Dynamical downscaling improves upon gridded precipitation products in the Sierra Nevada, California

Mimi Hughes[1], Jessica D. Lundquist[2], and Brian Henn[3]

[1]University of Colorado, Cooperative Institute for Research in Environmental Science, and NOAA/ESRL/PSD.

[2]University of Washington, Civil and Environmental Engineering.

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

Corresponding author: Mimi Hughes (mimi.hughes@noaa.gov); phone: (303)497-4865; fax: (303)497-6101

Keywords: precipitation, WRF, convection-permitting, downscaling

In preparation for Climate Dynamics (special issue on high-resolution climate modeling)

Abstract

Uncertainties in gridded and regional climate estimates of precipitation are large at high elevations, where observations are sparse and spatial variability is substantial. We explore these uncertainties for water year 2008 across California's Sierra Nevada in 10 datasets: six regional climate downscalings generated using the Weather, Research, and Forecast (WRF) model at convection-permitting resolution with differing lateral boundary conditions and microphysical parameterizations, and four gauge-based, interpolation-gridded precipitation datasets. Precipitation from these 10 datasets is evaluated against 95 snow pillows and a precipitation dataset inferred from stream gauges using a Bayesian inference method. During water year 2008, the gridded datasets tend to underestimate frozen precipitation on the windward slope of the Sierra Nevada, particularly in the vicinity of Yosemite National Park. The WRF simulations with single-moment microphysics tend to overestimate precipitation throughout much of the region, whereas the WRF simulations with double-moment microphysics tend to better agree with both the snow pillows and inferred precipitation estimates, although they somewhat overestimate the windward/leeside precipitation contrast in the northern Sierra Nevada. WRF simulations, in particular those with single-moment microphysics, better distinguish spatial patterns of wet-versus-dry pillows and watersheds over the water year than the gridded estimates. Our results suggest treating gauge-based datasets as 'truth' may give a misleading representation of model accuracy, since these gauge-based datasets often have issues of their own.

1 Introduction

Quantitative precipitation estimates in mountainous areas are essential for hydrologic modeling and water management. Despite being critical, accurate precipitation estimates are notoriously difficult to produce in areas of complex terrain due to limitations of observing systems and large spatial variability in these regions. Ground-based radars can provide reliable estimates of precipitation over regions with homogeneous topography, but suffer from beam blocking and other issues in complex terrain (Nelson et al. 2016; Willie et al 2016). Large spatial variability, due to meteorological response to terrain features, makes it difficult to translate point measurements into gridded estimates, a problem confounded by the difficulty of attaining a dense network of point measurements in the complex terrain. Thus hydrologists often turn to gridded precipitation datasets, created either through statistical methods applied to in situ data (hereafter gridded estimates) or through dynamical downscaling of reanalysis datasets with atmospheric models.

Several daily statistically gridded precipitation estimates exist for the continental United States at resolutions as fine as 1 km (e.g., see Table 1 in Lundquist et al. 2015, hereafter L15). These datasets interpolate gauge data to a grid using numerical methods; the majority of these scale their daily values such that their long-term monthly means match the Precipitation-elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 1994) long-term monthly means. Because they rely directly on in situ data, these datasets provide an estimate of precipitation with uncertainties commensurate with the uncertainties in the in situ data themselves where in situ data are dense, and with greater uncertainty where in situ data are sparse, such as at high elevations.

With the advent of more computing power, ever-improving atmospheric models, and better-constrained atmospheric reanalyses, dynamical downscaling of reanalysis datasets to convection-permitting resolution has become a viable, if more computationally expensive, alternative to statistical downscaling for high-resolution precipitation estimation. Dynamical downscaling uses a state-of-the-art numerical weather model to estimate the atmospheric state at high resolution given prescribed large-scale conditions (e.g., see review by Xue et al. 2014). To generate high resolution historical data, the prescribed large-scale conditions are generally from reanalysis datasets. Because dynamical downscaling uses a numerical weather model to calculate precipitation based on discretized equations of state, it can represent physical processes not captured by the linear precipitation-elevation regressions used by PRISM.

The skill of dynamical downscaling to accurately represent reality is a function of the model physics and accuracy of the large scale conditions, and has the largest potential to improve over coarser scale reanalyses in areas of complex terrain and coastlines (Xue et al. 2014; Feser et al. 2010). Prior studies have shown that microphysical parameterizations have large impacts on precipitation type/phase in the cloud and on the ground (Minder and Kingsmill 2013 and Jankov et al. 2007, 2009, 2011), and that precipitation sensitivity to microphysical scheme is often larger than to other physics parameterizations (Liu et al. 2011). Although not as well studied as sensitivity to physics parameterizations, dynamical downscalings are also sensitive to their lateral boundary conditions (Yang et al. 2012), so uncertainties in reanalysis datasets create uncertainties in the downscaled data.

Evaluating high-resolution gridded precipitation datasets is challenging: Statistically gridded datasets often ingest most available in situ precipitation gauge data, making determining their skill versus that of dynamical downscalings tricky. In fact, many studies use PRISM-based gridded datasets to evaluate their dynamical downscalings (e.g. Caldwell et al. 2009), despite the uncertainties in the gridded datasets. These uncertainties can be large: For example, when compared with independent precipitation observations, PRISM-based gridded datasets have been shown to be off by up to a factor of two in mountainous terrain (Gutmann et al. 2012; Jeton et al 2006). These uncertainties also have important implications for water resources because they can significantly impact water year total precipitation amounts: Henn et al. (2016b) showed using different gridded datasets in complex terrain could cause up to 40% differences in water year totals.

Two data sources that have been largely untapped for precipitation dataset evaluation are stream gauges and snow pillows. Streamflow datasets are not used in any of the gridded datasets, to our knowledge, and thus serve as a completely independent dataset representing basin-mean precipitation; however, using them to validate precipitation estimates requires accounting for hydrologic processes that carry the precipitation to the streams. Henn et al. (2016a; hereafter H16) infer basin-mean precipitation from 56 streamflow gauges in the Sierra Nevada with Bayesian modeling, providing a streamflow-inferred precipitation dataset independent from gridded estimates of precipitation; this dataset includes its own range of uncertainty as part of the Bayesian modeling technique. Snow pillows offer the second widely unused in situ dataset for gridded precipitation validation: L15 quality controlled 20 years of snow pillow data across the Sierra Nevada from the California Department of Water Resources (CA DWR); some of these pillows and nearby snow courses have been used to adjust PRISM climatologies (L15), but the daily SWE amounts are not used directly in any of the gridded datasets.

L15 evaluated two gridded datasets, Hamlet (Hamlet et al. 2010; Hamlet and Lettenmaier 2005) and Livneh (Livneh et al. 2013), against the CA DWR snow pillows and found median errors for the entire period of ±10%. L15 further identified that during some years the two gridded datasets severely underpredicted precipitation during individual storms by as much as 50%, leading to water year total errors of ~20%; water year 2008 (WY2008) showed the most severe underestimation in their 20-year study period, with two major storms having substantial snow underestimation in each dataset. L15 showed these errors occurred because of an increase in orographic precipitation gradient when the winds were more westerly/northwesterly than typical during precipitating days, and hypothesized that this shift was associated with more time spent in the storm's cold sector.

In this manuscript, we extend upon the results of L15 for WY2008, by comparing the snow pillow observations to four gridded datasets -- Hamlet, Livneh, Daymet (Thornton et al. 1997; Thornton et al. 2014), and Newman (Newman et al. 2015). We assess whether the additional two gridded datasets, Daymet and Newman, also underpredict snow and total precipitation at high elevations in WY2008. We hypothesize that these two datasets will suffer similar biases in WY2008 as those in Hamlet and Livneh, since they use similar, terrain-based interpolation techniques, although we expect some variation across the gridded datasets because of differing methodological choices. In addition, we identify whether six dynamical downscalings suffer from the same deficiencies in WY2008, and hypothesize that the dynamical downscalings will not suffer from similar biases on the windward slope of the Sierra Nevada because of their ability to represent orographic precipitation processes. We also investigate the sensitivity of the dynamical downscalings to their microphysical parameterizations and lateral boundary conditions to test whether the large scale uncertainties in reanalyses produce differences in precipitation comparable to those from microphysics parameterizations. Testing over the course of a water year (rather than performing case studies of individual storms) allows us to test whether differences in precipitation due to microphysical and large scale uncertainties accumulate over the course of the water year, and to see whether they are systematic or random in different parts of the Sierra Nevada. Finally, we compare the 10 precipitation estimates (4 gridded and 6 dynamical downscalings) to inferred basin-mean precipitation from H16. The manuscript is laid out as follows: Section 2 describes the datasets and methods used in the manuscript; Section 3 explores the differences between the 10 precipitation estimates and the snow pillow and inferred precipitation amounts; and Section 4 provides a summary and discussion of the results.

2 Datasets and methods

2.1 WRF simulations

Six dynamical downscalings of WY2008 were generated using the Weather, Research, and Forecast (WRF) model, version 3.6 (Skamarock et al. 2008). All simulations were initialized in July 2007 and run continuously through Oct. 1, 2008, with the first three months discarded as model spin-up. All six simulations were identically configured aside from the lateral boundary conditions and microphysics schemes used (Table 1). Half the simulations used lateral boundary conditions (LBCs) from the ERA Interim reanalysis (ERA-I; Dee et al. 2011), and half used the North American Regional Reanalysis (NARR; Mesinger et al. 2006). Each LBC was paired with one of three microphysics schemes: WRF Single-Moment 6-Class Scheme (WSM6; Hong and Lim 2006), the Morrison et al. (2009) double-moment scheme (Morr), or the Thompson et al. (2008) scheme (Thom), resulting in six simulations, which are hereafter identified as E.Morr, E.Thom., E.WSM6, N.Morr, N.Thom, and N.WSM6, where E. and N. refer to ERA-I and NARR, respectively. Approximate computational time per month of simulation is shown in Table 1.

The simulations used an 18 km outer domain covering much of the intermountain west and stretching west across the northeastern Pacific Ocean, with a 6 km inner domain that extended across all of California (Fig. 1). Both domains used the Rapid Refresh Transfer Model for GCM applications (RRTMG; Iacono et al. 2008) for shortwave and longwave radiation and the Yonsei University planetary boundary layer scheme (Hong et al. 2006) with revised Mesoscale Model version 5 surface layer physics (Jimenez et al. 2012). The 18 km domain used the Kain-Fritsch convective parameterization (Kain 2004), while in the 6 km domain only resolved convection could occur. Spectral nudging was used in the 18 km domain to prevent simulation drift; nudging was applied with strength 0.0003 s^{-1} on winds and temperature above the 40^{th} model level. The simulations used 82 vertical levels.

A yearlong test simulation which used the above E.Morr configuration with the Noah land surface model (Tewari et al. 2004) revealed extremely cold surface temperature biases that developed in springtime (not shown). These biases were attributed to the representation of snow within the Noah land surface model (e.g., Barlage et al. 2015; Pavelsky et al. 2011), and thus the simulations used in this manuscript use the more sophisticated Noah-MP land surface model (Niu et al. 2011), which eliminates the springtime biases (not shown).

WRF outputs total precipitation, snow, ice, and graupel. Thus for the comparison with snow pillow data, the sum of snow, ice, and graupel is used as frozen precipitation, whereas for the comparison with the Bayesian estimated precipitation, total precipitation is used.

2.2 Statistically gridded precipitation estimates

Four datasets that interpolate precipitation and temperature from gauge observations and estimate them on a grid that extends across the Continental United States are used in this manuscript.

As discussed in more detail in Lundquist et al. (2015), two of the gridded precipitation/temperature datasets, Livneh (Livneh et al. 2013) and Hamlet (Hamlet et al. 2010; Hamlet and Lettenmaier 2005) (and many other gridded datasets not shown in this manuscript), use the Parameter–Elevation Regressions on Independent Slopes Model (PRISM; Daly et al. 1994, 2008) climatology to interpolate precipitation over topography. Both Livneh and Hamlet are available on a 1/16° grid. Both datasets used gauges from the National Climatic Data Center (NCDC) Cooperative Observer (COOP) network, although they differ slightly in their criteria for station inclusion. Hamlet uses PRISM to rescale temperature over topography, whereas Livneh uses a constant lapse rate of 6.5 °C km^{-1} for topographic temperature adjustment. Hamlet has no data available in the northeastern quadrant of our focus region after 2006 (see greyed region in Fig. 3e), but most of our comparisons focus west and south of this area.

The third gridded precipitation dataset used is Daymet (Thornton et al. 1997; Thornton et al. 2014), which combines a Gaussian weighting filter centered at the observation locations with linear regression to account for elevation changes to solve for both daily gridded precipitation and temperature minimum and maximum. Daymet is available on a 1km grid; prior to interpolation to the WRF grid described below we smooth Daymet with a 5 km-wide centered average. Daymet includes both COOP precipitation stations and stations in the U.S. Natural Resources Conservation Service (NRCS) Snowpack Telemetry (SNOTEL) network.

The fourth gridded precipitation dataset, Newman (Newman et al. 2015), uses similar gridding methodologies as the other three datasets, but differs in its inclusion of uncertainty estimates by generating an ensemble of estimates following the methods of Clark and Slater (2006). Newman uses distance dependent weightings from nearby stations with regression methods to generate the gridded precipitation estimates, where the regression residuals are used to generate uncertainty estimates. Topographic slope information was included in the regressions to account in a simple way for windward and leeward slope precipitation differences. The individual Newman precipitation and temperature ensemble members are available on a 1/8° grid. Newman includes more gauge data than the other datasets, including COOP stations and SNOTEL as well as gauges from the Community Collaborative Rain, Hail, and Snow (CoCoRaHS) network; and the various automated airport weather stations.

