ratio, probability of detection (POD), and Brier skill score (BSS) to degrade to a given value for 5° latitude spans along the coast derived from Fig. 11.](image)
with a duration of 6 h) at 38°N, 123°W were 20–30 percentage points lower than forecasts for an event with two or three consecutive 6-h periods (i.e., with durations of 12–18 h) for lead times < 7 days (Fig. 14a). The highest average $P_{250}$ forecast values as a function of lead time prior to events with IVT magnitudes between 250 and 500 kg m$^{-1}$ s$^{-1}$ were ~10 percentage points lower than those prior to events with IVT magnitudes between 500 and 750 kg m$^{-1}$ s$^{-1}$ (Fig. 14b). The highest average $P_{250}$ forecast values as a function of lead time prior to events with IVT magnitudes $\geq 250$ kg m$^{-1}$ s$^{-1}$ were ~65% at lead times of five days and ~45% at lead times of seven days, which were 5 to 15 percentage points higher than forecasts prior to AR1 or AR2 events (Fig. 14c). Based on the way in which the AR scale is computed by comparing 1) the event-maximum IVT magnitude in 250 kg m$^{-1}$ s$^{-1}$ bins starting at 250 kg m$^{-1}$ s$^{-1}$ with 2) durations of IVT magnitudes $\geq 250$ kg m$^{-1}$ s$^{-1}$ that are either <24, 24–48, and $\geq$ 48 h, it is likely the higher $P_{250}$ forecast values prior to AR3, AR4, or AR5 events is primarily driven by the higher $P_{250}$ forecast values associated with the event-maximum IVT magnitude (cf.

![Fig. 14. Average of $P_{250}$ forecasts as a function of lead time at 38°N, 123°W prior to events with IVT magnitudes $\geq 250$ kg m$^{-1}$ s$^{-1}$ during an event of (a) variable duration, (b) variable maximum intensity, and (c) variable AR scale classification according to Ralph et al. (2019). (d) The average POD for threshold $P_{250}$ forecasts $\geq 50\%$ verifying during events of variable AR scale classification.](image-url)
Because we are focusing on the characteristics of events, it is appropriate to summarize how these characteristics influence the POD as a function of lead time. The average POD for $P_{250}$ forecasts $\geq 50\%$ of longer duration and higher intensity AR3, AR4, and AR5 events were $\sim 10–20$ percentage points higher than the POD for AR1 or AR2 events at lead times of three to eight days (Fig. 13d). The POD for AR3, AR4, and AR5 events was 0.73 at a lead time of five days and 0.52 at a lead time of seven days, suggesting that longer duration and higher intensity ARs were correctly forecast, on average, $>50\%$ of the time for $P_{250}$ forecasts $\geq 50\%$ at lead times up to seven days. These lead times are $\sim 1–2$ days longer than shorter duration and less intense ARs.

