Skill of Rain–Snow Level Forecasts for Landfalling Atmospheric Rivers: A Multimodel Assessment Using California’s Network of Vertically Profiling Radars

BRIAN HENN, RACHEL WEIHS, ANDREW C. MARTIN, F. MARTIN RALPH, AND TASHIANA OSBORNE

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

(Manuscript received 12 November 2018, in final form 4 December 2019)

ABSTRACT

The partitioning of rain and snow during atmospheric river (AR) storms is a critical factor in flood forecasting, water resources planning, and reservoir operations. Forecasts of atmospheric rain–snow levels from December 2016 to March 2017, a period of active AR landfalls, are evaluated using 19 profiling radars in California. Three forecast model products are assessed: a global forecast model downscaled to 3-km grid spacing, 4-km river forecast center operational forecasts, and 50-km global ensemble reforecasts. Model forecasts of the rain–snow level are compared with observations of rain–snow melting-level brightband heights. Models produce mean bias magnitudes of less than 200 m across a range of forecast lead times. Error magnitudes increase with lead time and are similar between models, averaging 342 m for lead times of 24 h or less and growing to 700–800 m for lead times of greater than 144 h. Observed extremes in the rain–snow level are underestimated, particularly for warmer events, and the magnitude of errors increases with rain–snow level. Storms with high rain–snow levels are correlated with larger observed precipitation rates in Sierra Nevada watersheds. Flood risk increases with rain–snow levels, not only because a greater fraction of the watershed receives rain, but also because warmer storms carry greater water vapor and thus can produce heavier precipitation. The uncertainty of flood forecasts grows nonlinearly with the rain–snow level for these reasons as well. High rain–snow level ARs are a major flood hazard in California and are projected to be more prevalent with climate warming.

1. Introduction

In the western United States, atmospheric river (AR) storms are responsible for much of the precipitation and snowpack (Ralph and Dettinger 2011; Guan et al. 2010), with most precipitation occurring in a few major events annually (Ralph et al. 2012). AR storms bring rain and snow for water resources and ecological services, but also pose significant flood risk from copious precipitation (Ralph et al. 2019). One factor that can distinguish between ARs that are beneficial (replenishing reservoirs, soil moisture deficits, and snowpack) and those that pose a flood hazard is the partitioning of rain and snow in mountain watersheds. This partitioning is controlled by temperatures during the AR, that is, the vertical level at which falling precipitation melts from snow into rain (the atmospheric rain–snow level \(Z_{RS}\)) and where it intersects topography.

Forecasting the atmospheric rain–snow level \(Z_{RS}\) during winter storms is critical for reservoir operations, flood control, and water supply management (White et al. 2002), because it determines the relative proportions of precipitation available for short-term streamflow generation and longer-term snow accumulation, and the likelihood of rain-driven melt of existing snowpack. In California cool-season precipitation events, atmospheric rain–snow levels typically vary from 1000 to 3500 m above sea level (Hatchett et al. 2017). Given that most topography of the Sierra Nevada mountain range, a key water supply region for the state, also falls within this range, proportions of watershed areas receiving rain and snow can vary between and within ARs. Warm ARs with relatively high \(Z_{RS}\) are associated with the largest floods in the Sierra Nevada (Dettinger 2004), have been observed to be more prevalent in recent decades (Hatchett et al. 2017), and are also projected to increase due to anthropogenic warming of the climate (Swain et al. 2018; Gershunov et al. 2019), exacerbating flood risk and reducing the reliability of California’s water supply.
provided by Sierra Nevada snowpack (Knowles and Cayan 2004).

To provide streamflow and flood forecasts, operational weather model forecasts of the atmospheric rain–snow level are used by the National Weather Service (NWS) river forecast centers in conjunction with precipitation forecasts. These streamflow forecasts are used by reservoir operators and emergency managers to make flood management decisions, such as during the sequence of ARs (with high observed rain–snow levels) that drove the spillway crisis at Oroville Dam in California in February 2017 (B. Henn et al. 2020, manuscript submitted to Geophys. Res. Lett.; White et al. 2019). The $Z_{RS}$ forecasts may come from global models with grid spacing on the scale of 25–50 km, or from higher-resolution (grid spacing of 5 km or less), regionally downscaled forecasts.

There has been limited evaluation of $Z_{RS}$ forecast skill during ARs. White et al. (2010) evaluated forecasts from one weather model at two locations in the Sierra Nevada, finding errors on the order of 300–900 m during ARs with high $Z_{RS}$. Neiman et al. (2014) reported error magnitudes averaging 500 m at one radar location in Washington State in winter 2014. Both studies evaluated the forecasts produced by NWS river forecast centers at 1–3-day lead times and used ground-based profiling radars to estimate rain–snow levels from the bright band produced by the reflectivity of snowflakes melting into rain (White et al. 2002). Using radar brightband height to estimate the rain–snow level (section 2b) is advantageous because it provides high temporal resolution observations within ARs over land, which are not reliably available from standard atmospheric soundings, ground weather stations, or from satellite observations (Cannon et al. 2017). However, atmospheric rain–snow level is often not directly output by forecast models and may need to be estimated from the height of the 0°C isotherm (section 2d).

Gaps remain in the understanding of atmospheric rain–snow level forecast skill. There has not been systematic evaluation of rain–snow levels forecast skill across models of different spatial scale; Minder and Kingsmill (2013) evaluated the rain–snow level simulations from the Weather Research and Forecasting (WRF) Model (Skamarock and Klemp 2008) against three profiling radar locations in the Sierra Nevada, finding that model microphysics parameterizations and (to a lesser extent) grid spacing had impacts on the forecast rain–snow level. Given that forecast products with a range of model grid spacings and forecast lengths are available operationally, it would be useful to know the extent to which higher-resolution models (typically with a shorter forecast period) provide more reliable forecasts. Additionally, previous work has suggested that there is correlation between rain–snow level forecast errors and AR characteristics, with warmer conditions and higher precipitation rates associated with greater rain–snow level errors (White et al. 2010; Neiman et al. 2014). This poses a threat to accurate flood forecasting due to the compounding nature of precipitation and rain–snow level errors. Finally, the reliability of ensemble rain–snow level forecasts, which estimate forecast uncertainty using perturbed model realizations, has not been evaluated. Thus, while $Z_{RS}$ forecasts are used by streamflow forecasters, reservoir operators, and emergency managers, there is limited understanding of their skill and uncertainty.

The goal this study is to assess forecast skill of $Z_{RS}$ and its relation to streamflow forecasting using vertically profiling radar sites in California during winter 2016/17. This period featured an unusually large number of strong AR storms (Vano et al. 2019; White et al. 2019; B. Henn et al. 2020, manuscript submitted to Geophys. Res. Lett.). We test three hypotheses: 1) that higher-resolution regional models provide more accurate $Z_{RS}$ forecasts than coarser global forecast models; 2) that $Z_{RS}$ forecast errors (both biases and absolute errors) are positively correlated with observed and forecast rain–snow levels, such that forecasters might have better information to interpret the reliability of certain forecasts; and 3) that a global forecast ensemble provides a reliable estimation of the uncertainty of $Z_{RS}$ forecast errors. Testing these hypotheses and synthesizing the results into expected hydrologic impacts will help water managers understand the flood forecast uncertainty associated with extreme $Z_{RS}$ events and ARs.