Since all four datasets are available on different grids, we interpolate them using nearest-neighbor interpolation to the 6 km WRF grid prior to our analysis. In addition, for the comparison to snow pillow data, daily frozen precipitation was calculated by summing precipitation only on days with minimum temperature (Tmin) less than 0°C (following L15). Because of its ensemble nature, the Newman dataset required additional processing: Calculation of frozen precipitation in Newman used Tmin (computed from the dataset-native temperature mean and range) for each ensemble member individually, constructing an ensemble of frozen precipitation. Throughout the manuscript, when only one value is shown for Newman precipitation or frozen precipitation, we are showing the ensemble median. We also show on a few figures the 25^{th} and 75^{th} percentiles of Newman precipitation or frozen precipitation, in addition to the median, to characterize the uncertainty captured by the ensemble.

2.3 Snow pillows

The CA DWR manages a network of 125 snow pillows, 103 in the Sierra Nevada (Fig. 1, data available from California Data Exchange Center 2014); 95 of these pillows report enough quality data in 2008 for comparison to our frozen precipitation datasets. These are generally located in flat clearings and measure the weight of snow accumulating over an area of about 7 m^2 to determine snow water equivalent (SWE). Because pillows can experience several hours delay in responding to changes in SWE (Beaumont 1965; Johnson and Marks 2004), they are not as reliable at sub-daily resolution, and thus data were analyzed at daily increments. All positive daily changes in measured snow water equivalent, ΔSWE, were taken to be a measure of daily snowfall. An increase in SWE was attributed to snow falling on the pillow, or to liquid water falling on snow already on the pillow and freezing into the snowpack, thereby increasing its density. In freezing locations where a snow pillow was co-located with a precipitation gauge, the timing and amount of ΔSWE closely tracked the total accumulated precipitation. Exceptions occurred where the precipitation gauge suffered severe undercatch (in those cases ΔSWE exceeds measured precipitation) or during warm rain events (when rainwater passes through the snowpack and drains away from the pillow, and measured precipitation exceeds ΔSWE). Snowmelt and/or sublimation also may decrease SWE. Wind redistribution of snow can either augment or decrease SWE, but this effect is slight because most California snow pillows are in sheltered locations (Farnes 1967). In summary, snow pillows are a reliable measure of high-elevation snowfall, and they do not suffer from the undercatch that standard precipitation gauges suffer in such environments (Yang et al. 2005). However, because Sierra snowpacks are typically warm and isothermal, most rain falling on a snow pillow is not retained and therefore, not measured (Lundquist et al. 2008). All snow pillow data were quality controlled as described in L15.

2.4 Bayesian precipitation estimates

In order to provide another independent estimate against which to validate the modeled precipitation, we use daily streamflow observations and a method for inferring basin-mean precipitation given streamflow. Streamflow observations provide an indirect representation of precipitation patterns, as each basin integrates spatially-distributed precipitation inputs into the streamflow response.

Of the 56 stream gauges identified by H16, which measure streamflow from basins that are largely free of upstream diversions and flow regulation, we use a subset of 31 with data in

WY2008. We then apply a Bayesian methodology (Henn et al. 2015) to infer the probability distribution of the basin-mean precipitation total for WY2008, given the observed streamflow in each basin. The methodology uses lumped hydrologic models forced by daily precipitation time series, which are scaled using multiplier parameters. These parameters, along with the other hydrologic model parameters, are then inferred in Bayesian model calibration to streamflow observations. Thus, the inferred precipitation from streamflow, $P_{inferred}$, is the WY2008 precipitation total that yields the best match to observed streamflow in each basin. $P_{inferred}$ is given as an ensemble resulting from using six different hydrologic model structures, in order to represent the uncertainty associated with this approach. While the uncertainty of $P_{inferred}$ is substantial, we note that streamflow represents a spatially-integrated response to precipitation, unlike precipitation gauge-based datasets that are derived from point measurements. In areas of high spatial variability of precipitation and sparse gauge networks, streamflow-derived $P_{inferred}$ may capture aspects of this variability missed by gauge-based datasets. For more information on the methodology used to infer precipitation from streamflow, see Henn et al. (2015, 2016a).

3 Differences in frozen and total precipitation across datasets

3.1 Snow pillow comparisons

3.1.1 Differences in annual frozen precipitation

In this section, we examine how frozen precipitation varies across the different datasets, and how each dataset's frozen precipitation compares with that of the snow pillows and the multi-product mean. We begin by comparing the gridded datasets to the multi-product mean (Figs. 2 and 3); the multi-product mean is the average of the six WRF and four gridded datasets (the Newman median frozen precipitation is used). WY2008 had ~13% lower-than-average snow totals compared with the 20-year average across the pillows of L15, and ~30% lower-than-average precipitation totals for the Sierra Nevada in the Sierra Nevada 8-station index (Ralph et al. 2016); a larger-than-average percentage of WY2008's precipitation fell during westerly wind situations (L15). There are stark differences in where the gridded datasets and WRF tend to put the largest frozen precipitation amounts: WRF places more precipitation on the windward slope of the Sierra Nevada, just east of the 1000 m terrain contour, with less precipitation than the multi-product mean east of the crest and in the foothills. This tendency is most pronounced in the WSM6 simulations, which have frozen precipitation amounts 400 to 500 mm larger on the western slope than the multi-product mean, and least pronounced in EThom, followed by NThom, and EMorr, respectively. All four gridded datasets have a nearly inverse pattern, with less precipitation along the windward slope (up to 300-400 mm less in Livneh and Hamlet) and more precipitation along the foothills and east of the crest (largest in Daymet and Newman).

Individual comparison of frozen precipitation at the nearest gridpoint to each snow pillow observation (Fig. 4 and Fig. 5) reveals general biases of the individual datasets. When scattered against the snow pillow water year totals (Fig. 4a), the reduced precipitation along the windward slope in the gridded datasets shows up as a general tendency for these datasets to fall below the 1:1 line, especially at pillow locations with greater than 800 mm of snow in WY2008. The bias of the WRF simulations depends strongly on the microphysics scheme: The two WRF simulations with WSM6 microphysics have a large wet bias across much of the region (10-20%, on average, and up to 50-60% at a few locations). The other four WRF simulations with the double-moment microphysics schemes generally fall between the gridded datasets and the WSM6 simulations, with frozen precipitation amounts that are frequently more in line with that observed at the snow pillows, although large biases still exist at some locations. These overall tendencies are reflected by summary statistics of the differences of the datasets' total WY2008 frozen precipitation with the snow pillow observations (Table 2). The WRF simulations with double-moment microphysics schemes overall have mean and median differences that are closer to zero than the other datasets, with the smallest median differences in EThom, NThom, and EMorr, respectively. The two WRF simulations with WSM6 microphysics have mean and median differences that are greater than 100 mm or ~13% of gauge-median total precipitation, reflecting their wet bias. The four gridded datasets all have mean and median differences that are negative, with Daymet and the Newman datasets having differences about half as large as the differences of Livneh and Hamlet. The standard deviations of the differences are rather large (greater than 200 mm, or ~26% of gauge-median total precipitation) for all datasets, indicating the large variation of comparisons with individual snow pillows.

Linear fits of each dataset with the snow pillow data (Fig. 4b) make clear a few additional details. First, the linear fits for all of the datasets tend to have a slope less than 1, indicating that they do not have a large enough difference between the 'dry' pillows and 'wet' pillows (i.e., those pillows with a small and large annual total frozen precipitation, respectively). This characteristic is generally worse in the gridded datasets – particularly Daymet – than in the WRF simulations. The two WSM6 WRF simulations, despite their large wet bias, seem to suffer the least of all datasets from this effect. Some of the inability to represent wet versus dry pillows could be due to spatial variability occurring on scales smaller than the 6 km grid, but the variance of this across datasets suggests some datasets are missing the general areas of heaviest and lightest precipitation. The linear fits also reveal a systematic difference between the WRF simulations with different lateral boundary conditions not easily visible in the scatterplot: The NARR-forced runs tend to be slightly wetter than their ERA-I-forced counterparts, and have a steeper (thus more in agreement with the snow pillows) slope, with more frozen precipitation at the stations with more observed snowfall, although this effect is clearly second-order when compared with the effects of microphysics.

To tease out the impact terrain forcing has on the distribution of frozen precipitation and errors in the downscaled and gridded datasets, Fig. 5 shows the snow pillow precipitation totals along with percent differences in each datasets' frozen precipitation plotted as a function of zonal terrain gradient. The zonal terrain gradient in longitude/latitude space is shown for reference in Fig. 5a, and has been calculated from the WRF terrain, smoothed by a 7 gridpoint filter to focus on larger-scale terrain features, and multiplied by -1 to facilitate the plotting: Eastward-directed gradients, i.e., those on the windward or west-facing slope, are thus negative values in Fig. 5a and appear on the left half of each panel Fig. 5b-l (in agreement with the gradients in longitude/latitude space), with westward-directed gradients on the right half of each panel. In the northern Sierra Nevada, the snow pillows generally have considerably more precipitation on the windward slope than in the lee, with annual frozen precipitation totals greater than 800 mm where the gradient is sloping up to the east and less than 800 mm where it is sloping up to the west. All 6 WRF downscalings tend to overdo this windward/leeside contrast, with slight (double-moment runs) to moderate (WSM6 runs) positive biases at the windward locations ranging in magnitude from near 0 to 60%, and dry leeside biases of up to 60% in all 6 simulations. The windward/leeside contrast in the southern Sierra is generally smaller in the snow pillow totals, and the WRF simulations similarly do not show as consistent of a pattern in their biases in this region. Unlike the WRF simulations, the biases in the gridded datasets do not seem to have any strong relationship with terrain gradient.

3.1.2 Differences in event and daily frozen precipitation

Our comparison thus far has focused on differences in total WY2008 frozen precipitation, but now we turn our attention to biases in daily precipitation, to get a sense for how these biases evolve over the water year. To do this, we examined cumulative traces of daily snow pillow and dataset snow pillow precipitation at each pillow (not shown). The behavior of the datasets' frozen precipitation with respect to the snow pillows varied substantially from pillow to pillow, with snow at some pillows being very over- or under-estimated by all datasets, well-represented in WRF but underestimated by gridded datasets, or well-represented by most datasets. These large differences at different pillows were not apparent when all pillows were lumped together for error statistics calculation. Thus, we subjectively divided the snow pillows into four groups, based on the general characteristics of the amount of snow pillow frozen precipitation with respect to that of the other datasets. Our four groups are:

Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets

Group B (25% of pillows): snow pillow > 8 or more datasets

Group C (28% of pillows): snow pillow near center of all datasets

Group D (25% of pillows): snow pillow < 8 or more datasets

A list of which snow pillows are in each group is provided in Table 3, and a map of their distribution is included on Fig. 1b. See H16a for a complete list and map of watersheds. Group A pillows largely run down the crest of the central Sierra, with a secondary cluster south of the Merced watershed. Group B is mostly in the lee of the northern Sierra and at a few of the lowest-elevation, windward side locations. Group C has a cluster of locations across the Tuolumne and Cherry-Eleanor watersheds, with additional locations scattered across the entire region. Group D is largely confined to the southern Sierra.

Example cumulative frozen precipitation traces from each of the four groups are shown in Fig. 6. These traces are representative of their individual groups, although the 'best' and 'worst' datasets at each pillow vary substantially. All four pillows shown, and 92% of all pillows (not shown), receive more than 50% of their total WY2008 precipitation during three, 3-11 day periods: These three events were identified in L15 as coincident 'missed storms' – i.e., underestimated snow amounts – in Hamlet and Livneh. That such a large fraction of WY2008 snow fell during 3 short periods is consistent with previous work showing that a substantial portion of annual California precipitation tends to fall during a few synoptic-scale events, often containing atmospheric rivers (L15; Dettinger et al. 2011): Dettinger et al. 2011 show that 50% of each year's precipitation accumulates over less than 15 days in the Northern Sierra and less that 10 days in the Southern Sierra, on average. Although each dataset has small errors during most of the precipitation events, the errors are larger for the largest three events, and in many cases large under/overestimation of precipitation in one event leads to the bulk of the WY disagreement. Fig. 6 also reveals that the timing of the precipitation is very similar in all datasets. This agreement is not surprising for the gridded datasets, which rely on gauge data. However, the WRF simulations, which are run in a regional climate framework (i.e., initialized in July 2007 and then integrated continuously from then through Oct. 1, 2008) could potentially diverge in their evolution of synoptic features after these features enter the edges of the outer domain. Since the timing agrees so well across the simulations and with observations, the domain configuration and nudging used on the outermost domain are providing strong constraints and keeping the simulations in agreement with the reanalysis lateral boundary conditions. Finally, Fig. 6 reveals information about the uncertainty estimate from the Newman dataset: The interquartile range (IQR) of the Newman precipitation at 3 of the 4 pillows shown exceeds the total range of the 10 frozen precipitation estimates and observed snowfall, and this is true for 86% of all pillows.

