Data Adoption Justification Memo (for California's Fifth Climate Change Assessment)

LOCA Version 2 - Training Data

David W. Pierce, Stefan Rahimi, Sam Iacobellis Daniel R. Cayan, Julie Kalansky, Scripps Institution of Oceanography, UCSD & UCLA

CEC-funded agreement: EPC-20-006 Development of Climate Projections for California and Identification of General Use Projections

December 2023

Methods and Prior Relevant Work

All versions of the Localized Constructed Analogs (LOCA) statistical downscaling method use observationally-based training data in the bias correction and spatial downscaling processes. Since LOCA is designed to reproduce the statistics (e.g., seasonal variation, seasonal averages, etc.) of the training data in the final historical portion of the downscaled result, the training data plays a key role in the downscaling process and determines the climatology, annual cycle, and variability of the downscaled result.

In California's Fourth Climate Change Assessment (Pierce et al. 2018), LOCA version 1 was used to downscale global climate models (GCM) from CMIP5[1]. The training data in that work were based on station observations interpolated to a regular grid using a form of nearest-neighbor interpolation, then imposed on a specified monthly gridded climatology (Livneh et al., 2015; Livneh hereafter). Temperature was additionally corrected for elevation using a fixed lapse rate. Although this methodology has been used for decades to construct gridded data sets, it has some drawbacks. In particular, interpolating scattered station observations across regions of varied topography, such as found in California, may yield errors since the local topography at an unobserved location does not figure into the final result except insofar as the assumed gridded climatology reflects topography. The gridded climatology itself is an estimate since station observations in many regions are not available at the fine spatial resolution used for downscaling. In sum, the final gridded result supplies estimates of meteorological variables at unobserved locations that are informed by the values at neighboring stations, the elevation, and the estimated climatology.

LOCA version 2 (LOCA2 hereafter) is used in the current CMIP6[2]-based effort for California's Fifth Climate Change Assessment. Differences between LOCA version 1 and 2 are not described in this memo, which focuses exclusively on the training data. In California's Fifth Climate Change Assessment we have used an alternative method to construct the training data: the Weather Research and Forecasting model (WRF; Skamarock et al. 2019) is used to interpolate between stations rather than using nearest neighbor interpolation (details are given below). The motivation for this approach is the hypothesis that WRF's simulation of physical processes and topographical effects yields a better estimate of meteorological variables at

[1] Climate Model Intercomparison Project version 5

[2] https://wcrp-cmip.org/cmip-phase-6-cmip6/

unobserved locations than does nearest neighbor interpolation. Our evaluation of this hypothesis is described below.

Constructing the training data begins with the ERA5[3] reanalysis (Hersback et al. 2020), a state-of-the-art global atmospheric reanalysis at ~30 km spatial resolution that incorporates a large volume of weather observations in its generation. The ERA5 reanalysis is then dynamically downscaled by WRF to the 3 km spatial resolution used by LOCA2.

Even though ERA5 ingests large quantities of observations, both ERA5 and WRF, like all models, have biases that need to be corrected before the ERA5-WRF data can be used to train LOCA2. We therefore use station observations to bias correct the ERA5-WRF data. At each station and for each of the 12 months we calculate the difference between the station-observed value and the value from ERA5-WRF. We calculate these differences for all integer percentiles of the data from 1 to 99. Then for each combination of month and percentile we construct, over the California domain, a best-fit surface of the ERA5-WRF errors using the Generic Mapping Tools (GMT) "surface" function (Wessel et al. 2019). This surface yields an estimate of the ERA5-WRF bias at all locations, for each month-of-year and percentile. The bias is then removed arithmetically for temperature (and other non-positive definite variables) or multiplicatively for precipitation (and other positive definite variables). This methodology is similar to that used in Brown et al. 2016, although WRF-based interpolation between stations is not used in that work. We refer to the final bias corrected training data as ERA5-WRF-BC.

Downward surface solar radiation is the exception to this process. Our evaluations of WRF showed that coastal clouds are poorly represented in the ERA5-WRF simulations. Since these clouds have a strong bearing on rooftop solar photovoltaic electricity generation, we deemed the WRF data unsuitable for use as this variable's training data. We instead used GOES satellite observations as the basis for the surface downward solar radiation training data (Clemesha et al. 2016).

QA/QC and Uncertainty

We evaluated the ability of our methodology to estimate precipitation in unobserved locations using a cross-validation method. Twenty stations that have many decades of observations and cover a range of climate conditions across California were selected for analysis (Figure 1). For each station, we first constructed an entirely new gridded station data set using the Livneh nearest-neighbor methodology but leaving out the station in question. We then compared the estimated time series at the omitted station's location to the actual time series from the station. We then repeated the process for the 20 cross-validation stations using the ERA5-WRF data, performing the complete surface-fitting bias correction process but again omitting the station in question. This process could only be done for precipitation since that was the variable we had the facilities to process due to the work in Pierce et al. 2021. Nonetheless, this is a reasonable evaluation of the methodology since precipitation is heavily influenced by topography, covers a wide dynamic range across the landscape, and can be locally patchy.