The remainder of the paper describes forecast and observational rain–snow level and other data and the methods used to evaluate the forecasts (section 2), the tests of the three hypotheses about rain–snow level forecasts (section 3), a discussion of rain–snow level forecast robustness and the flood risk associated with forecast errors (section 4), and a summary of key findings (section 5).

2. Data and methods

a. Atmospheric rain–snow level forecasts

We evaluate $Z_{RS}$ forecasts from three sources over California during winter 2016/17: regional WRF Model downscaled forecasts at 3-km grid spacing produced by the Center for Western Weather and Water Extremes (CW3E) at Scripps Institution of Oceanography, University of California, San Diego, called West-WRF (Martin et al. 2018); $Z_{RS}$ forecasts provided by the NWS California Nevada River Forecast Center (herein referred to as
CNRFC forecasts) used in operational streamflow forecasting, and ensemble $Z_{RS}$ forecasts of the NWS Global Ensemble Forecast System (GEFS), version 2 (Hamill et al. 2013).

1) WEST-WRF FORECASTS

Since winter 2015/16, CW3E has produced downscaled weather forecasts over California from December to March using the WRF-ARW model (version 3.8), in a configuration called West-WRF (Martin et al. 2018). We use 117 forecast initializations from 12 December 2016 to 28 March 2017, initialized either daily (1200 UTC) or twice daily (0000 and 1200 UTC) from NWS Global Forecasting System (GFS) boundary and initial conditions at 0.5° grid spacing. The GFS boundary conditions drive the outer West-WRF domain, which at 9-km grid spacing covers much of the northeast Pacific approximately from 164°W to 108°W and from 19° to 48°N. In turn, the outer domain is used to produce a 3-km inner domain, which covers California and the coastal Pacific as far offshore as 130°W. The inner domain is used for all forecasts evaluated here and is vertically discretized with 36 levels and a terrain-following coordinate system, resulting in about 15 vertical points in the lowest 5 km of the atmosphere where the rain–snow transition occurs, though most of these points are within the lowest 2–3 km. The newer Thompson microphysics parameterization (Thompson et al. 2008) is used. Model output is provided every 3 h to between 168 and 216 h, depending on the initialization. For more details on West-WRF including other parameterizations, see Martin et al. (2018).

We extract the height of the 0°C isotherm, or $Z_{0°C}$, a value that is computed by WRF via vertical interpolation, using the highest 0°C height in cases of multiple occurrences. Three-hourly forecast time series of $Z_{0°C}$ are extracted from the 3-km grid cells containing the 19 radar profilers used in this study (section 2b). To establish the offset between $Z_{0°C}$ and $Z_{RS}$, we also analyze other West-WRF forecast variables, including the mixing ratios of rain, snow, ice, and graupel in the vertical model columns of the radar profiler grid cells (see section 2d).

2) CALIFORNIA NEVADA RIVER FORECAST CENTER OPERATIONAL FORECASTS

CNRFC forecasts of $Z_{0°C}$ are made using multiple model guidance products and forecaster input. Depending on the lead time, these include global forecast models (GFS and that of the European Centre for Medium-Range Weather Forecasting), as well as regional models. CNRFC issues $Z_{0°C}$ forecasts four times per day on a 4-km grid. The forecasts are 6-hourly and most extend up to 6 days’ (144 h) lead time, though some extend to 126, 132, or 138 h. We use 286 forecast initializations from 1 December 2016 to 31 March 2017. On all but one day, the 1200 and 1800 UTC archived forecasts were available; the 0000 (26 days) and 0600 UTC (20 days) forecasts were also intermittently archived and were used as well. These archived $Z_{0°C}$ forecasts are available from the CNRFC’s website (cnrfc.noaa.gov).

3) GLOBAL ENSEMBLE FORECAST SYSTEM REFORECAST VERSION 2 FORECASTS

To evaluate an ensemble forecast of $Z_{RS}$, we use the Global Ensemble Forecast System Reforecast version 2 (GEFSRv2: Hamill et al. 2013). The GEFSRv2 is a long-term reforecast using version 9.0.1 of the GEFS, with 10 ensemble members plus a control simulation (11 forecasts). It is provided by the National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Physical Sciences Division (NOAA ESRL/PSD; esrl.noaa.gov/psd/forecasts/reforecast2). The horizontal grid spacing of the forecasts is approximately 50 km, and forecasts are provided daily. GEFSRv2 data (as opposed to archived GEFS data) are used here because GEFS $Z_{0°C}$ is not archived, and archived temperatures are provided at insufficient vertical grid points for robustly estimating $Z_{0°C}$ in that dataset. GEFSRv2 does not archive $Z_{0°C}$, but it does provide temperatures at more vertical levels (five pressure levels between 1000 and 500 hPa plus eight hybrid eta/pressure levels generally below 700 hPa), allowing for better estimation of $Z_{0°C}$ via linear interpolation of the pressure- and eta-level temperatures against geopotential height, which are provided as model outputs. This results in variable vertical grid spacing of 200–500 m in the lower atmosphere. As with West-WRF, when the temperature profile crosses 0°C multiple times, the greatest height is used; in general, this occurs during ground-level temperature inversions associated with dry weather that have limited hydrologic impacts. We use 106 daily forecast initializations between 12 December 2016 and 27 March 2017. Each forecast provides 6-hourly $Z_{0°C}$ to 192-h lead time.

b. Atmospheric rain–snow level observations from profiling radars

We use vertically profiling radar brightband observations (White et al. 2002) as observations of $Z_{RS}$ that is, the level at which snowflakes melt as they fall through the atmospheric column. Brightband heights estimated using the White et al. (2002) algorithm, which scans radar range gates for reflectivity maxima and increases in Doppler fall speed associated with melt, are made available by NOAA ESRL/PSD (ftp1.esrl.noaa.gov).
These heights estimate the middle of the bright band where approximately half of hydrometeor mass has melted; the bright band itself may be tens or even hundreds of meters thick.

California has an array of profiling radars capable of observing bright bands (White et al. 2013). Figure 1a shows locations of 19 profiling radars in California at which $Z_{RS}$ was computed during winter 2016/17. These sites include four types of radars: 915- and 449-MHz wind profiling radars, 3-GHz (S-band) precipitation radars, and low-power frequency modulated continuous wave (FMCW; 915 MHz) radars that observe bright bands. Table 1 provides a description of each radar’s name, site code shown on Fig. 1, frequency, reporting time step, elevation, and number of 6-h periods from December 2016 to March 2017 in which bright bands were observed.

Figure 1b shows the distributions of $Z_{RS}$ at each radar, ordered from north to south. This shows a climatic gradient with colder (lower $Z_{RS}$) air in the north during ARs relative to Southern California, with about 700 m average difference. Inland or mountainous locations also appear to have lower mean $Z_{RS}$ relatively to coastal locations at similar latitudes. However, of greater importance here is the large range of $Z_{RS}$ at each site, from a few hundred to >3500 m above sea level. Thus, the variability of $Z_{RS}$ between and within ARs is much larger than spatial differences in median $Z_{RS}$, suggesting that $Z_{RS}$ is critical to forecasting storm impacts across the Sierra Nevada.