The question remains as to whether each datasets' biases are consistent for each storm leading to the total precipitation, or rather, if individual storm biases tend to vary in sign and cancel out over the WY. We address this question using histograms of errors for each group (Fig. 7) and boxplots of the errors for the three largest storms periods for each group (Fig. 8). For all datasets in all groups, the histogram peaks lie between -10 and 0 mm, indicating a general tendency for all datasets to underestimate small precipitation events. However, the majority of error in WY total frozen precipitation is driven by larger events, which impact the skewness and width of the histograms, and show up more clearly in boxplots from the three largest storms (Fig. 8).

Biases in group B in all 10 datasets seem to be fairly systematic, with the histograms of daily errors and the errors for the three largest storms lying mostly below zero, indicating that all the datasets have a tendency to underestimate frozen precipitation at these locations. In contrast, biases in group D are quite skewed, with a heavy positive tail (Fig. 7), and the overprediction at these pillows is mainly due to overprediction of frozen precipitation during the 3 big storm events of the WY (Fig. 8); the errors in this group are a bit worse in the WRF simulations than the gridded datasets, and are largest for the first storm period, especially in the NARR-forced simulations. The histograms and boxplots for group A and, to a lesser degree group C, illustrate the tendency for all the gridded datasets to underestimate precipitation in the central Sierra Nevada, and the daily error statistics suggest this is largely a systematic problem for these two groups of stations, with the majority of the histogram probability lying very close to or below zero for all four datasets in these groups. The difference between group A and group C in the gridded datasets is largely a result of Storm 1: in group A the boxes and most of the whiskers for all three storms are consistently below zero, and the net result is large underestimation of WY total frozen precipitation by the gridded datasets. In group C the gridded datasets' Storm 1 errors are more consistently positive, thus compensating for the general underestimation and resulting in smaller biases in WY total frozen precipitation. In both groups A and C, the biases in all 6 WRF simulations are more centered around zero than those of the gridded datasets, although in group C there are more large positive outliers. The positive WY total biases in the two WSM6 simulations appear as a slight shift in the histograms (Fig. 7) and storm total barplots (Fig. 8). The storm total barplots also reveal large differences in biases from storm to storm in the NARR-forced simulations, with Storm 1 showing very large positive errors in Groups A, C, and D for all microphysics parameterizations; the ERA-Interim-forced WRF biases are more consistent from storm to storm.

These statistics for the daily and storm-total errors suggest that error outliers contribute strongly to the overall biases for each ensemble member, and in some cases these biases can vary significantly from one storm to another. This lack of a more systematic bias possibly suggests that case studies of individual storm events of microphysics scheme performance may sometimes lead to incorrect conclusions about their overall tendencies, since the biases are reflected in the higher-order statistics rather than being systematic. California's tendency to receive most of each water year's precipitation in a few concentrated storm periods means that these errors in individual storms can strongly influence the water year biases. In addition, the different behavior of the daily errors in the different groups, which cluster in localized regions (Fig. 1), mean that bias tendencies can be highly variable spatially; thus conclusions about overall bias need to either take this spatial variability into account or be drawn for rather large areas.

3.2 Comparison to Bayesian estimated precipitation

We now turn our attention to a comparison of the datasets' WY2008 precipitation with $P_{inferred}$ (Section 2.4). Although $P_{inferred}$ is limited temporally to WY total precipitation and spatially to basin-mean precipitation amounts, it provides an independent verification when combined with the snow pillow dataset used in the previous section. We begin this comparison with maps of the differences between $P_{inferred}$ and basin-mean precipitation for each dataset (Fig. 9). Because $P_{inferred}$ includes an uncertainty estimate, we use this estimate in our comparison, and rather than showing absolute or percentage-wise difference maps, we categorically compare with $P_{inferred}$. The precipitation estimates for a large number of basins for all datasets fall within the IQR of $P_{inferred}$. However, for the differences falling outside this uncertainty range, we see some patterns quite similar to those we saw with the snow pillow comparisons. For instance, in the central Sierra, the Yosemite-area watersheds of the Tuolumne River and/or Cherry and Eleanor Creeks are generally drier than $P_{inferred}$ in the four gridded datasets. The pillows in this region are mostly Group C (i.e., snow pillow near center of all datasets); however, for most of the pillows in these regions, at least 2 of the gridded datasets greatly underestimated frozen precipitation (not shown), and Hamlet and Daymet both underestimate frozen precipitation at more pillows than Livneh and Newman, consistent with $P_{inferred}$. In addition, the two WSM6 WRF runs are wetter than $P_{inferred}$ in several basins across the region, similar to the snow pillow comparisons. The North Fork of the American River, in the northern Sierra, is consistently underestimated by the gridded datasets and also tends to be underestimated by the WRF simulations, although in WRF the underestimation is within the range of uncertainty for all but one ensemble member. Many of the WRF simulations and gridded datasets overestimate precipitation in several of the smaller watersheds throughout the region, in particular, the San Joaquin basins of Pitman, Bear, and Bishop Creeks, and the Mokelumne basins; the pattern of which watersheds are overestimated is more consistent across the WRF simulations than across the gridded datasets. Finally, in the southern Sierra, the Kern River watershed precipitation is overestimated in all WRF simulations but those using Thompson microphysics, as well as in Daymet, consistent with the differences against snow pillows seen earlier in Fig. 5.

We see similar correspondence to the snow pillow comparisons when the datasets' precipitation is scattered against $P_{inferred}$ (Fig. 10), although as we saw with the mapped differences, many of the datasets' estimates fall within the IQR of $P_{inferred}$. Displayed in this way it becomes clear that all datasets tend to overestimate basin-mean precipitation in watersheds with the smallest amounts of WY2008 precipitation inferred from streamflow. The WSM6 simulations systematically overestimate precipitation in many watersheds, whereas the Morrison and Thompson simulations more consistently fall within the IQR of $P_{inferred}$. The NARR-driven simulations are consistently wetter than the ERA Interim-driven simulations, although again, this effect is secondary compared to the impact of microphysics on total precipitation. The gridded datasets consistently underestimate precipitation in the wettest watersheds; this underestimation, when combined with the overestimation of drier watersheds, is reflected in a flatter slope of linear fits against $P_{inferred}$ (Fig. 10b) similar to that seen in the snow pillow comparisons. All the WRF simulations do a consistently better job in distinguishing wet and dry basins; curiously, despite their consistent wet bias and consistent with the snow pillow results, the slope of the WSM6 simulations best matches that seen in $P_{inferred}$.

4 Summary and Discussion

In this manuscript, we explore the uncertainties during WY2008 in the Sierra Nevada of California's high-elevation precipitation in 10 datasets: six WRF regional climate downscalings with differing lateral boundary conditions and microphysical parameterizations, and four gauge-based, interpolation-gridded precipitation datasets: Livneh, Hamlet, Daymet, and Newman. We first compare frozen precipitation from these 10 datasets with positive daily changes in snow water equivalent from a network of 95 snow pillows across the Sierra Nevada, then follow this with a comparison of total precipitation with a precipitation dataset inferred from stream gauges using a Bayesian inference method. Most of the manuscript focuses on comparisons of WY total precipitation, but we also compare daily error statistics with the snow pillow data.

During WY2008, the gridded datasets, especially Livneh and Hamlet, tend to underestimate frozen precipitation on the windward slope of the Sierra Nevada, particularly in the vicinity of Yosemite National Park (Fig. 1). The WRF simulations consistently place more precipitation on the windward slope than the gridded datasets, although the amount of precipitation along the windward slope depends strongly on microphysical parameterization: the WRF simulations with single-moment microphysics tend to overestimate precipitation along the windward slope, whereas those with double-moment microphysics tend to better agree with the snow pillows at a large proportion of the snow pillow locations. WRF simulations with NARR as lateral boundary conditions are slightly wetter than those with ERA Interim boundary conditions, but this effect is second order compared to microphysical differences.

All six WRF simulations somewhat overestimate the windward/leeside precipitation contrast in the northern Sierra Nevada. This problem is more pronounced in the single-moment simulations, which produce significantly more graupel than the double-moment schemes (not shown, e.g., see Jankov et al. 2009), suggesting it could be partially due to too-efficient fallout of hydrometeors on the windward side of the Sierra Nevada. However, since the issue appears in all six WRF simulations irrespective of microphysics, another factor is probably also contributing, and we speculate that this may be due to insufficient embedded convection in these simulations during post-cold-frontal periods: Since embedded convection during post-cold-frontal storms tends to result in an increase in leeside snow (Geerts et al. 2011), if these simulations have too little embedded convection (due perhaps to their horizontal grid spacing) it could lead to not enough precipitation being lofted into the lee. This hypothesis is also consistent with the location of this bias in the northern Sierra Nevada, since the trajectory of most wintertime extratropical cyclones hitting California means the northern Sierra Nevada spends more time in the cold sector than the southern Sierra Nevada. The position of the most prominent 'leeside' underestimation east of Lake Tahoe also raises the possibility that an additional issue is insufficient resolution of the very narrow Carson range to the east of Lake Tahoe along with potential lake effects.

Daily errors were investigated by sorting the pillows into four different groups based on how much snow pillow precipitation fell with respect to the other datasets; three of these four groups clustered in highly localized regions. Daily and storm-total error statistics for the gridded datasets were fairly consistently dry-biased, with outliers determining the overall site biases; WRF's biases were more centered around zero, and similarly, outliers contributed to the overall tendencies for the different schemes. All 10 datasets underestimated small precipitation events. WRF's zero-centered daily biases and varying storm-total biases suggest case studies of microphysical bias need to be interpreted carefully, since individual events may not sample the distribution adequately to capture the error distribution. In addition, geographical tendencies of biases varied widely with topographic characteristics. Thus conclusions about region-mean bias need to be drawn for rather large areas, although region-mean bias is likely not adequate to understand the dynamics leading to the biases.

Finally, the WRF simulations, in particular those with single-moment microphysics, better distinguish wet-versus-dry pillows and watersheds than the gridded estimates. Even though the WRF simulations with single-moment microphysics have a large wet bias, this wet bias was fairly systematic across the region. The double-moment WRF simulations were less biased overall, but the differences between wet and dry pillows/watersheds were not as large as observed in these simulations. The gridded datasets have the least contrast between wet and dry pillows and watersheds, with positive biases at dry pillows and watersheds and large negative biases at the wettest pillows and watersheds; this result was also seen in H16. The differences between the gridded datasets and WRF simulations in this respect are likely caused in part by limitations of the statistical gridding methodologies used by the gridded datasets: These datasets use linear regression techniques based on climatology to extrapolate gauge precipitation amounts to regions without measurements, and thus embed these climatological relationships in their estimates. When precipitation patterns differ from climatology, these gridded datasets would tend toward climatology, and that introduces biases; in general, any statistical interpolation technique will likely produce a smoother solution than the underlying data. All 10 datasets somewhat underestimated the wet and dry contrast, and this consistent underestimation may be related to small-scale terrain features unrepresented by the 6km grid (Minder et al. 2008). More work should be done to understand whether this wet-versus-dry bias is systematic across multiple water years. Further, until this issue is better understood, uncertainty in precipitation should be explicitly considered when conducting research that uses precipitation as an input (e.g., hydrological or ecological science). The reason for better wet-versus-dry ratios in the single-moment than double-moment WRF simulations is unclear, and will require an in-depth investigation of storm dynamics and cloud microphysical properties, which is beyond the scope of the present manuscript.

The present study found systematic differences in the error statistics of WRF simulations based on microphysical scheme sophistication, and also between these WRF dynamical downscalings and gridded estimates of precipitation, with double-moment microphysics WRF simulations outperforming both single-moment WRF simulations and the gridded datasets in most respects in the Sierra Nevada. We highlight three significant results of this manuscript: 1) Many model evaluations use PRISM-based gridded datasets as truth; had we taken this approach our conclusion would be that all the WRF datasets have a wet bias, which the comparisons to snow pillows and $P_{inferred}$ show is not the case. 2) Our investigation of both water year total and single storm precipitation biases revealed that the water year total biases were in some cases quite dependent on biases from one major water year storm: Case studies of model configuration performed for individual storm events could lead to incorrect conclusions about the model's overall tendencies, since precipitation biases are reflected in the higher-order statistics rather than being systematic. 3) Our focus both Sierra-wide and at smaller scales (e.g., watershed scale) reveals that very different biases can exist at highly localized scales. These three results provide guidance for future research, suggesting care be taken regarding spatial and temporal scales and with the "observations" used for model evaluation: Without this care, studies may reach incorrect conclusions about model performance and where to focus future model development. Our results for the Sierra Nevada should be largely transferrable to other mid-latitude mountainous regions that receive most of their precipitation from orographically enhanced synoptic scale events, with that caveat that the performance of the gauge-based datasets is likely sensitive to gauge density, and that California's tendency to receive most precipitation in a few large events per year increases the risk that individual case studies may not represent overall biases.

Our results are limited by the single water year of available WRF output. Further, we emphasize that the water year presented was particularly problematic for gridded datasets (L15); further work is needed to examine whether the patterns in biases that we show here are consistent during water years with different conditions, particularly years with more extreme precipitation, and to investigate the dynamic and thermodynamic causes for interannual changes in orographic precipitation gradient. Finally, although using WRF to downscale reanalysis data is shown in this manuscript to improve over gridded datasets during certain water years, it is a computationally expensive option, and may not always be feasible for all applications. Two possible and promising alternative approaches are hybrid techniques that combine statistical and dynamical downscaling approaches (e.g., Sun et al. 2015) or simpler and less computationally demanding dynamical models, such as the Intermediate Complexity Atmospheric Research Model (Gutmann et al. 2016), although more work is needed in the development and testing of such tools. These approaches could also potentially be used to improve gauge-based gridded datasets, incorporating dynamical information to improve upon the weaknesses of purely statistical and elevation-based gridding techniques.