The forecast assessment herein also considered strict verification metrics as necessary to compute the contingency-based skill metrics. This study largely focused on $P_{250}$ threshold forecasts for probability values $\geq 50\%$ and only considered a forecast successful if it resulted in a “hit” at the exact predicted forecast hour at the exact predicted latitude. A $P_{250}$ forecast may still be considered useful for decision-making purposes and to-some-degree successful if it results in a hit within a given region (e.g., $\pm 1.0^\circ$ latitude) and within a given time window (e.g., $\pm 6$ h). These allowances are arbitrary, but in practice should decrease as a function of decreasing lead time and be tailored to specific applications. For example, allowances of $\pm 1.0^\circ$ latitude and $\pm 6$ h are likely not sufficient for some water resource applications at short lead times where differences of 100–200 km may result in one watershed receiving a lot of precipitation and another receiving very little (e.g., Neiman et al. 2011) and differences of 6–12 h may be crucial in reservoir drawdown in advance of extreme precipitation events (e.g., Jasperse et al. 2017). At longer lead times, however, these allowances applied to $P_{250}$ forecasts may still provide enhanced situational awareness for various applications. Similarly, a minimum $P_{250}$ threshold of 50% may still provide too much uncertainty for some decision-making applications as compared to minimum $P_{250}$ thresholds of 75% or 90%. These thresholds are likewise arbitrary, but in practice should increase as a function of decreasing lead time and be tailored to specific applications and risk tolerances. For example, using allowances of both $\pm 1.0^\circ$ latitude and $\pm 6$ h to score events at 38°N, 123°W increases the 5-day success ratio of $P_{250}$ forecasts $\geq 50\%$ from 0.66 to 0.86 (Fig. 15a). Similarly, higher probability $P_{250}$ forecasts using thresholds of $\geq 75\%$ and $\geq 90\%$
produced increases from 0.81 to 0.93 and from 0.84 to 0.94, respectively (Figs. 15b,c). These day-5 increases in the success ratio represent 23%, 15%, and 11% improvements in the success ratio as compared to the success ratio without any allowances (Fig. 15d). These allowances could also be applied to individual ensemble members prior to the calculation of $P_{250}$ (and other ensemble forecast metrics). This approach would account for ensemble member variability in the timing, duration, and locations of maximum IVT magnitudes. A full assessment of these forecast and verification allowances is beyond the scope of the current investigation and is the focus of future work.

6. Conclusions

This study focuses on summarizing the cool-season WY skill for 2017–20 of the NCEP–GEFS forecasts of enhanced IVT magnitudes along the U.S. West Coast that are commonly observed during landfalling ARs. The skill is summarized for ensemble probability-over-threshold forecasts of IVT magnitudes $\geq 250\ kg\ m^{-1}\ s^{-1}$ (referred to in this study as $P_{250}$) in a pseudo-Hovmöller forecast lead time–latitude framework spanning the West Coast of North America ($25^\circ$–$55^\circ$N) in a similar fashion to a forecast tool referred to as the “AR Landfall Tool” (Figs. 1 and 2). Individual case study illustrations of this tool from February 2017 and February 2019 demonstrated differences in the lead-time prediction of $P_{250}$ forecasts on the synoptic scale associated with two high-impact landfalling ARs (e.g., Fig. 6; ~8–10 days as compared to ~4–5 days, respectively). This type of synoptic-scale variability and also subseasonal and WY-to-WY variability was also illustrated across all four WYs in North-Coastal California at 38°N, 123°W (Fig. 5). On average, the $P_{250}$ forecast values prior to events with IVT magnitudes $\geq 250\ kg\ m^{-1}\ s^{-1}$ at 38°N, 123°W increased from 20% at lead times of 10 days to 70% at lead times of 3 days, with WY-to-WY variability of approximately ±5% (Fig. 7). Overall these $P_{250}$ forecasts were reliable with $P_{250}$ forecasts of 50% at lead times of 7, 8, and 9 days verifying on average 46% of the time (Fig. 8a). Threshold $P_{250}$ forecasts were also relatively accurate and successful at lead times of ~8–9 days with an average success ratio $>0.5$ for $P_{250}$ forecasts $\geq 50\%$ at lead times of 8 days (Fig. 8b), an average ROC score $>0.7$ at a lead time of 10 days (Fig. 8d), and an average ETS and BSS $>0.2$ at a lead time of 7–8 days and $>0.0$ at lead times $>10$ days (Fig. 9). The skill and accuracy of the $P_{250}$ forecasts varied as a function of latitude with the highest success ratios and POD values for $P_{250}$ forecasts $\geq 50\%$ occurring on average across Northern California and Oregon ($38^\circ$–$44^\circ$N; Figs. 11a,b) where the frequency of events with IVT magnitudes $\geq 250\ kg\ m^{-1}\ s^{-1}$ was also highest (Fig. 10a). The lowest success ratio and POD values occurred on average across Southern California ($30^\circ$–$34^\circ$N; Figs. 11a, and b, and 13) where the frequency of events was lowest (Fig. 10a). When adjusted for skill due to random chance, and also considering forecasts of nonevents, the ETS and BSS suggest that the skill of $P_{250}$ forecasts was relatively higher (lower) across coastal locations across Northern (Southern) California (Figs. 11c,d and 13), with noticeably WY-to-WY variability (Fig. 12). This study also demonstrated that average $P_{250}$ forecast values prior to events and their associated POD at 38°N, 123°W (Fig. 14) both increase $>10$–20 percentage points for more intense and longer duration consecutive periods of IVT and for more intense and longer duration ARs as characterized by the Ralph et al. (2019).