We use all radar $Z_{RS}$ observations over this time period, without instituting a requirement that they be made during an AR storm in particular. Other work has shown that December 2016–March 2017 was a period many AR landfalls in California (Vano et al. 2019; White et al. 2019; Hatchett et al. 2017; Ralph et al. 2019), in which “AR conditions” (local atmospheric integrated vapor transport of at least 250 kg m$^{-1}$ s$^{-1}$; Rutz et al. 2014) were frequently observed over California.

c. Watershed topography and precipitation

Figure 1a also shows outlines of the watersheds of eight major flood control and water supply reservoirs in the Sierra Nevada, and an outline of the entire mountain range. Figure 1c shows the elevation distributions of each watershed and the entire range. Watershed topographic distributions use a 30-m resolution U.S. Geological Survey digital elevation model. There is overlap between the range of the watersheds’ elevations and the range of observed $Z_{RS}$, with the watersheds trending higher to the south at a rate
similar to the climatic gradient of $Z_{RS}$. Thus, most major reservoir watersheds of the Sierra Nevada are sensitive to variability in $Z_{RS}$; during an AR, a majority of the precipitation may be either snow or rain, depending on $Z_{RS}$. The sensitivity of hydrologic impacts to $Z_{RS}$ also will be specific to the watershed and AR.

We use precipitation data from December 2016 through March 2017 to evaluate the relationship between $Z_{RS}$ and hydrologic impacts. These are the CNRFC’s grid-ded precipitation estimates at 4-km grid spacing at 6-hourly intervals (archived at cnrfc.noaa.gov). The precipitation grids are based on gauge observations interpolated using climatological relationships between precipitation and topography (Daly et al. 2008). We compute 6-hourly precipitation accumulations by averaging grid cells inside each watershed.

d. Evaluating atmospheric rain–snow level forecasts against observations

1) TEMPORAL AVERAGING

We evaluate forecasts from each of the models against brightband $Z_{RS}$ observations at each profiling radar. The differences in time steps between profiling radar types (10 min, 15 min, and hourly) and between models (3-hourly for West-WRF and 6-hourly for CNRFC and GEFSRv2) requires averaging for consistent comparison. Additionally, the intermittent nature of brightband observations means that they are often not available at model valid times. Thus, we average $Z_{RS}$ observations using a nearest neighbor approach; for example, for 6-hourly model output, we average observations occurring 3 h before or 3 h after the model valid time. These criteria ensure enough valid observation times for evaluation across all radar sites. However, they may also introduce sampling biases and mismatches due to timing errors, particularly for extreme events where timing errors may result in forecasts underestimating their magnitude. Diagnosing timing errors is challenging and beyond the scope of this study.

2) ACCOUNTING FOR DIFFERENCES BETWEEN $Z_{RS}$ AND $Z_{0^\circ C}$

One important difference is between $Z_{0^\circ C}$ (the height of the $0^\circ C$ isotherm) and $Z_{RS}$ (the height at which approximately half of the mass of frozen hydrometeors have melted, which is estimated as the radar bright-band height). Studies comparing free atmosphere sounding profiles with radar bright bands have shown that $Z_{0^\circ C}$ tends to be 100–300 m above $Z_{RS}$; the average difference between the two was found to be 192 m (White et al. 2002), with $Z_{RS}$ at a temperature between 0°C and 3°C (Lundquist et al. 2008; White et al. 2010). At times, this difference may be larger due to airflow and microphysical dynamics (Marwitz 1983; Minder et al. 2011; Stewart 1985). However, because the CNRFC and GEFSRv2 output $Z_{0^\circ C}$ but not $Z_{RS}$, an approximation of $Z_{RS}$ using $Z_{0^\circ C}$ is required for intermodel comparison.

We use vertical profiles of the mixing ratios of rain, snow, ice, and graupel from West-WRF at radar grid cells to compute profiles of the frozen fraction of

<table>
<thead>
<tr>
<th>Site name</th>
<th>Site code</th>
<th>Radar type</th>
<th>Time step</th>
<th>Latitude (°)</th>
<th>Longitude (°)</th>
<th>Elevation (m)</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy Camp</td>
<td>hcp</td>
<td>FMCW</td>
<td>10 min</td>
<td>41.79</td>
<td>−123.39</td>
<td>366</td>
<td>197</td>
</tr>
<tr>
<td>McKinleyville</td>
<td>acv</td>
<td>449 MHz</td>
<td>1 h</td>
<td>40.97</td>
<td>−124.11</td>
<td>56</td>
<td>136</td>
</tr>
<tr>
<td>Shasta Dam</td>
<td>std</td>
<td>FMCW</td>
<td>10 min</td>
<td>40.72</td>
<td>−122.43</td>
<td>202</td>
<td>163</td>
</tr>
<tr>
<td>Chico</td>
<td>cco</td>
<td>915 MHz</td>
<td>1 h</td>
<td>39.70</td>
<td>−121.91</td>
<td>42</td>
<td>95</td>
</tr>
<tr>
<td>Oroville</td>
<td>ovl</td>
<td>FMCW</td>
<td>10 min</td>
<td>39.53</td>
<td>−121.49</td>
<td>114</td>
<td>146</td>
</tr>
<tr>
<td>Colfax</td>
<td>cff</td>
<td>FMCW</td>
<td>10 min</td>
<td>39.08</td>
<td>−120.94</td>
<td>644</td>
<td>138</td>
</tr>
<tr>
<td>Hopland</td>
<td>hop</td>
<td>S band</td>
<td>15 min</td>
<td>39.00</td>
<td>−123.09</td>
<td>253</td>
<td>169</td>
</tr>
<tr>
<td>Middletown</td>
<td>mdt</td>
<td>S band</td>
<td>15 min</td>
<td>38.75</td>
<td>−122.71</td>
<td>972</td>
<td>122</td>
</tr>
<tr>
<td>Cazadero</td>
<td>cuz</td>
<td>S band</td>
<td>15 min</td>
<td>38.61</td>
<td>−123.22</td>
<td>478</td>
<td>155</td>
</tr>
<tr>
<td>Saint Helena</td>
<td>sth</td>
<td>FMCW</td>
<td>10 min</td>
<td>38.55</td>
<td>−122.49</td>
<td>135</td>
<td>54</td>
</tr>
<tr>
<td>Santa Rosa</td>
<td>str</td>
<td>S band</td>
<td>15 min</td>
<td>38.52</td>
<td>−122.80</td>
<td>32</td>
<td>133</td>
</tr>
<tr>
<td>Bodega Bay</td>
<td>bby</td>
<td>449 MHz</td>
<td>1 h</td>
<td>38.32</td>
<td>−123.07</td>
<td>15</td>
<td>69</td>
</tr>
<tr>
<td>New Exchequer</td>
<td>ner</td>
<td>FMCW</td>
<td>10 min</td>
<td>37.60</td>
<td>−120.28</td>
<td>259</td>
<td>121</td>
</tr>
<tr>
<td>San Luis Reservoir</td>
<td>slr</td>
<td>FMCW</td>
<td>10 min</td>
<td>37.06</td>
<td>−121.07</td>
<td>81</td>
<td>113</td>
</tr>
<tr>
<td>Pine Flat Dam</td>
<td>pfd</td>
<td>FMCW</td>
<td>10 min</td>
<td>36.83</td>
<td>−119.33</td>
<td>184</td>
<td>113</td>
</tr>
<tr>
<td>Visalia</td>
<td>vis</td>
<td>915 MHz</td>
<td>1 h</td>
<td>36.31</td>
<td>−119.39</td>
<td>81</td>
<td>73</td>
</tr>
<tr>
<td>Point Sur</td>
<td>pts</td>
<td>449 MHz</td>
<td>1 h</td>
<td>36.30</td>
<td>−121.89</td>
<td>10</td>
<td>42</td>
</tr>
<tr>
<td>Kernville</td>
<td>knv</td>
<td>FMCW</td>
<td>10 min</td>
<td>35.75</td>
<td>−118.42</td>
<td>816</td>
<td>72</td>
</tr>
<tr>
<td>San Bernardino</td>
<td>sbo</td>
<td>FMCW</td>
<td>10 min</td>
<td>34.20</td>
<td>−117.34</td>
<td>602</td>
<td>105</td>
</tr>
</tbody>
</table>