Acknowledgments and Data

This work was supported by the National Science Foundation (NSF) grant EAR – 1344595. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research. The authors acknowledge the NOAA

Research and Development High Performance Computing Program for providing computing and storage resources that have contributed to the research results reported within this paper. URL: http://rdhpcs.noaa.gov. We would like to thank three anonymous reviewers as well as Rob Cifelli and Andrea Thorstensen for providing comments which improved the manuscript.

All data used in this study are publicly available. ERA Interim data were downloaded from ECMWF using tools available on their website (http://apps.ecmwf.int/datasets/data/interim-full-daily/levtype=sfc/). The North American Regional Reanalysis data were retrieved from the Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://rda.ucar.edu/datasets/ds608.0/, accessed 15 Aug 2014. Daymet (Thornton et al. 2014) was retrieved from the Distributed Active Archive Center of Oak Ridge National Laboratory. Hamlet data are housed by the University of Washington Climate Impacts Group (http://cses.washington.edu/cig/data/wus.shtml). Livneh data are available from an ftp site (www.hydro.washington.edu/Lettenmaier/Data/livneh/livneh.et.al.2013.page.html). Newman data are available on NCAR's Earth System Grid (https://www.earthsystemgrid.org/dataset/gridded_precip_and_temp.html). The California Department of Water Resources snow pillow data are available from the California Data Exchange Center (CDEC; http://cdec.water.ca.gov). The WRF simulations are available through personal communication with the corresponding author (mimi.hughes@noaa.gov).

References

Accessed 10 May 2014.: California Data Exchange Center—Snow.

Barlage, M., M. Tewari, F. Chen, G. Miguez-Macho, Z. L. Yang, and G. Y. Niu, 2015: The effect of groundwater interaction in North American regional climate simulations with WRF/Noah-MP. Climatic Change, 129, 485-498.

Beaumont, R., 1965: Mt. Hood pressure pillow snow gage. Journal of Applied Meteorology, 4, 626-631.

Caldwell, P., H. N. S. Chin, D. C. Bader, and G. Bala, 2009: Evaluation of a WRF dynamical downscaling simulation over California. Climatic Change, 95, 499-521.

Clark, M. P., and A. G. Slater, 2006: Probabilistic quantitative precipitation estimation in complex terrain. Journal of Hydrometeorology, 7, 3-22.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A STATISTICAL TOPOGRAPHIC MODEL FOR MAPPING CLIMATOLOGICAL PRECIPITATION OVER MOUNTAINOUS TERRAIN. Journal of Applied Meteorology, 33, 140-158.

Daly, C., and Coauthors, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. International Journal of Climatology, 28, 2031-2064.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society, 137, 553-597.

Dettinger, M. D., F. M. Ralph, T. Das, P. J. Neiman, and D. R. Cayan, 2011: Atmospheric Rivers, Floods and the Water Resources of California. Water, 3, 445-478.

Farnes, P., 1967: Criteria for determining mountain snow pillow sites. Proc. 35th Western Snow Conf, 59-62.

Feser, F., B. Rockel, H. von Storch, J. Winterfeldt, and M. Zahn, 2011: Regional Climate Models Add Value To Global Model Data A Review and Selected Examples. Bulletin of the American Meteorological Society, 92, 1181-1192.

Geerts, B., Q. Miao, and Y. Yang, 2011: Boundary Layer Turbulence and Orographic Precipitation Growth in Cold Clouds: Evidence from Profiling Airborne Radar Data. Journal of the Atmospheric Sciences, 68, 2344-2365.

Gutmann, E. D., and Coauthors, 2012: A Comparison of Statistical and Dynamical Downscaling of Winter Precipitation over Complex Terrain. Journal of Climate, 25, 262-281.

Gutmann, E. D., I. Barstad, M. P. Clark, J. R. Arnold, and R. M. Rasmussen (2016), The Intermediate Complexity Atmospheric Research Model, J. Hydrometeor, 17, 957–973, doi: 10.1175/JHM-D-15-0155.1.

Hamlet, A., and Coauthors, 2010: Final project report for the Columbia basin climate change scenarios project. Report, University of Washington, Seattle, WA. Available at: http://www. hydro. washington. edu/2860/report/.[Accessed November 20, 2015].

Hamlet, A. F., and D. P. Lettenmaier, 2005: Production of temporally consistent gridded precipitation and temperature fields for the continental United States. Journal of Hydrometeorology, 6, 330-336.

Henn, B., M. P. Clark, D. Kavetski, and J. D. Lundquist, 2015: Estimating mountain basin-mean precipitation from streamflow using Bayesian inference. Water Resources Research, 51, 8012-8033. doi:10.1002/2014WR016736

Henn, B., M. P. Clark, D. Kavetski, A. J. Newman, M. Hughes, B. McGurk, and J. D. Lundquist, 2016a: Spatiotemporal patterns of precipitation inferred from streamflow observations across the Sierra Nevada mountain range. Journal of Hydrology (in press). doi:dx.doi.org/10.1016/j.jhydrol.2016.08.009.

Henn, B., Newman, A. J., Livneh, B., Daly, C., & Lundquist, J. D., 2016b. An assessment of differences in gridded precipitation datasets in complex terrain. Journal of Hydrology (in press). doi:http://dx.doi.org/10.1016/j.jhydrol.2017.03.008.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). J. Korean Meteor. Soc, 42, 129-151.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. Monthly Weather Review, 134, 2318-2341.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. Journal of Geophysical Research-Atmospheres, 113.

Jankov, I., P. J. Schultz, C. J. Anderson, and S. E. Koch, 2007: The impact of different physical parameterizations and their interactions on cold season QPF in the American River basin. Journal of Hydrometeorology, 8, 1141-1151.

Jankov, I., J. W. Bao, P. J. Neiman, P. J. Schultz, H. L. Yuan, and A. B. White, 2009: Evaluation and Comparison of Microphysical Algorithms in ARW-WRF Model Simulations of Atmospheric River Events Affecting the California Coast. Journal of Hydrometeorology, 10, 847-870.

Jankov, I., and Coauthors, 2011: An Evaluation of Five ARW-WRF Microphysics Schemes Using Synthetic GOES Imagery for an Atmospheric River Event Affecting the California Coast. Journal of Hydrometeorology, 12, 618-633.

Jeton, A. E., S. A. Watkins, and J. Huntington, 2006: Evaluation of Precipitation Estimates from PRISM for the 1961-90 and 1971-2000 Data Sets, Nevada. US Geological Survey.

Jimenez, P. A., J. Dudhia, J. F. Gonzalez-Rouco, J. Navarro, J. P. Montavez, and E. Garcia-Bustamante, 2012: A Revised Scheme for the WRF Surface Layer Formulation. Monthly Weather Review, 140, 898-918.

Johnson, J. B., and D. Marks, 2004: The detection and correction of snow water equivalent pressure sensor errors. Hydrological Processes, 18, 3513-3525.

Kain, J. S., 2004: The Kain-Fritsch convective parameterization: An update. Journal of Applied Meteorology, 43, 170-181.

Liu, C. H., K. Ikeda, G. Thompson, R. Rasmussen, and J. Dudhia, 2011: High-Resolution Simulations of Wintertime Precipitation in the Colorado Headwaters Region: Sensitivity to Physics Parameterizations. Monthly Weather Review, 139, 3533-3553.

Livneh, B., and Coauthors, 2013: A Long-Term Hydrologically Based Dataset of Land Surface Fluxes and States for the Conterminous United States: Update and Extensions. Journal of Climate, 26, 9384-9392.

Lundquist, J. D., P. J. Neiman, B. Martner, A. B. White, D. J. Gottas, and F. M. Ralph, 2008: Rain versus snow in the Sierra Nevada, California: Comparing Doppler profiling radar and surface observations of melting level. Journal of Hydrometeorology, 9, 194-211.

Lundquist, J. D., M. Hughes, B. Henn, E. D. Gutmann, B. Livneh, J. Dozier, and P. Neiman, 2015: High-Elevation Precipitation Patterns: Using Snow Measurements to Assess Daily Gridded Datasets across the Sierra Nevada, California. Journal of Hydrometeorology, 16, 1773-1792.

Mesinger, F., and Coauthors, 2006: North American regional reanalysis. Bulletin of the American Meteorological Society, 87, 343-+.

Minder, J. R., and D. E. Kingsmill, 2013: Mesoscale Variations of the Atmospheric Snow Line over the Northern Sierra Nevada: Multiyear Statistics, Case Study, and Mechanisms. Journal of the Atmospheric Sciences, 70, 916-938.

Minder, J. R., D. R. Durran, G. H. Roe, and A. M. Anders, 2008: The climatology of small-scale orographic precipitation over the Olympic Mountains: Patterns and processes. Quarterly Journal of the Royal Meteorological Society, 134, 817-839.

Morrison, H., G. Thompson, and V. Tatarskii, 2009: Impact of Cloud Microphysics on the Development of Trailing Stratiform Precipitation in a Simulated Squall Line: Comparison of One- and Two-Moment Schemes. Monthly Weather Review, 137, 991-1007.

Nelson, B. R., O. P. Prat, D. J. Seo, and E. Habib, 2016: Assessment and Implications of NCEP Stage IV Quantitative Precipitation Estimates for Product Intercomparisons. Weather and Forecasting, 31, 371-394.

Newman, A. J., and Coauthors, 2015: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States. Journal of Hydrometeorology, 16, 2481-2500.

Niu, G. Y., and Coauthors, 2011: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. Journal of Geophysical Research-Atmospheres, 116.

Pavelsky, T. M., S. Kapnick, and A. Hall, 2011: Accumulation and melt dynamics of snowpack from a multiresolution regional climate model in the central Sierra Nevada, California. Journal of Geophysical Research-Atmospheres, 116.

Skamarock, W., J.B. Klemp, J. Dudhia (more), 2008: A Description of the Advanced Research WRF Version 3. . NCAR Technical Note NCAR/TN-475+STR.

Sun F, D Walton, and A Hall, 2015: A hybrid dynamical-statistical downscaling technique, part II: End-of-century warming projections predict a new climate state in the Los Angeles region. Journal of Climate, 28(12): 4618-4636. DOI: 10.1175/JCLI-D-14-00197.1.

Tewari, M., and Coauthors, 2004: Implementation and verification of the unified NOAH land surface model in the WRF model. 20th conference on weather analysis and forecasting/16th conference on numerical weather prediction, pp. 11–15.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit Forecasts of Winter Precipitation Using an Improved Bulk Microphysics Scheme. Part II: Implementation of a New Snow Parameterization. Monthly Weather Review, 136, 5095-5115.

Thornton, P. E., S. W. Running, and M. A. White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology, 190, 214-251.

Thornton, P.E., M.M. Thornton, B.W. Mayer, N. Wilhelmi, Y. Wei, R. Devarakonda, and R.B. Cook. 2014. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. ORNL DAAC, Oak Ridge, Tennessee, USA. Accessed February 01, 2016. Time period: 2007-10-01 to 2008-09-30. http://dx.doi.org/10.3334/ORNLDAAC/1219

Willie, D., and Coauthors, 2016: Evaluation of Multisensor Quantitative Precipitation Estimation in Russian River Basin. Journal of Hydrologic Engineering, E5016002.

Xue, Y. K., Z. Janjic, J. Dudhia, R. Vasic, and F. De Sales, 2014: A review on regional dynamical downscaling in intraseasonal to seasonal simulation/prediction and major factors that affect downscaling ability. Atmospheric Research, 147, 68-85.

Yang, D. Q., D. Kane, Z. P. Zhang, D. Legates, and B. Goodison, 2005: Bias corrections of long-term (1973-2004) daily precipitation data over the northern regions. Geophysical Research Letters, 32.

Yang, H. W., and B. Wang, 2012: Reduction of systematic biases in regional climate downscaling through ensemble forcing. Climate Dynamics, 38, 655-665.Barlage, Michael, et al. "The effect of groundwater interaction in North American regional climate simulations with WRF/Noah-MP." *Climatic Change* 129.3-4 (2015): 485-498.

Figure 1. (a) WRF terrain (m) and extent of 18 km (edge of color fill) and 6 km (red outline) domains. Green line shows focus area of manuscript. (b) 2-min USGS terrain (black contours at 0, 1000, and 3000m), watershed extent for stream gauges (purple outlines and green color fill), and locations of snow pillows (colored markers) for green-outlined region of (a). Color and shape of snow pillow markers indicate their group in section 3: Group A, blue circles; Group B, green triangles; Group C, red diamonds; Group D, orange squares. Letters A, B, C, and D indicate locations of example pillows. Numbers 1-5 indicate watersheds highlighted in the text, with names given in upper right.

Figure 2. WY2008 total (mm): a) snow pillow observed snowfall, b) multi-product mean frozen precipitation, c) difference of multi-product mean frozen precipitation and snow pillow observed snowfall, d) gridded dataset mean frozen precipitation, e) WRF mean frozen precipitation, f) gridded dataset difference from multi-product mean, and g) WRF difference from multi-product mean. WRF amounts (e, g) are the sum of snow and graupel; gridded datasets (d, f) are the sum of precipitation on all days when Tmin<0 ℃ 2-minute terrain is plotted every 1000m starting at 0m.