The results from this study illustrate that the ensemble-derived NCEP–GEFS $P_{250}$ forecasts contain skill over random chance at lead times of ~8–10 days and are consistent with previous results by DeFlorio et al. (2018) of AR prediction skill of hindcast forecasts by the ECMWF (see their Fig. 7a) and by Nardi et al. (2018) of multimodel AR-related occurrence-based forecast skill (see their Fig. 5). The DeFlorio et al. (2018) study also related synoptic and subseasonal variability of prediction skill to subseasonal climate modes, such as those influenced by ENSO, the Arctic Oscillation, and the Pacific–North American pattern, which may be responsible for the observed synoptic, subseasonal, and WY-to-WY variability in the NCEP–GEFS $P_{250}$ forecasts in this study. For example, it is possible that these subseasonal climate modes may influence the annual frequency of longer duration and more intense landfalling ARs which was identified in Fig. 14 to influence the skill of GEFS $P_{250}$ forecasts in this study. Future work is aimed at identifying those physical processes on different spatial and temporal scales that influence the synoptic, subseasonal, and WY-to-WY variability observed in this study.

The DeFlorio et al. (2018) study also recommended that future studies “evaluate forecasts of greater than 7 days using time windows greater than 24 h, which could yield greater skill than the values reported here.” In this study, we briefly assess the success of different threshold $P_{250}$ forecast values allowing for flexibility of ±1.0° latitude and ±6 h in verifying location and time and demonstrate notable increases (i.e., percent improvements of 10%–30%) in the success ratio at 5–10-day lead times relative to strict verification metrics (Fig. 15). Relaxing the prediction requirements such as those mentioned above may still provide situational awareness, especially at longer lead times for various forecasting applications, which was the ultimate goal in creating the ensemble-derived “AR Landfall Tool” (e.g., Cordeira et al. 2017; Waliser and Cordeira 2020). In future development of these AR-related forecast tools, these allowances should be informed by individual stakeholders in order to account for application-based risk tolerances and also should be informed by characteristic errors as a function of lead time in AR landfall position (e.g., as quantified by Wick et al. 2013; Nardi et al. 2018), timing, and duration. Unfortunately, these characteristic errors are also fundamentally related to the chosen NWP model and, for an ensemble probability-over-threshold forecasts, should contain enough ensemble spread to represent the standard errors of the ensemble mean. Future work should also seek to compare the NCEP–GEFS results herein based on version 11.0.0 to version 12.0.0 of the NCEP–GEFS that became operational in September 2020 and other ensemble NWP models (e.g., the ECMWF Ensemble Prediction System), develop methods to combine output from multiple ensemble prediction systems, and calibrate their output in order to correct for underdispersion and improve their statistical consistency (e.g., Chapman et al. 2019). These studies could then inform the development of
more reliable and accurate tools for situational awareness in advance of landfalling ARs along the U.S. West Coast.

Acknowledgments. Support for this project was provided by awards from the State of California, Department of Water Resources (4600013361) and the U.S. Army Corps of Engineers (W912HZ-15-2-0019, W912HZ-19-2-0023) as part of broader projects led by the Center for Western Weather and Water Extremes (CW3E) at the University of California, San Diego Scripps Institution of Oceanography. We greatly acknowledge feedback by Luca Delle Monache (CW3E), Michael DeFlorio (CW3E), William Chapman (CW3E), and three anonymous reviewers that improved the quality of this manuscript.

Data availability statement. Data analyzed in this study were a reanalysis and derivation of existing data, which are openly available at locations cited in the data and methods section.

REFERENCES