Total 2216
hydrometeor mass, and the height at which it declines to 50%. This level is referred to as $Z_{Q50}$, and should approximate $Z_{RS}$ (Minder et al. 2011). We compare West-WRF $Z_{Q50}$ and $Z_0$ in Fig. 2, and in Fig. A1 in the appendix. Figure 2a shows both variables in a forecast at the Oroville radar from January 2017. Figure 2b displays relative occurrences of both and indicates a consistent offset between them. Offsets are similar between radar grid cells (Figs. 2c, A1), with variability of about ±25 m between sites. The average across all sites and forecasts is 207 m; we also estimate the uncertainty in the offset using its interquartile range (138–268 m).

Due to the similarities between the estimation of the $Z_{RS} - Z_0$ offset and past observational studies, we use 207 m (138–268 m) as an approximation. Therefore, we evaluate all $Z_{RS}$ forecasts in this study by

\[ Z_{RS,\text{forecast}} = Z_{0,\text{forecast}} - 207 \text{ m (138–268 m)}, \]  

where $Z_{RS,\text{forecast}}$ is the forecast to be compared against radar $Z_{RS}$ observations, and $Z_{0,\text{forecast}}$ is the model 0°C isotherm height, and values in parentheses indicate uncertainty ranges.

3) EVALUATION METRICS

We evaluate $Z_{RS}$ forecast skill for each radar site, with respect to forecast lead time, quintiles of $Z_{RS}$ (both forecast and observed), and observed precipitation rates. We subset the data by quartiles for both forecast and observed $Z_{RS}$ to present error statistics conditional on a forecast of high or low $Z_{RS}$ (i.e., from the perspective of a forecaster), and conditional on observed high or low $Z_{RS}$ events. Quartiles of $Z_{RS}$ are computed across a pooled dataset of all profiling radar sites to reduce the impact of outliers.

We present root-mean-square error (RMSE), bias, and correlation coefficients between forecasts and observations. Bias is calculated as the mean error:

\[ \text{bias} = \text{mean}(Z_{RS,\text{forecast}} - Z_{RS,\text{observed}}), \]  

where $i$ indexes observation–forecast pairs. For GEFSRv2, these statistics are calculated using forecasts from all ensemble members. For each metric, 95% confidence intervals (CIs) are calculated, assuming forecast errors follow a normal distribution, and that the sum of squared errors is chi-square distributed. The CIs allow...
for assessing the significance of differences between model metrics.

3. Results

a. Evaluation of atmospheric rain–snow level forecast skill across models

1) WEST-WRF

Figure 3 shows scatterplots of 3-km West-WRF forecasts against $Z_{RS}$ observed at the 19 profiling radars, with colors indicating forecast lead time. The 1:1 (dashed) and linear regression line for 0–24-h lead times (dotted, with West-WRF as predictor) are also shown. Shading indicates areas that are below the ground elevation of the profiling radar.

![Figure 3: Scatterplots of radar-observed $Z_{RS}$ against West-WRF forecast $Z_{RS}$, for each of the 19 profiling radars, with colors indicating forecast lead time. The 1:1 (dashed) and linear regression line for 0–24-h lead times (dotted, with West-WRF as predictor) are also shown. Shading indicates areas that are below the ground elevation of the profiling radar.](image-url)

We summarize RMSE and bias of West-WRF $Z_{RS}$ forecasts across sites and lead times (Fig. 4). Figure 4a shows that West-WRF bias is generally between 0 and $-300$ m, though there is variability between sites: the two southernmost sites [Kernville (knv) and San Bernardino (sbo)]...
Bernardino (sbo) exhibit positive biases (West-WRF forecast $Z_{RS}$ too high), particularly at longer lead times. The 95% CIs (error bars) show that for most sites and lead times, bias is significantly different from zero. RMSE (Fig. 4b) averages over 700 m for lead times over 144 h, but declines with decreasing lead times, to about 350 m for 24 h or less. Similar to bias, greater RMSE is seen for San Bernardino and Point Sur (pts) than at the central and Northern California sites. Comparing the ~350-m RMSE magnitude to the observed $Z_{RS}$ range of ~3000 m suggests the model has skill over simple climatology in forecasting $Z_{RS}$.

Next, we present West-WRF evaluations stratified by observed $Z_{RS}$ quintiles. The quintiles are computed by

Fig. 4. (a) Bias in West-WRF $Z_{RS}$ forecasts with respect to profiling radar and lead time. Dashed lines indicate average of data from all sites. (b) As in (a), but with RMSE for West-WRF $Z_{RS}$ forecasts. (c) Bias in West-WRF $Z_{RS}$ forecasts averaged over all profiling radars, separated into quintiles of observed $Z_{RS}$ values and by lead time. (d) As in (c), but for RMSE. (e) Bias in West-WRF $Z_{RS}$ forecasts, separated into quintiles of forecast $Z_{RS}$ values and by lead time. (f) As in (e), but for RMSE. Shaded areas on bias plots indicate West-WRF $Z_{RS}$ uncertainty.
sorting observed $Z_{RS}$ values for which there are valid forecasts and dividing the observations into equally sized groups. The observed quintiles average from 918 m (lowest, colder ARs) to 2776 m (highest, warmer ARs). West-WRF $Z_{RS}$ bias depends on the quintile (Fig. 4c); at lead times of up to 24 h, the lowest quintile has bias of $-16$ m, within the range of uncertainty in the offset (shaded regions in Fig. 4), decreasing to $-330$ m for the highest quintile, a greater magnitude than the uncertainty range. For lead times of $>144$ h, bias ranges from 168 m for the lowest to $-508$ m for the highest quintile; the middle quintiles are within the range of uncertainty. Thus, there is a tendency to underforecast observed high $Z_{RS}$ events. RMSE is less sensitive to observed $Z_{RS}$ quintile (Fig. 4d), though for longer lead times RMSE is greater for warmer ARs.