Figure 3: Precipitation difference of individual dataset WY2008 total frozen precipitation from multi-product mean (Fig. 2a; mm). WRF amounts (a, b, c, f, g, h) are the sum of snow and graupel; gridded datasets (d, e, I, j) are the sum of precipitation on all days when Tmin<0 ℃. 2-minute terrain is plotted at 0, 1000, and 3000 m.

Figure 4. (a) Scatterplot of frozen precipitation versus snow pillow data at nearest gridpoint (mm). b) Linear regressions for scatterplots of (a).

Figure 5. a) Meridional gradient of smoothed terrain (color fill, m 6km-1), terrain from WRF simulation (black contours, every 1000m), and locations of snow pillows (black dots). b) Snow pillow water year total snow (mm) versus smoothed zonal terrain gradient (x-axis, as in (a)) and latitude (y-axis). (c-l) As in (b), but percent difference between frozen precipitation and snow pillow snow (%).

Figure 6. Cumulative traces of daily snow pillow snow and frozen precipitation for examples from each of the 4 groups of snow pillows (A-D) outlined in the text. Purple and cyan arrows show start dates of ‘missed storms’ from Lundquist et al. (2015) in Livneh and Hamlet datasets, respectively. Locations of example pillows are shown with letters in Fig. 1b. Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets; group B (25% of pillows): snow pillow > 8 or more datasets; group C (28% of pillows): snow pillow near center of all datasets; group D (25% of pillows): snow pillow < 8 or more datasets.

Figure 7. Histogram of errors (gridded-snow pillow) of smoothed daily ‘frozen’ precipitation for each of the 4 groups of snow pillows (A-D) outlined in the text, on days with smoothed observed snow > 5 mm. Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets; group B (25% of pillows): snow pillow > 8 or more datasets; group C (28% of pillows): snow pillow near center of all datasets; group D (25% of pillows): snow pillow < 8 or more datasets.

Figure 8. Boxplots of errors (gridded-snow pillow) of 'storm' total 'frozen' precipitation (mm) for each of the 4 groups of snow pillows (A-D) outlined in the text, for three major storm periods highlighted with arrows in Fig. 5: Storm1: Jan 3-8, 2008; Storm 2: Jan. 26 – Feb. 5, 2008; and Storm 3: Feb. 19-26, 2008. Outliers shown with red + signs are +/- 2 standard deviations. Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets; group B (25% of pillows): snow pillow > 8 or more datasets; group C (28% of pillows): snow pillow near center of all datasets; group D (25% of pillows): snow pillow < 8 or more datasets.

Figure 9. (a) Median basin-mean WY2008 Pinferred (mm). (b-k) Categorical difference of gridded dataset basin-mean precipitation (P) and Pinferred. Categories are: 1:P < min Pinferred, 2: min Pinferred < P < 25th % Pinferred, 3: 25th % Pinferred < P < 50th % Pinferred, 4: 50th % Pinferred < P < 75th % Pinferred, 5: 75th % Pinferred < P < max Pinferred, 6: max Pinferred < P. Note that categories 3 and 4 are within the interquartile range of uncertainty.

Figure 10. (a) Basin-mean precipitation (mm) from gridded datasets (see legend), as a function of Pinferred (black crosses). Large black crosses show median and gray shading bounded by smaller black crosses show interquartile range (IQR) of Pinferred. (b) Linear fit for each dataset. Black solid line shows Bayesian median and gray shading bounded by black dashed lines show IQR.

Table 1: Details of the WRF simulations. References for microphysics schemes and lateral boundary conditions can be found in Section 2.1.

Simulation name	**Summary description**	**Resolution of lateral boundary conditions (LBCs)**	**Compute hours per 30 days of simulation (run on 120 CPUs)**
E.Morr	ERA Interim LBCs and Morrison microphysics	~80 km	~4000
E.Thom	ERA Interim LBCs and Thompson microphysics	~80 km	~4000
E.WSM6	ERA Interim LBCs and WSM6 microphysics	~80 km	~3700
N.Morr	NARR LBCs and Morrison microphysics	~38 km	~4000
N.Thom	NARR LBCs and Thompson microphysics	~38 km	~4000
N.WSM6	NARR LBCs and WSM6 microphysics	~38 km	~3700

Table 2: Mean, median, and standard deviation of differences between total WY2008 frozen precipitation for each dataset and the snow pillow observations.

Dataset	Mean Difference (mm)	Median Difference (mm)	Standard Deviation of Difference (mm)
E.Morr	-2.4	36.5	222.0
E.Thom	-62.8	-16.6	210.4
E.WSM6	103.0	132.3	255.5
N.Morr	29.5	68.6	236.1
N.Thom	-22.9	19.0	221.0
N.WSM6	146.1	174.2	274.3
Livneh	-142.8	-111.1	259.2
Hamlet	-164.3	-141.1	222.1
Daymet	-60.7	-47.7	227.7
Newman	-81.1	-59.0	215.6

Table 3. Snow pillows in each of the 4 groups of Figs. 5 and 6. Group A: snow pillow and WRF > at least 2 gridded datasets; Group B: snow pillow > 8 or more datasets; Group C: snow pillow near center of all datasets; Group D: snow pillow < 8 or more datasets

	Name	**Elev.**	**Latitude**	**Longitude**
Group A pillows	Agnew Pass	2880	37.728	-119.143
	Bloods Creek	2195	38.45	-120.033
	Caples Lake (DWR)	2438	38.71	-120.042
	Gianelli Meadow	2560	38.205	-119.892
	Green Mountain	2408	37.555	-119.238
	Hagans Meadow	2438	38.853	-119.94
	Chilkoot Meadow	2179	37.41	-119.49
	Dana Meadows	2987	37.897	-119.257
	Ebbetts Pass	2652	38.561	-119.808
	Highland Meadow	2652	38.49	-119.805
	Mud Lake	2408	38.615	-120.14
	Poison Flat	2408	38.501	-119.631
	Poison Ridge	2103	37.403	-119.52
	Schneiders	2667	38.747	-120.068
	Stanislaus Meadow	2362	38.5	-119.937
	Squaw Valley Gold Coast	2499	39.194	-120.276
	Van Vleck	2042	38.945	-120.305
	Ward Creek 3	2057	39.137	-120.22
	Echo Peak 5	2377	38.849	-120.079
	Graveyard Meadow	2103	37.465	-119.29
Group B pillows	Alpha (SMUD)	2316	38.805	-120.215
	Bucks Lake	1753	39.85	-121.242
	Blue Canyon	1609	39.276	-120.708
	Casa Vieja Meadows	2530	36.2	-118.268
	Cottonwood Lakes	3094	36.483	-118.177
	Dismal Swamp	2149	41.993	-120.165
	Four Trees	1570	39.813	-121.321
	Forni Ridge	2316	38.805	-120.213
	Gem Pass	3277	37.78	-119.17
	Heavenly Valley	2682	38.929	-119.917
	Independence Lake (SCS)	2576	39.435	-120.322
	Kettle Rock	2225	40.14	-120.715
	Lobdell Lake	2804	38.44	-119.377
	Monitor Pass	2545	38.67	-119.615
	Marlette Lake	2438	39.173	-119.905
	Mount Rose Ski Area	2713	39.326	-119.902
	Pascoes	2789	35.967	-118.35
	Quaking Aspen	2195	36.117	-118.54
	Rattlesnake	1859	40.125	-121.043
	Slate Creek	1737	41.045	-122.478
	Snow Mountain	1814	40.778	-121.782
	Upper Burnt Corral	2957	37.183	-118.937
	Big Meadows (SCS)	2652	39.458	-119.946
	Grizzly Ridge	2103	39.917	-120.645
	Name	**Elev.**	**Latitude**	**Longitude**

Group C pillows	Adin Mountain	1890	41.237	-120.792
	Blackcap Basin	3139	37.067	-118.77
	Chagoopa Plateau	3139	36.497	-118.442
	Central Sierra Snow Laboratory	2103	39.325	-120.367
	Deadman Creek	2819	38.332	-119.653
	Gin Flat	2149	37.767	-119.773
	Huntington Lake (USBR)	2134	37.228	-119.221
	Horse Meadow	2560	38.158	-119.662
	Independence Creek	1981	39.494	-120.293
	Kaiser Point	2804	37.3	-119.1
	Mammoth Pass (USBR)	2835	37.61	-119.033
	Pilot Peak (Dwr)	2073	39.786	-120.875
	Robbs Saddle	1798	38.912	-120.378
	Lower Relief Valley	2469	38.243	-119.758
	Slide Canyon	2804	38.092	-119.43
	Sonora Pass Bridge	2667	38.318	-119.601
	Tunnel Guard Station	2713	36.367	-118.288
	Volcanic Knob	3063	37.388	-118.903
	Cedar Pass	2164	41.583	-120.303
	Blue Lakes	2438	38.613	-119.931
	Black Springs	1981	38.375	-120.192
	Humbug	1981	40.115	-121.368
	Huysink	2012	39.282	-120.527
	Lower Kibbie Ridge	2042	38.032	-119.877
	Medicine Lake	2042	41.592	-121.61
	Paradise Meadow	2332	38.047	-119.67
	Ostrander Lake	2499	37.637	-119.55
Group D pillows	Beach Meadows	2332	36.127	-118.293
	Crabtree Meadow	3261	36.563	-118.345
	Giant Forest (USACE)	2027	36.562	-118.765
	Independence Camp	2134	39.454	-120.299
	Leavitt Meadows	2195	38.305	-119.552
	Robbs Powerhouse	1570	38.903	-120.375
	Rock Creek Lakes	3048	37.455	-118.743
	Rubicon Peak 2	2286	39.001	-120.14
	South Lake	2926	37.176	-118.562
	Sawmill	3109	37.162	-118.562
	Tahoe City Cross	2057	39.172	-120.154
	Truckee 2	1951	39.3	-120.194
	Virginia Lakes Ridge	2835	38.077	-119.234
	West Woodchuck Meadow	2774	37.03	-118.918
	Big Pine Creek	2987	37.128	-118.475
	Bishop Pass	3414	37.1	-118.557
	Mitchell Meadow	3018	36.737	-118.712
	Silver Lake	2164	38.678	-120.118
	Spratt Creek	1875	38.667	-119.818
	Tamarack Summit	2301	37.165	-119.2
	Tuolumne Meadows	2621	37.873	-119.35
	Upper Tyndall Creek	3475	36.65	-118.397
	Wet Meadows	2728	36.348	-118.572
	Gold Lake	2057	39.675	-120.615

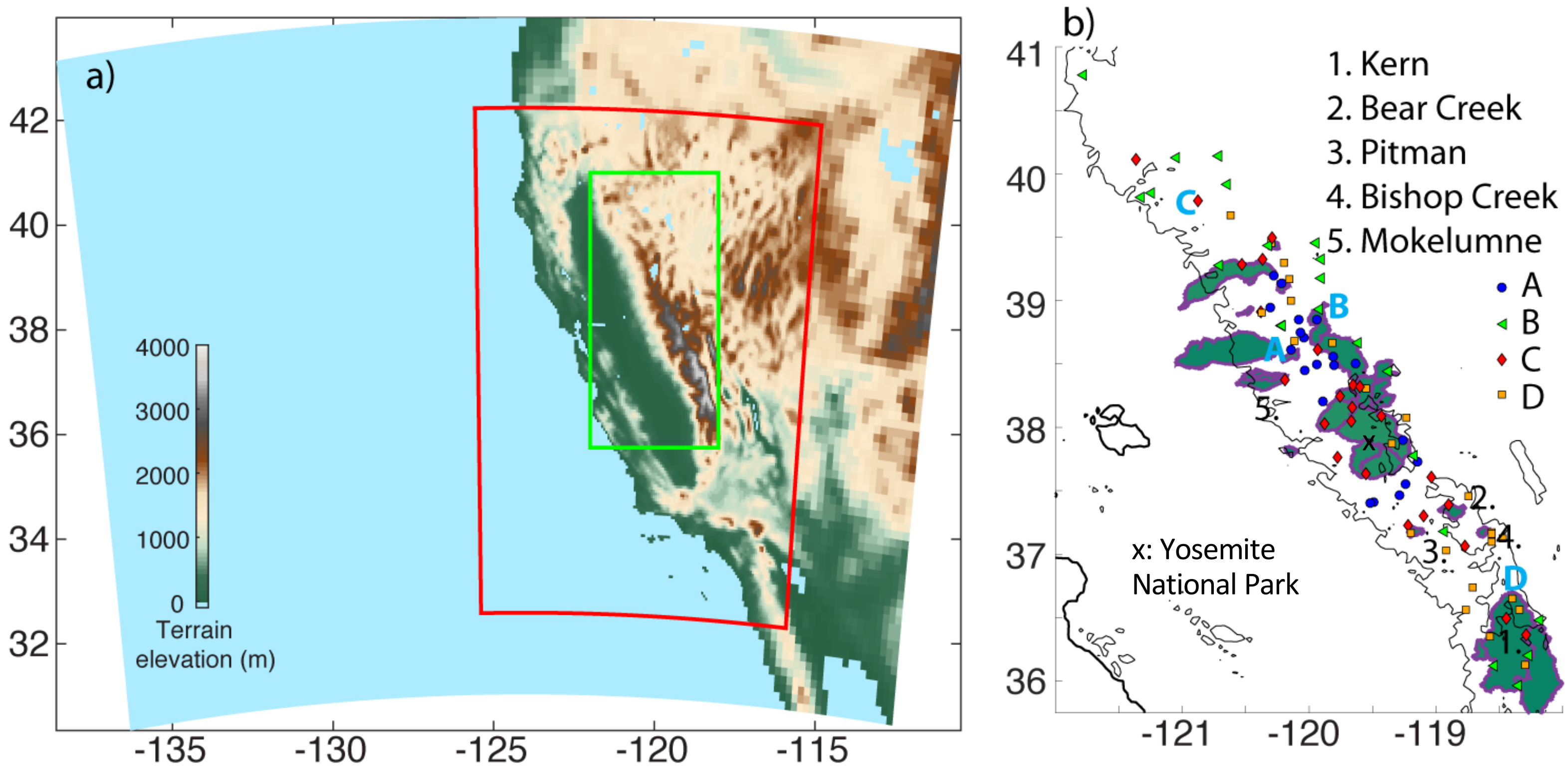

Fig. 1: (a) WRF terrain (m) and extent of 18 km (edge of color fill) and 6 km (red outline) domains. Green line shows focus area of manuscript. (b) 2-min USGS terrain (black contours at 0, 1000, and 3000m), watershed extent for stream gauges (purple outlines and green color fill), and locations of snow pillows (colored markers) for green-outlined region of (a). Color and shape of snow pillow markers indicate their group in section 3: Group A, blue circles; Group B, green triangles; Group C, red diamonds; Group D, orange squares. Letters A, B, C, and D indicate locations of example pillows. Numbers 1-5 indicate watersheds highlighted in the text, with names given in upper right.