In Figs. 4e and 4f, the evaluation is stratified by forecast $Z_{RS}$ (as opposed to observed). Here, all forecasts with valid observations are sorted into five equal groups. In the case of longer lead times, the sign of the bias is opposite that of the observed $Z_{RS}$ quintiles: 377-m bias for the highest forecast quintile and $-620$ m for the lowest forecast quintile for lead times $>144$ h. Thus, high $Z_{RS}$ forecasts tend to overforecast observed conditions, and low forecasts underforecast observations. Also, the highest and lowest quintiles have the greatest RMSEs (Fig. 4f), that is, the most extreme events. The CIs indicate that these differences are significant at $p < 0.05$. Together, these results suggest challenges in skillfully forecasting $Z_{RS}$ anomalies, particularly higher $Z_{RS}$ (warm) AR events.

2) CNRFC

Figure 5 presents forecast skill as with Fig. 4, but for the CNRFC $Z_{RS}$ forecasts. The forecast metrics are qualitatively similar between CNRFC and West-WRF. Forecast bias averaged across all sites is slightly negative for each lead time category ($-61$, $-56$, and $-73$ m, for 0–24, 30–72, and 78–144 h, respectively, within the uncertainty range). Like West-WRF, regional variability is most prominent in Southern California (Point Sur, Kernville, and San Bernardino) where biases are positive, whereas other sites are negative. However, CNRFC forecast biases are slightly smaller in magnitude than for West-WRF. RMSE (Fig. 5b) increases with lead time, from 328 m for lead times of 24 h or less to 506 m for lead times of 78–144 h, with the smallest magnitude errors in Northern California [e.g., McKinleyville (acv) and Chico (cco)].

Also like West-WRF, CNRFC forecast biases exhibit opposite sensitivities with respect to forecast and observed $Z_{RS}$ quintiles (Figs. 5c and 5e). The bias is positive for lower observed quintiles and negative for higher observed quintiles, though they are small and within the uncertainty range. For forecast $Z_{RS}$ quintiles, the opposite relationship is observed, significantly for lead times of 78–144 h. CNRFC forecast RMSEs are not strongly sensitive to the observed quintile (Fig. 5d), but are greatest for the highest forecast quintile, and for the lowest forecast quintile at lead times of 78–144 h (Fig. 5f). Again, this suggests that anomalous events, particularly ARs with high forecast $Z_{RS}$, suffer from the largest forecast errors.

3) GEFSRv2

GEFSRv2 $Z_{RS}$ forecast evaluation is summarized in Fig. 6, suggesting that GEFSRv2 skill is also broadly like West-WRF and CNRFC. Forecast biases (Fig. 6a) are less negative than for West-WRF and CNRFC, though the difference is within the range of uncertainty of the offset. A similar geographic distribution of biases is seen as with the other models. Forecast RMSE (Fig. 6b) increases with lead time from 200 to 400 m at less than 24 h lead to 700–800 m for lead times of 144 h or greater. Long-lead bias ranges from 376 m for the lowest observed $Z_{RS}$ quintile to $-255$ m for the second-highest (Fig. 6c). RMSE is less sensitive to observed $Z_{RS}$ quintile (Fig. 6d) but has the greatest values for the lower or second-lowest quintiles. Forecast $Z_{RS}$ bias ranges from $-578$ m for the lowest quintile to 889 m for the highest for $>144$ h lead time (Fig. 6e). Forecast RMSE is highest for the lowest and highest quintiles at longer lead times (Fig. 6f).

After considering each model’s error statistics, we compare them in Fig. 7. Bias (Fig. 7a) varies in sign between models, with West-WRF (from $-108$ to $-187$ m) and CNRFC (from $-90$ to $-110$ m) averaging negative biases that exceed the uncertainty range. (Note that no forecast skill metrics are available for CNRFC forecasts beyond $144$ h and those values are omitted from Fig. 7). GEFSRv2 biases are within the uncertainty range. Overall, the magnitude of the average biases (generally less than 200 m, even at longer lead times) when compared to the large observed variability in $Z_{RS}$ (~3000 m, Fig. 1) suggests that challenges in forecasting $Z_{RS}$ are not due to average biases. The magnitudes of RMSE (Fig. 7b), however, are larger than the biases. These magnitudes are similar across models, with West-WRF and the CNRFC having slightly lower RMSE than GEFSRv2. For example, at 27–72-h lead time, RMSE is 363 m for West-WRF, 370 m for CNRFC, and 410 m for GEFSRv2.

When considering that West-WRF has both a greater (negative) bias magnitude and a lower RMSE, it must be that West-WRF has a higher correlation with observations than the other coarser models, and this is verified in Fig. 7c. Again considering 27–72-h lead times, the average correlation with observations is 0.866 for West-WRF, 0.854 for CNRFC, and 0.828 for
These differences are small, but statistically significant at $p = 0.05$. The higher West-WRF correlation is evidence for marginally improved $Z_{RS}$ forecast skill relative to the coarser models.

b. Precipitation rate and rain–snow level forecast relationship

During ARs with heavy precipitation, $Z_{RS}$ is hydrologically most significant, as it dictates partitioning into rain and snow and the balance of flood hazards and water supply benefits. All observations of $Z_{RS}$ used here occur during precipitation periods, as the profiling radars observe $Z_{RS}$ via radar reflection of melting snowflakes. However, the relationship between precipitation rate and forecast skill of $Z_{RS}$ is important because much of the West Coast's precipitation occurs during the largest ARs (Dettinger et al. 2011) that are often associated with warmer
temperatures (high $Z_{RS}$) and enhanced flood risk (Neiman et al. 2014).

We evaluate the relationship between precipitation rate and both $Z_{RS}$ and the forecast skill of $Z_{RS}$ in Fig. 8. Six-hourly watershed-average precipitation rates for five major reservoirs in the Sierra Nevada (Shasta Lake, Lake Oroville, Folsom Lake, Lake McClure, and Pine Flat Lake) are plotted against observed $Z_{RS}$ at adjacent profiling radars (Fig. 8a; the radars are Shasta Dam, Oroville, Colfax, New Exchequer, and Pine Flat Dam, respectively). For each watershed, we identify terciles of precipitation rates (vertical gray lines in Fig. 8a; terciles are used here instead of quintiles because of the smaller sample sizes of the 6-hourly precipitation observations). For times when a minimum precipitation rate threshold was met (0.8 mm h$^{-1}$, to screen out times of zero or very light precipitation), and $Z_{RS}$ was observed, an exponential regression model was fit (blue line; this type of model was chosen because of exponential Clausius–Clapeyron relationship between temperature

![Fig. 6. As in Figs. 4 and 5, but for the GEFSRv2 $Z_{RS}$ forecasts. Statistics shown here reflect all GEFSRv2 ensemble members.](https://example.com/f6.png)
and atmospheric water vapor capacity). The greatest precipitation rates (e.g., >5 mm h\(^{-1}\)) occur almost exclusively for relatively high observed \(Z_{\text{RS}}\) over 2000 m. Though the relationship between temperature and precipitation rates is weak, there is a robust physical link between the two, and it emphasizes the importance of warm ARs in flood forecasting.

We evaluate each model for each precipitation rate tercile (Figs. 8b–d). For each model forecast, the largest and most negative biases are found during the highest precipitation tercile for the majority of these watersheds, though most differences between precipitation terciles are not significant at \(p = 0.05\). Second, the RMSE of each forecast model tends to increase from the lower precipitation terciles to the highest; in most cases, the third tercile RMSE is significantly bigger than the first tercile RMSE at \(p = 0.05\). These results are consistent with Figs. 4–6 and 8a: larger forecasts errors are correlated with high \(Z_{\text{RS}}\), and high \(Z_{\text{RS}}\) is correlated with high precipitation rates. In other words, warm ARs have the highest precipitation rates, the most negative \(Z_{\text{RS}}\) forecast biases, and the largest forecast RMSE.

c. Reliability of ensemble forecasts

Of the forecast products considered here, only GEFSRv2 is an ensemble. We assess the degree to GEFSRv2 reliably estimates the uncertainty of \(Z_{\text{RS}}\) by comparing its standard deviation against the magnitude of observed forecast \(Z_{\text{RS}}\) errors (Leutbecher and Palmer 2008).

For each profiling radar and lead time category, we plot GEFSRv2 average standard deviation (RMS differences) against RMSE from GEFSRv2 evaluation against \(Z_{\text{RS}}\) observations. For all sites and lead times, RMS differences are less than RMSE (points above 1:1 line, Fig. 9a). While there is significant site-to-site variability, RMSE most greatly exceeds RMS differences at the shortest lead times. The ratio of RMSE to RMS differences is also plotted (Fig. 9b). For lead times of 24 h or less, the ratios vary between 2.3 and 10.7, averaging 4.0. This average ratio declines with lead time, reaching 1.9 for lead times of 144 h or greater, and is significantly greater than 1 for all sites and lead times using an \(F\) test at \(p = 0.05\). Thus, GEFSRv2 (and likely thus operational GEFS also) ensemble uncertainty is insufficient to represent actual \(Z_{\text{RS}}\) uncertainty during ARs, by a factor of 4 for short lead times.

4. Discussion

a. Effects of model scale on forecast skill

As shown in Figs. 4–7, we find relatively similar performance across three forecast models with different grid spacing. West-WRF’s inner domain (3 km), operational CNRFC forecasts (derived from \(1/4\) GEFS forecasts and other, higher-resolution models, and GEFSRv2 (~50 km) all produce \(Z_{\text{RS}}\) forecasts that have mean biases with magnitudes of 0–200 m, and with RMS errors that average 350 m at short (24 h or less) lead times, and increase with lead time to 400–600 m for 75–144-h (3–6 days) lead time. At longer (>144 h) lead times, West-WRF produced forecasts with lower RMSE (averaging 718 m and ranging across sites from 550 to 952 m) than GEFSRv2 (average RMSE 755 m, range 694–1303 m). However, the similarity of the forecast models’ skill, particularly at shorter lead times where many metric differences are not significant at \(p = 0.05\), suggests that there may not be benefits to smaller grid spacing for prediction of \(Z_{\text{RS}}\), which is derived from the temperatures within the AR. Unlike precipitation, which is driven by processes that are highly scale dependent, and where increases in model resolution have produced improvements in forecast skill (e.g., Davis et al. 2006), AR temperatures fields are smoother and likely better represented at the 25–50-km scales of global models. However, we show that there is significantly lower RMSE and higher correlation with observations (but larger bias magnitude) for West-WRF than for the other two models. The improvement is small relative to the magnitude of the forecast errors, and further experiments are needed to evaluate impact of grid resolution on \(Z_{\text{RS}}\) forecast skill.
Additionally, we find that each model produced similar bias structures relative to forecast and observed \( Z_{\text{RS}} \) values. The models tended to overforecast \( Z_{\text{RS}} \) when it was observed to be relatively low, and underforecast it for higher observed values. This suggests a similar representation of variability and extremes in \( Z_{\text{RS}} \) in forecast models. Relative to forecast \( Z_{\text{RS}} \), all three models at longer lead times had positive bias for lower quintiles and negative bias for higher quintiles. In other words, conditional on a forecast of an anomalously high (or low) \( Z_{\text{RS}} \), model errors are likely to regress to climatology, such that the expected outcome is less extreme. Forecasters are left in a difficult position due to this shortcoming in model skill: observed extreme events tend to be underforecast, while high or low \( Z_{\text{RS}} \) forecasts at longer lead times tend to overestimate the true magnitude. To some extent, these challenges may be due to AR timing and location errors, which we do not evaluate here, beyond the 3- and 6-h temporal aggregation used for model evaluation. We further discuss the risk of flood forecasting errors below.

b. Relationships between \( Z_{\text{RS}} \) forecast skill and precipitation

One key finding of this study is that there is an association both between \( Z_{\text{RS}} \) and precipitation rate and
between precipitation rate and errors in $Z_{RS}$ forecasts. First, $Z_{RS}$ and precipitation rate observations suggest a positive association between the two (Fig. 8), consistent with nonlinear increases in the atmosphere’s saturated water vapor capacity with temperature. Thus, as $Z_{RS}$ increases, the upper limit for precipitation rates increases nonlinearly. This is presumably one reason why most of the large historical floods in California resulted from ARs, which tend to have temperatures above cool-season wet day averages (e.g., Guan et al. 2010). Thus, errors in $Z_{RS}$ will have greater impact on forecasts of water available for streamflow when $Z_{RS}$ is high than during colder events, regardless of the magnitude of those errors. And during high $Z_{RS}$ events, the largest fractions of key Sierra Nevada watersheds receive rain and not snow (Fig. 1), such that the immediate streamflow response will nearly always be larger for these events than for colder ARs with lower $Z_{RS}$.

The increasing streamflow forecast uncertainty with higher $Z_{RS}$ is compounded by a second finding from this study: $Z_{RS}$ forecast errors tend have more negative model biases for higher observed $Z_{RS}$. Figures 4c, 5c, and 6c also suggest that underforecast biases worsen with observed $Z_{RS}$ in each model. For the events with the greatest potential for streamflow generation, there is both a tendency for larger errors in $Z_{RS}$, and for the models to underestimate $Z_{RS}$. We find that forecasts of watershed-averaged rain rates (not including solid precipitation), have error magnitudes that grow nonlinearly with forecast $Z_{RS}$ and have underestimation biases for this critical input to flood forecast models. These effects were not resolved with forecast models with finer grid spacing.

c. Implications for operational streamflow forecasting

Figure 10 schematically describes the uncertainties in flood forecasting due to $Z_{RS}$ forecast uncertainty described in this paper. During an AR, the rain–snow level intersects the topography of the Sierra Nevada, but due to forecast uncertainty, the possible elevation of the intersection spans a range (Fig. 10a). Thus, the proportion of the watershed area receiving rain versus snow is uncertain. Figure 1c shows that much of a given watershed’s topography may fall within a relatively small range (as little as 500–1000 m for some watersheds, such as that of the Feather River above Lake Oroville). Given that forecast RMSE of $Z_{RS}$ averages several hundred meters, even at short lead time, this uncertainty can cause major errors in forecast precipitation partitioning and streamflow forecasts.