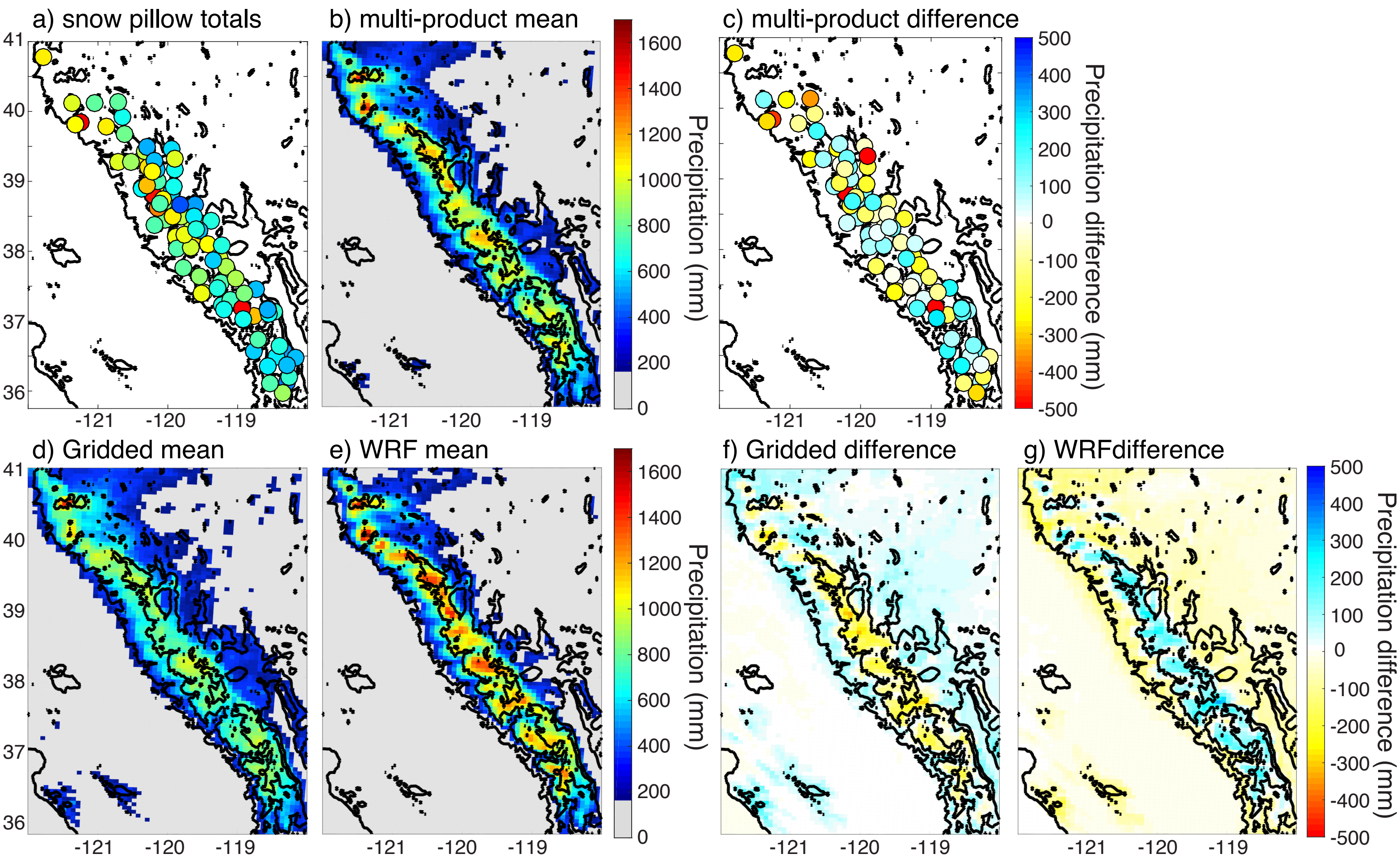

Fig. 2: WY2008 total (mm): a) snow pillow observed snowfall, b) multi-product mean frozen precipitation, c) difference of multi-product mean frozen precipitation and snow pillow observed snowfall, d) gridded dataset mean frozen precipitation, e) WRF mean frozen precipitation, f) gridded dataset difference from multi-product mean, and g) WRF difference from multi-product mean. WRF amounts (e, g) are the sum of snow and graupel; gridded datasets (d, f) are the sum of precipitation on all days when Tmin<0 °C 2-minute terrain is plotted every 1000m starting at 0m.

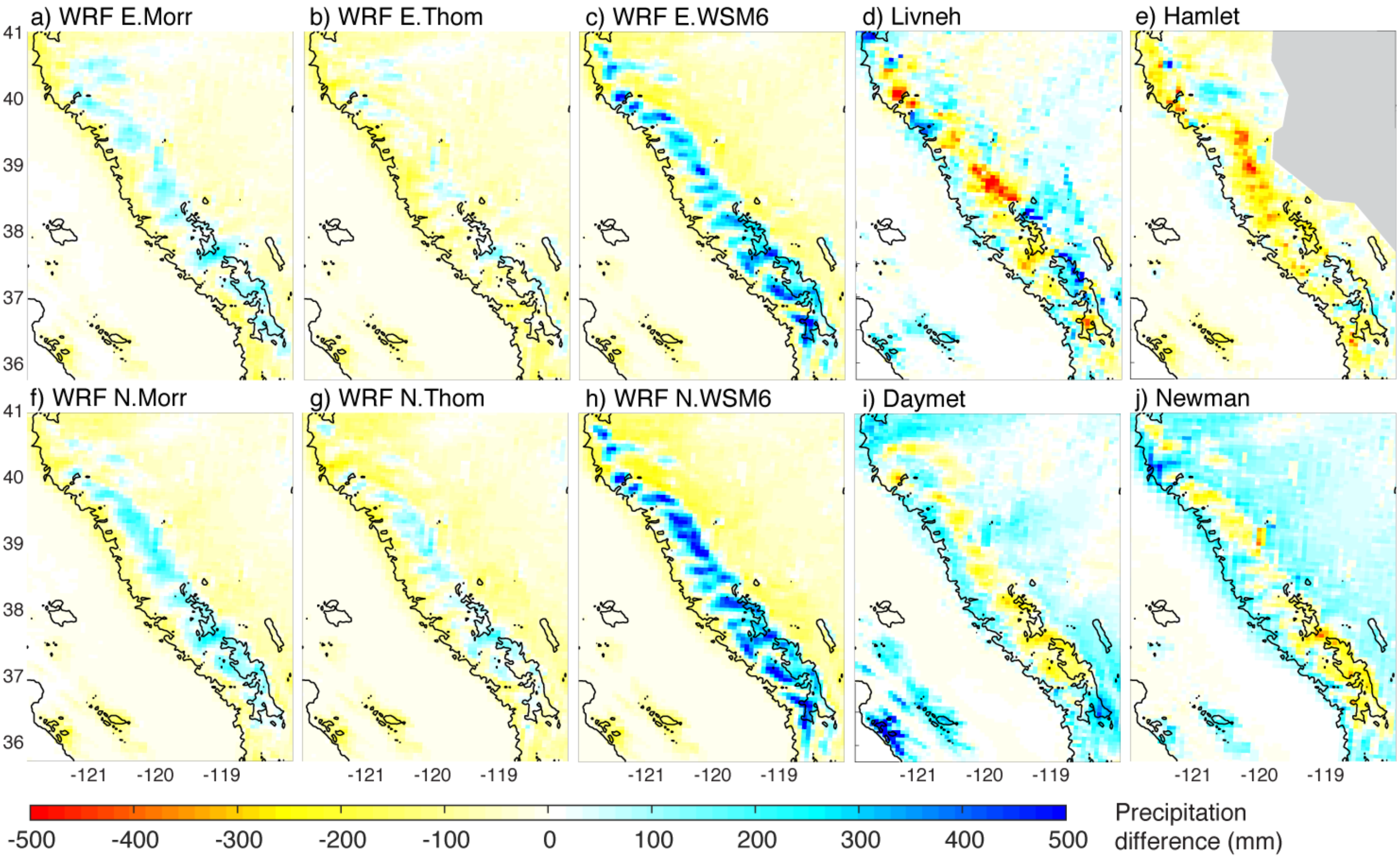

Fig. 3: Precipitation difference of individual dataset WY2008 total frozen precipitation from multi-product mean (Fig. 2a; mm). WRF amounts (a, b, c, f, g, h) are the sum of snow and graupel; gridded datasets (d, e, I, j) are the sum of precipitation on all days when Tmin<0 °C. 2-minute terrain is plotted at 0, 1000, and 3000 m.

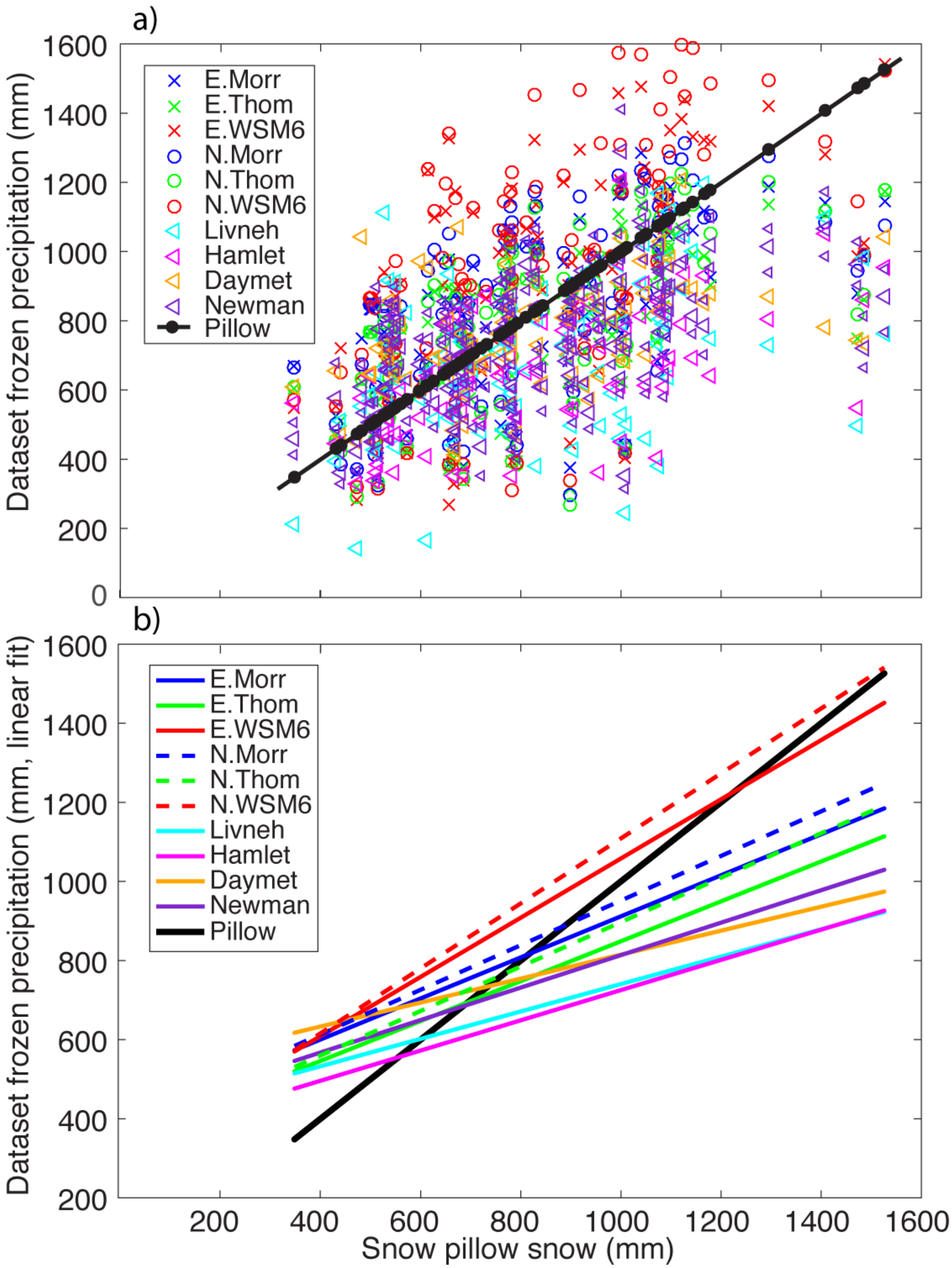

Fig. 4: a) Scatterplot of frozen precipitation versus snow pillow data at nearest gridpoint (mm). b) Linear regressions for scatterplots of (a).