Figures 10b–e illustrate how $Z_{RS}$ forecast errors, the tendency of $Z_{RS}$ to be correlated with precipitation rates, the tendency of $Z_{RS}$ forecast errors to increase
with $Z_{RS}$, and the topography of Sierra Nevada watersheds interact to exacerbate the risk of floods and flood forecasting errors. To demonstrate this context, we use $Z_{RS}$ observations from the New Exchequer radar, and topographic and precipitation data from the adjacent watershed of Lake McClure on the Merced River, which spans the central Sierra Nevada and much of Yosemite National Park. In Fig. 10b, the distribution of West-WRF

FIG. 10. (a) Schematic diagram of the rain–snow level impinging upon the Sierra Nevada during an AR. Uncertainty in the forecast of $Z_{RS}$ is bound by the dashed lines, indicating the portion of the land surface where the precipitation phase is uncertain. (b)–(e) Lake McClure (Merced River) watershed topography and precipitation and New Exchequer (ner) profiling radar observations illustrate sensitivity of flood risk to $Z_{RS}$ and its forecast errors (West-WRF forecasts with 27–72 h lead times are shown). (b) Scatterplot of $Z_{RS}$ forecast errors against observed $Z_{RS}$. For each observed $Z_{RS}$ quintile, the 90% interval of forecast errors is shown (line), with the 5th (star), median (circle), and 95th (star) percentile errors. (c) For Lake McClure watershed, the fraction of the watershed expected to receive rain as a function of $Z_{RS}$ (black line), and the 90% interval of rain fraction for each $Z_{RS}$ quintile due to the errors shown in (b), with actual values in black. (d) Scatterplot of observed $Z_{RS}$ and observed precipitation rates for Lake McClure watershed, with 90% intervals of precipitation for each $Z_{RS}$ quintile. (e) Combining watershed rain fraction and its uncertainty from (c) and median precipitation rates for each quintile from (d), watershed rain rate 90% confidence intervals for each observed $Z_{RS}$ quintile, with actual values in black.
Z\textsubscript{RS} forecast errors is shown with respect to observed Z\textsubscript{RS} (for forecasts with 24–72-h lead time, when critical flood forecast information is being developed). Negative biases in the median forecast of Z\textsubscript{RS} is apparent for all quintiles, but much more significantly for the higher quintiles. Combining these Z\textsubscript{RS} errors distributions with the topography of the Lake McClure watershed, the forecast fraction of the basin expected to receive rain for each observed Z\textsubscript{RS} quintile is plotted in Fig. 10c, again showing CIs. The uncertainty in Z\textsubscript{RS} results in uncertainties in the fraction of the basin area receiving rain. Across the quintiles, the CI of forecast rain fraction spans between 18% and 42% of the Lake Maclure watershed, highlighting the uncertainty in hydrologic response due to Z\textsubscript{RS} forecasting errors. The true values (black squares) show that rain fraction also tends to be underforecast for the warmest events.

In Figs. 10d and 10c, we also account for the observed relationship between observed Z\textsubscript{RS} and precipitation rate. Figure 10d shows this relationship for the New Exchequer profiling radar and the Lake McClure watershed. Both median precipitation rates (circles) and ranges of precipitation rates increase with Z\textsubscript{RS}. Figure 10e combines watershed rain fraction uncertainty with the increases in precipitation rates to show the total uncertainty in watershed rain input as a function of observed Z\textsubscript{RS}. This is the product of watershed rain fractions (Fig. 10c) and median rain rates for each quintile (circles, Fig. 10d), reflecting uncertainty solely due to uncertainties in Z\textsubscript{RS} (i.e., the median precipitation rate reflects the average effect of Z\textsubscript{RS}, not the additional uncertainties in forecasting variability in the precipitation rate itself). Figure 10e shows that the watershed rain rate increases in a nonlinear fashion with observed Z\textsubscript{RS} due to the combined effects described above. Uncertainty in watershed rain rate increases with Z\textsubscript{RS}: for example, CI of the watershed rain rate for the fifth quintile is 4.2–5.8 thousand acre-ft h\textsuperscript{-1} (5.2–7.1 million m\textsuperscript{3}h\textsuperscript{-1}), spanning 1.6 thousand acre-ft h\textsuperscript{-1} (1.9 million m\textsuperscript{3}h\textsuperscript{-1}). The observed average rain rate for the highest quintile is 6.0 thousand acre-ft h\textsuperscript{-1} (7.4 million m\textsuperscript{3}h\textsuperscript{-1}), illustrating the underforecasting of warm AR streamflow potential. In contrast, the middle quintile has a median watershed rain rate of 2.3 thousand acre-ft h\textsuperscript{-1} (1.1 million m\textsuperscript{3}h\textsuperscript{-1}) and CI spans 0.92 thousand acre-ft h\textsuperscript{-1} (1.1 million m\textsuperscript{3}h\textsuperscript{-1}). The lowest Z\textsubscript{RS} quintile also has significant rain rate uncertainty but the rain rate magnitude is less. For reference, Lake McClure has a flood control volume of approximately 350 thousand acre-ft (432 million m\textsuperscript{3}).

Thus, not only are warmer ARs subject to larger Z\textsubscript{RS} forecast errors, but these errors combine with watershed topography in the Sierra Nevada and precipitation rates to exacerbate flood forecast uncertainty. While each watershed and AR will have unique uncertainties, the tendency of both the topography and the climatological value of Z\textsubscript{RS} to increase from north to south means that these findings generally hold throughout the range. Warm, moisture-laden ARs pose both the greatest flood hazards and forecasting challenges, and these types of events are projected to become more severe due to growing climate variability (Swain et al. 2018).

d. Limitations of this study

We examined skill in Z\textsubscript{RS} forecasts using 19 profiling radars across California during winter 2016/17, which included 2216 six-hourly periods for forecast evaluation. Despite this sample size, there are several methodological limitations to this study. First, the study does not account for temporal or spatial mismatches in the forecasts of Z\textsubscript{RS}, since they are evaluated at the time and grid cell of the observations. The temporal averaging of Z\textsubscript{RS} observations may also introduce error into our results. Timing and location errors in forecasts of AR storms are a common feature of dynamical models (Martin et al. 2018), and as such some of the errors reported here may be due to storms in which peak Z\textsubscript{RS} may have been accurately forecasted, but the timing was incorrect by more than the 3- or 6-h evaluation time steps. Forecasters may be able to mitigate the impacts of these errors, particularly since timing errors in Z\textsubscript{RS} and precipitation may be strongly correlated (i.e., models forecast high Z\textsubscript{RS} and heavy precipitation associated with the AR to be both too early or too late). Nonetheless, timing and location errors still cause problems for operators, even if the large-scale characteristics of the AR are correctly forecast.

It is beyond the scope of this study to diagnose physical mechanisms for Z\textsubscript{RS} forecast errors. Such interpretation is important for improving models and requires examination of the physical processes associated with AR storms and their temperature. For example, our findings suggest less skill in forecasting Z\textsubscript{RS} in Southern California across the models (Figs. 3b, 4b, and 5b), which is probably only partially explained by the smaller sample of ARs in this region (Table 1). Future work should focus on the mechanisms driving Z\textsubscript{RS} forecast errors and their regional variability.

Our approach may also contain small biases in forecast skill statistics for low Z\textsubscript{RS} due to the influence of the ground. When forecast Z\textsubscript{RS} is below the ground surface, there is no valid forecast, and thus potential low model biases are excluded. Similarly, when snowfall reaches close to the ground at the radar sites, no bright band can
be observed and these cases are excluded as well. As a result, the lowest quantiles of observed (forecast) $Z_{RS}$ (Figs. 4c–f, 5c–f, and 6c–f) may overestimate (underestimate) bias due to this sampling effect. But these effects are likely small, as most radar grid cells are at low elevation (all but four sites are below 0.5 km, Fig. 1); only about 1% of West-WRF forecast $Z_{RS}$ values would be below the ground surface during the study period. Nearly all AR events in California are too warm to produce low-elevation snow, and so the fraction of observed $Z_{RS}$ that are missed for this reason is similarly small.

Additionally, because we evaluate $Z_{RS}$ forecasts using radars that are upwind of the Sierra Nevada, and not ground observations, we may not account for the differences in $Z_{RS}$ between the two. It has been shown that $Z_{RS}$ may lower near the crest of mountain ranges as AR winds cross them (Marwitz 1987; Medina et al. 2005; Minder et al. 2011). This descent of $Z_{RS}$ may result in high biases when using profiling radars to partition rain and snow in mountain watersheds (Mizukami et al. 2013). To account for this in operational streamflow models, the CNRFC uses offsets from model-output $Z_{0^\circ C}$ that are larger than the 207-m offset used in this study; these offsets are derived from calibration of hydrologic model output to streamflow observations (as opposed to direct observations of rain–snow partitioning). Due to $Z_{RS}$ descent, error statistics reported here may be somewhat different than those experienced on the ground. Ideally, model evaluation could be carried out with on the ground rain–snow partitioning as the forecast variable. However, due to the lack of comprehensive ground-based observations of $Z_{RS}$, radar-derived observations are currently the best available method for forecast evaluation. Bending of $Z_{RS}$ and its variability with AR characteristics is another area needing further research.

5. Conclusions

We evaluated forecasts of the atmospheric rain–snow level $Z_{RS}$ during December 2016–March 2017 at 19 profiling radars in California, using three forecast model products: 3-km West-WRF downscaling of a global forecast model, 4-km CNRFC operational forecasts, and 50-km GEFSRv2 ensemble forecasts. We use model output of the height of the 0°C isotherm $Z_{0^\circ C}$ minus an offset (207 m, shown to be appropriate from prior studies and model output), to compare model forecasts of the rain–snow level against observations of rain–snow bright bands at the profiling radars. We find that each model product tends to produce relatively unbiased forecasts (mean biases of less than ±250 m) across a range of forecast lead times, from 0–24 h to greater than 144 h. West-WRF is shown to have a slight cold bias at some lead times and locations, while other models have no significant bias. We find that $Z_{RS}$ RMSE increases with lead time and is consistent between models, with 0–24 h RMSEs averaging 350 m and growing to 700–900 m for lead times of greater than 144 h. We find only marginal improvements in $Z_{RS}$ correlation with finer grid spacing (West-WRF slightly more correlated). We also find that GEFSRv2 ensemble spread is insufficient to span observed forecast errors, particularly for shorter lead times though it remains insufficient at lead times of 144 h or greater.

We find that observed $Z_{RS}$ extremes are underestimated for warm (high $Z_{RS}$) ARs, and that the magnitude of errors in $Z_{RS}$ increases with $Z_{RS}$; these findings were consistent across the three forecasts models. We also find that high $Z_{RS}$ events are correlated with the highest observed precipitation rates in Sierra Nevada watersheds. Synthesizing these findings, we show that flood risk increases with $Z_{RS}$ in these watersheds, and that the uncertainty in flood risk from ARs grows nonlinearly with $Z_{RS}$, due to larger errors in rain–snow partitioning for high $Z_{RS}$ events, and the tendency for precipitation rates to increase with $Z_{RS}$.

Due to this unpredictability, high $Z_{RS}$ ARs represent a major driver of flood forecast error in California. To better predict and manage these events, improvements must be made in the representation of AR dynamics, temperature, position, and timing in operational forecast models. We did not find evidence that reducing model grid spacing resulted in improved $Z_{RS}$ forecast skill. Additional offshore observations of landfalling AR characteristics via airborne and satellite-based instruments and improved techniques to assimilate these observations into near-real-time forecasts may yield improvement in $Z_{RS}$ forecast skill. Additionally, model ensemble reliability needs to be improved, whether by statistical postprocessing or by improvements in underlying ensemble methodologies, to better represent forecast uncertainty. In the projected warmer, more variable climate, these improvements may help California better manage the risk of floods and its valuable water resources.

Acknowledgments. CW3E acknowledges support for this work from the California Department of Water Resources atmospheric river research program. We thank Arthur Henkel and the staff of the
NOAA/NWS California Nevada River Forecast Center for their help in accessing archived data and for providing manuscript feedback. We thank Gary Bates and Michael Scheuerer of NOAA ESRL/PSD for making the augmented eta-level GEFSRv2 temperature and pressure data available. We thank Caroline Papadopoulos of Scripps Institution of Oceanography, UC San Diego for assistance in running and archiving West-WRF forecasts. We are grateful to members of CW3E, including Forest Cannon, Brian Kawzenuk,
and Anna Wilson, for their helpful feedback on this study.

APPENDIX

Effects of $Z_{0^\circ C}$ Offset Uncertainty

As an extension of Fig. 2b, we calculate the two-dimensional histogram of the occurrences of $Z_{0^\circ C}$ versus $Z_{Q50}$ in West-WRF for each radar grid cell (Fig. A1). These results show broad consistency with the all-site average histogram (Fig. 2b). One common feature is the tendency for the difference between $Z_{RS}$ and $Z_{0^\circ C}$ to become small or negative for the warmest events (rain–snow levels of 2500–3000 m). To some extent, this is likely an artifact of the coarser vertical grid spacing in this region, as the spacing increases from approximately 200–300 m below this height to 500–600 m above it. Because $Z_{RS}$ is interpolated and the frozen hydrometeor fraction is much less smooth than temperature, the reduction in the difference between the two variables for warmer events may be because of the less robust interpolation of $Z_{RS}$.

Figure A2 shows West-WRF error statistics similar to those in Fig. 4, but computed using the West-WRF
output $Z_{RS}$ instead of the offset $Z_{0C}$. Statistics based on offset $Z_{0C}$ computed with the same sample set are also shown for comparison. Qualitatively, the model performance is similar using $Z_{RS}$ and offset $Z_{0C}$. There is some evidence of improved bias RMSE for longer lead times and warmer AR events; however, for short lead time the $Z_{RS}$ forecasts average slightly worse performance than offset $Z_{0C}$.

The $Z_{RS}$ is available only for West-WRF, and even then, only for the subset of cases when West-WRF predicted sufficient precipitation in order for $Z_{RS}$ to be calculated robustly (about one-third of the samples available for evaluation with $Z_{0C}$). Therefore, we present these results as an appendix but use offset $Z_{0C}$ instead of $Z_{RS}$ as the main forecast variable in this study.

REFERENCES


Skamarock, W. C., and J. B. Klemp, 2008: A time-split non-hydrostatic atmospheric model for weather research and