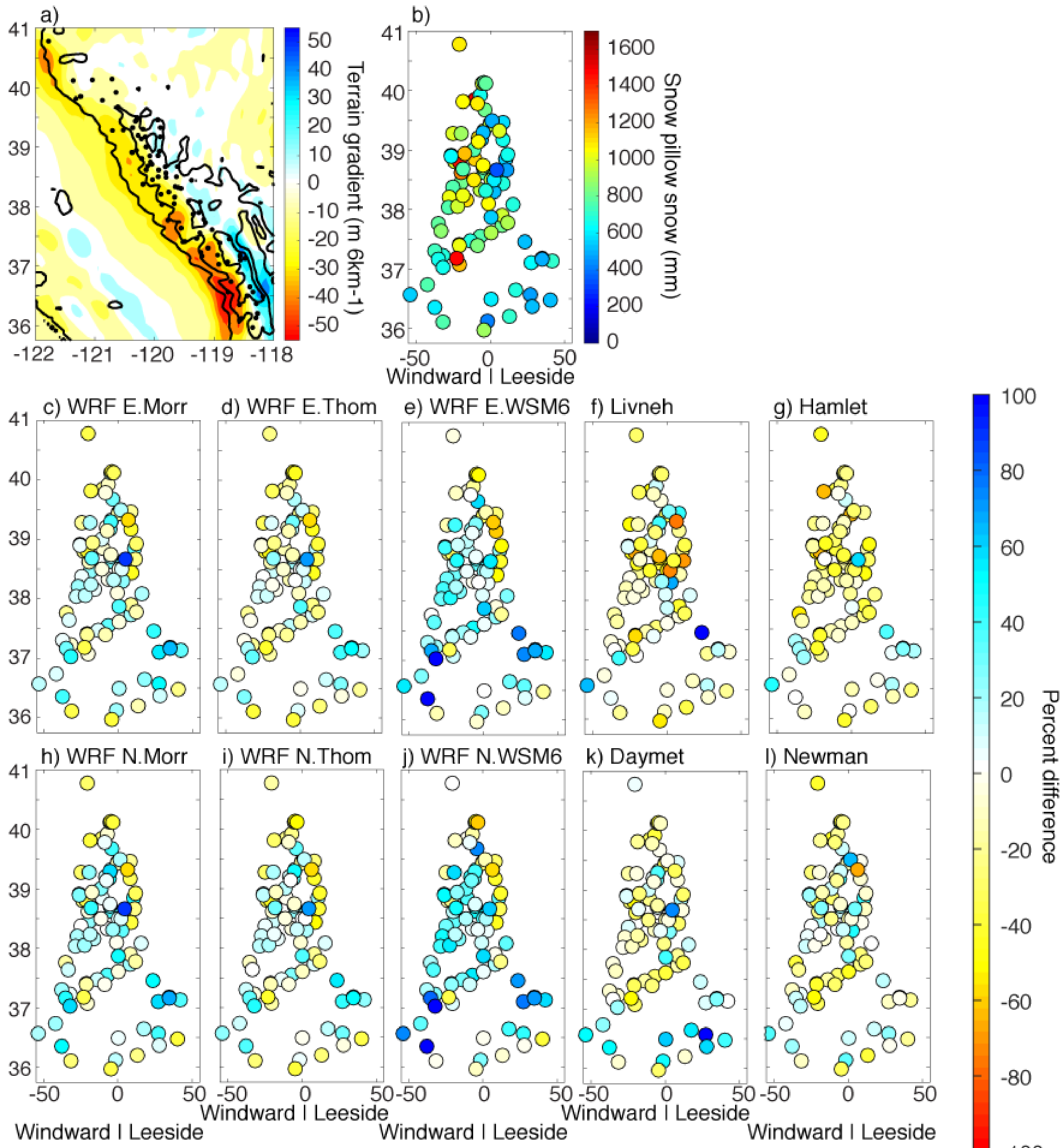

Fig. 5: a) Meridional gradient of smoothed terrain (color fill, m 6km^{-1}), terrain from WRF simulation (black contours, every 1000m), and locations of snow pillows (black dots). b) Snow pillow water year total snow (mm) versus smoothed zonal terrain gradient (x-axis, as in (a)) and latitude (y-axis). (c-l) As in (b), but percent difference between frozen precipitation and snow pillow snow (%).

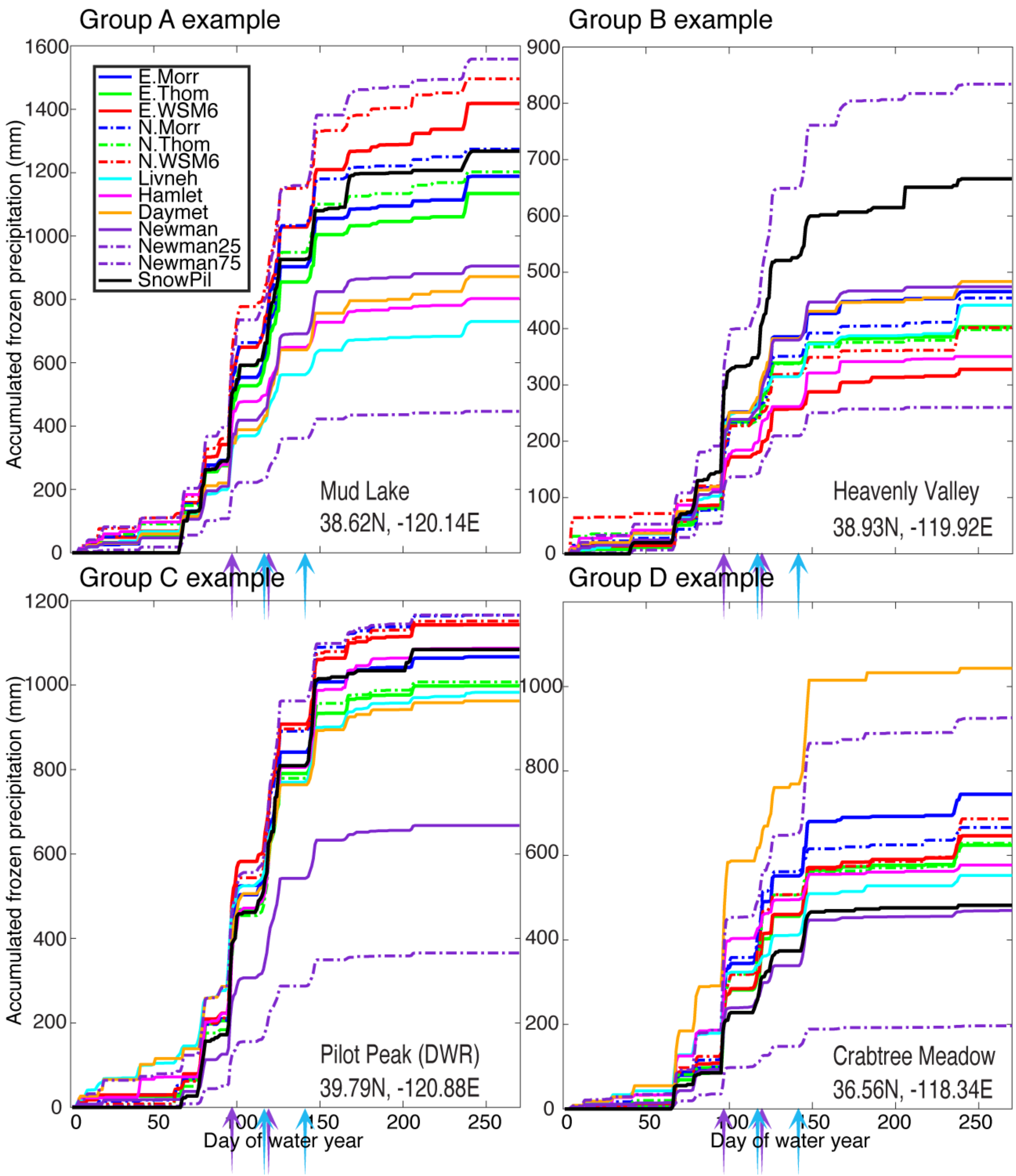

Fig. 6: Cumulative traces of daily snow pillow snow and frozen precipitation for examples from each of the 4 groups of snow pillows (A-D) outlined in the text. Purple and cyan arrows show start dates of 'missed storms' from Lundquist et al. (2015) in Livneh and Hamlet datasets, respectively. Locations of example pillows are shown with letters in Fig. 1b. Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets; group B (25% of pillows): snow pillow > 8 or more datasets; group C (28% of pillows): snow pillow near center of all datasets; group D (25% of pillows): snow pillow < 8 or more datasets.

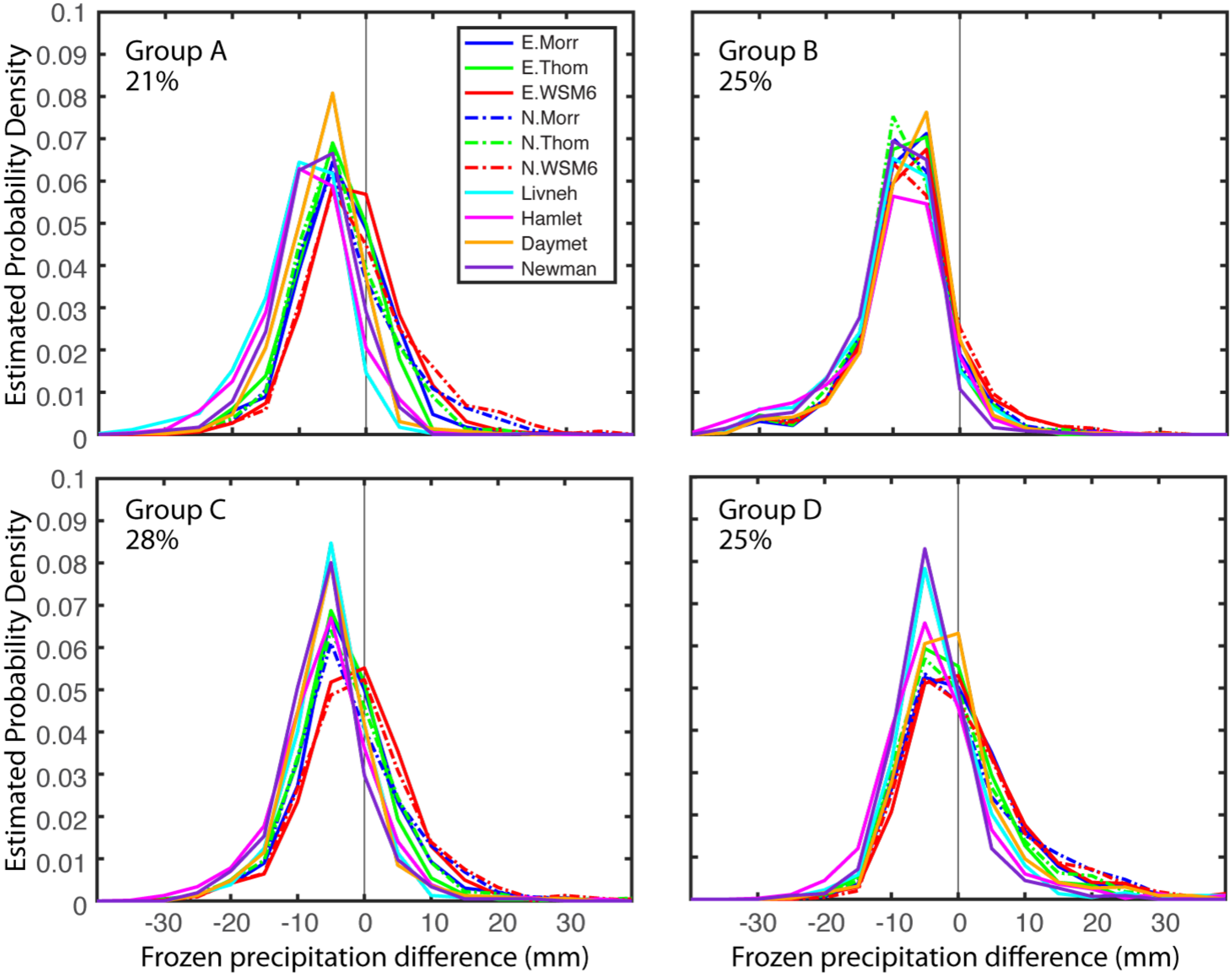

Fig. 7: Histogram of errors (gridded-snow pillow) of smoothed daily ‘frozen’ precipitation for each of the 4 groups of snow pillows (A-D) outlined in the text, on days with smoothed observed snow > 5 mm. Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets; group B (25% of pillows): snow pillow > 8 or more datasets; group C (28% of pillows): snow pillow near center of all datasets; group D (25% of pillows): snow pillow < 8 or more datasets.

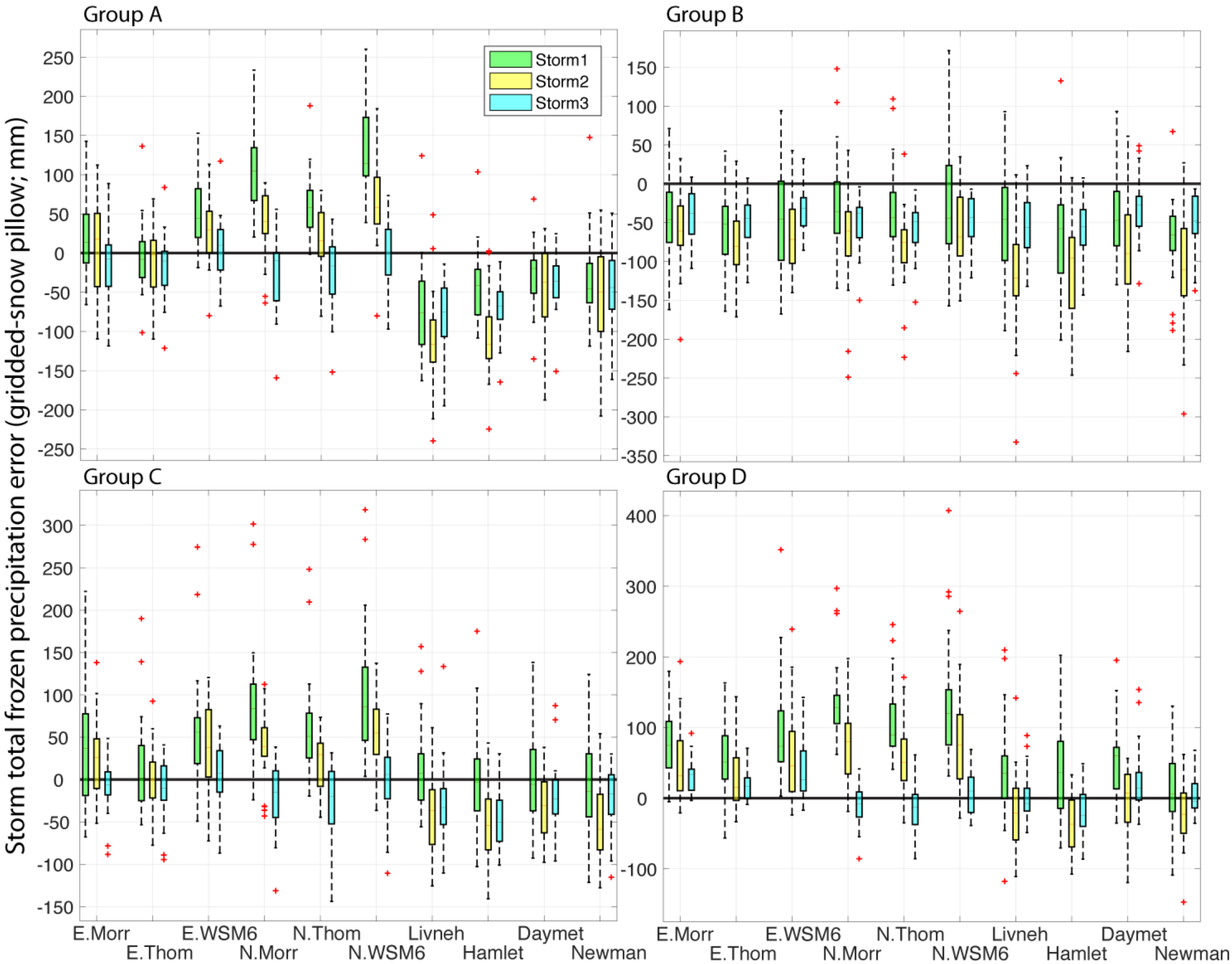

Fig. 8: Boxplots of errors (gridded-snow pillow) of 'storm' total 'frozen' precipitation (mm) for each of the 4 groups of snow pillows (A-D) outlined in the text, for three major storm periods highlighted with arrows in Fig. 5: Storm1: Jan 3-8, 2008; Storm 2: Jan. 26 – Feb. 5, 2008; and Storm 3: Feb. 19-26, 2008. Outliers shown with red + signs are +/- 2 standard deviations. Group A (21% of pillows): snow pillow and WRF > at least 2 gridded datasets; group B (25% of pillows): snow pillow > 8 or more datasets; group C (28% of pillows): snow pillow near center of all datasets; group D (25% of pillows): snow pillow < 8 or more datasets.

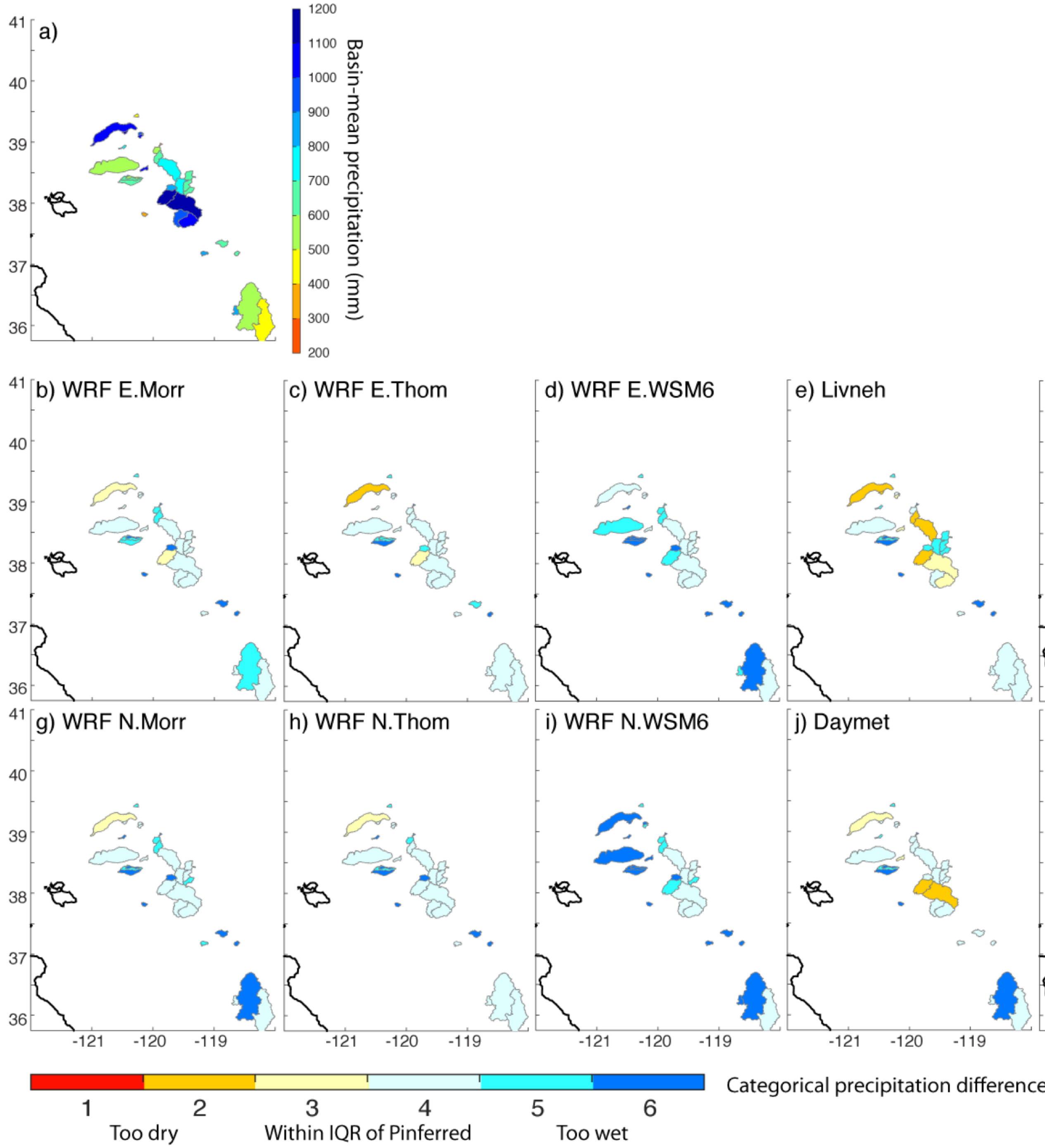

Fig. 9: (a) Median basin-mean WY2008 $P_{inferred}$ (mm). (b-k) Categorical difference of gridded dataset basin-mean precipitation (P) and $P_{inferred}$. Categories are: 1:$P <$ min $P_{inferred}$, 2: min $P_{inferred} < P < 25^{th}$ % $P_{inferred}$, 3: 25^{th} % $P_{inferred} < P < 50^{th}$ % $P_{inferred}$, 4: 50^{th} % $P_{inferred} < P < 75^{th}$ % $P_{inferred}$, 5: 75^{th} % $P_{inferred} < P <$ max $P_{inferred}$, 6: max $P_{inferred} < P$. Note that categories 3 and 4 are within the interquartile range of uncertainty.

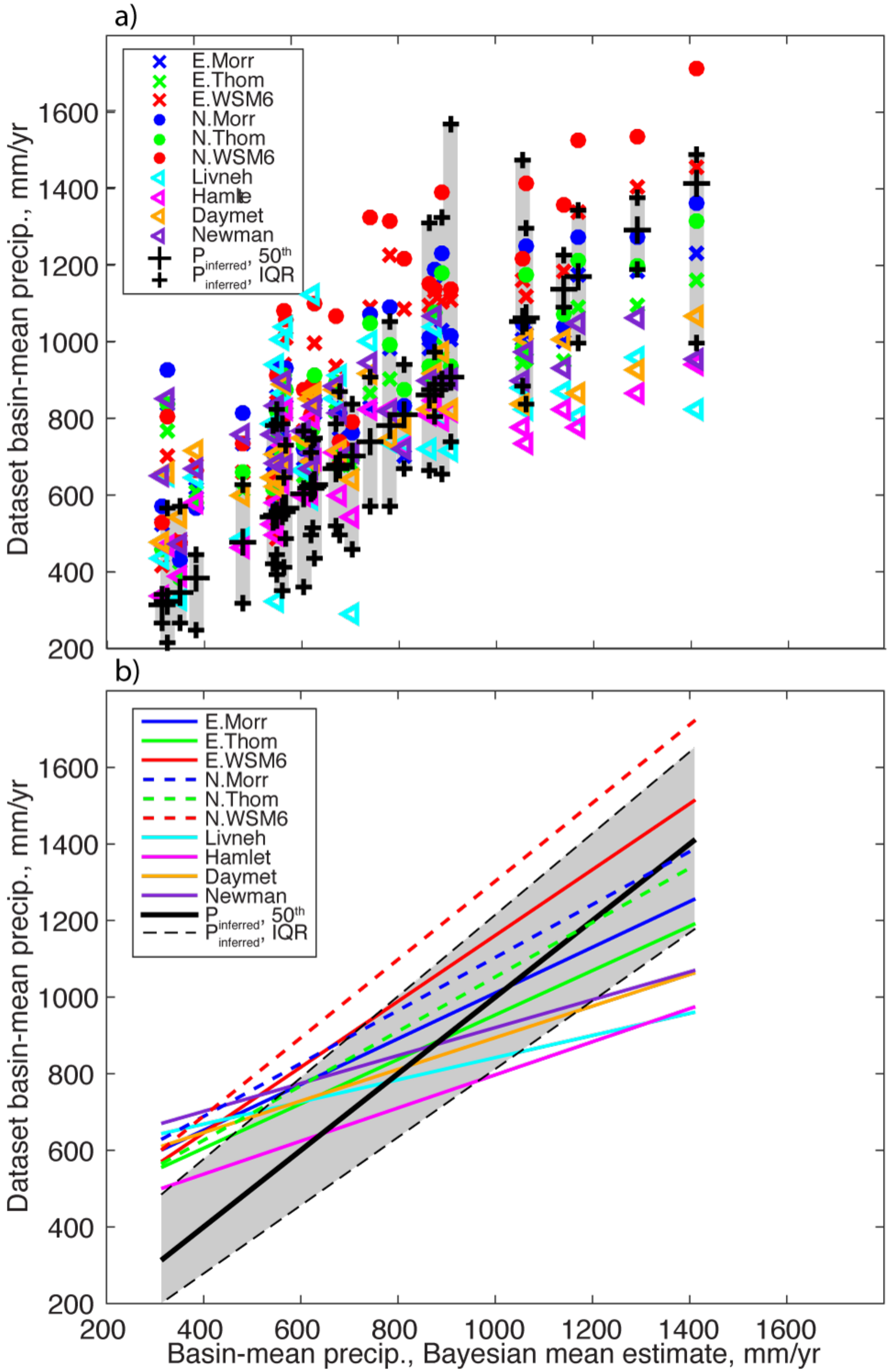

Fig. 10: (a) Basin-mean precipitation (mm) from gridded datasets (see legend), as a function of $P_{inferred}$ (black crosses). Large black crosses show median and gray shading bounded by smaller black crosses show interquartile range (IQR) of $P_{inferred}$. (b) Linear fit for each dataset. Black solid line shows Bayesian median and gray shading bounded by black dashed lines show IQR.