[3] ERA 5 is the fifth generation European Center for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the global climate covering the period from Jan 1940 to present.

Results from the cross-validated analysis show that the Livneh gridded nearest-neighbor methodology is superior to ERA5-WRF-BC for capturing the specific day-by-day evolution of historically observed precipitation, presumably because neighboring stations are likely to experience precipitation at the same time as the omitted station. By contrast ERA5-WRF-BC produces precipitation via simulated processes that are close to, but not perfectly synchronized with observed precipitation on any given day. Therefore, if one wishes to know whether precipitation fell on some particular day (for example, for forensic meteorology applications), the nearest-neighbor regridding is preferable. On the other hand, over longer periods, errors in ERA5-WRF-BC tend to cancel out – Figure 2 shows that the RMS error in yearly mean precipitation in cross-validated ERA5-WRF-BC is 5.9% across the 20 stations, about half the error seen in the cross-validated Livneh style nearest-neighbor interpolation (10.2%). Similar results are seen in winter (DJF), spring (MAM), and autumn (SON). Errors in summer are large in both methods (~20%) but the dry summer conditions in California make the summer evaluation of little interest compared to the performance in seasons when precipitation occurs. ERA5-WRF-BC likewise outperforms the nearest-neighbor method for extreme precipitation values (Figure 3), specifically for daily precipitation return periods between 2 and 20 years. By 50 years the difference in RMSE between the two methods falls to only about 1 percentage point.

Guidance or Caveats on Best Practices for Use of Data Products

The purpose of the LOCA2 training data is twofold: 1) to provide unbiased estimates of meteorological values across the year and at different quantiles for use in the LOCA2 bias correction process; 2) to provide spatial patterns of the meteorological variable with the correct spatial patterns, means, and variability for use in the LOCA2 spatial downscaling analog day matching process. According to our cross-validation analysis the ERA5-WRF-BC data succeeds well for these purposes, outperforming the nearest-neighbor gridding approach when estimating data in unobserved locations. On the other hand, if an investigator needs to know if precipitation occurred at a particular location on a specific day for historical event-based or forensic meteorology applications, then the nearest-neighbor approach provides better results.

References

Brown, T., G. Mills, S. Harris, D. Podnar, H. Reinbold, and M. Fearon, 2016: A bias corrected WRF mesoscale fire weather dataset for Victoria, Australia 1972-2012.

Clemesha, R. E. S., A. Gershunov, S. F. Iacobellis, A. P. Williams, and D. R. Cayan, 2016: The Northward March of Summer Low Cloudiness along the California Coast. Geophys. Res. Lett., 43, DOI:10.1002/2015GL067081

Hersbach, H, Bell, B, Berrisford, P, et al. The ERA5 global reanalysis. *Q J R Meteorol Soc*. 2020; 146: 1999–2049. https://doi.org/10.1002/qj.3803

Livneh, B., T. J. Bohn, D. W. Pierce, F. Munoz-Arriola, B. Nijssen, R. Vose, D. R. Cayan, and L. Brekke, 2015: A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950-2013. Scientific Data, v. 2, article 150042 (2015). doi:10.1038/sdata.2015.

Pierce, D. W., J. F. Kalansky, and D. R. Cayan, 2018: Climate, drought, and sea level rise scenarios for California's fourth climate change assessment. California's Fourth Climate Change Assessment, California Energy Commission and California Natural Resources Agency, report CCCA4-CEC-2018-006

Pierce, D. W., L. Su, D. R. Cayan, M. D. Risser, B. Livneh, and D. P. Lettenmaier, 2021: An Extreme-Preserving Long-Term Gridded Daily Precipitation Dataset for the Conterminous United States. *Journal of Hydrometeorology*, v. 22, p. 1883-1895. https://doi.org/10.1175/JHM-D-20-0212.1

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, and X.-Y. Huang, 2019: A Description of the Advanced Research WRF Version 4. *NCAR Tech. Note NCAR/TN-556+STR*, 145 pp. doi:10.5065/1dfh-6p97

Wessel, P., Luis, J. F., Uieda, L., Scharroo, R., Wobbe, F., Smith, W. H. F., & Tian, D. (2019). The Generic Mapping Tools version 6. *Geochemistry, Geophysics, Geosystems*, 20, 5556–5564. https://doi.org/10.1029/2019GC008515

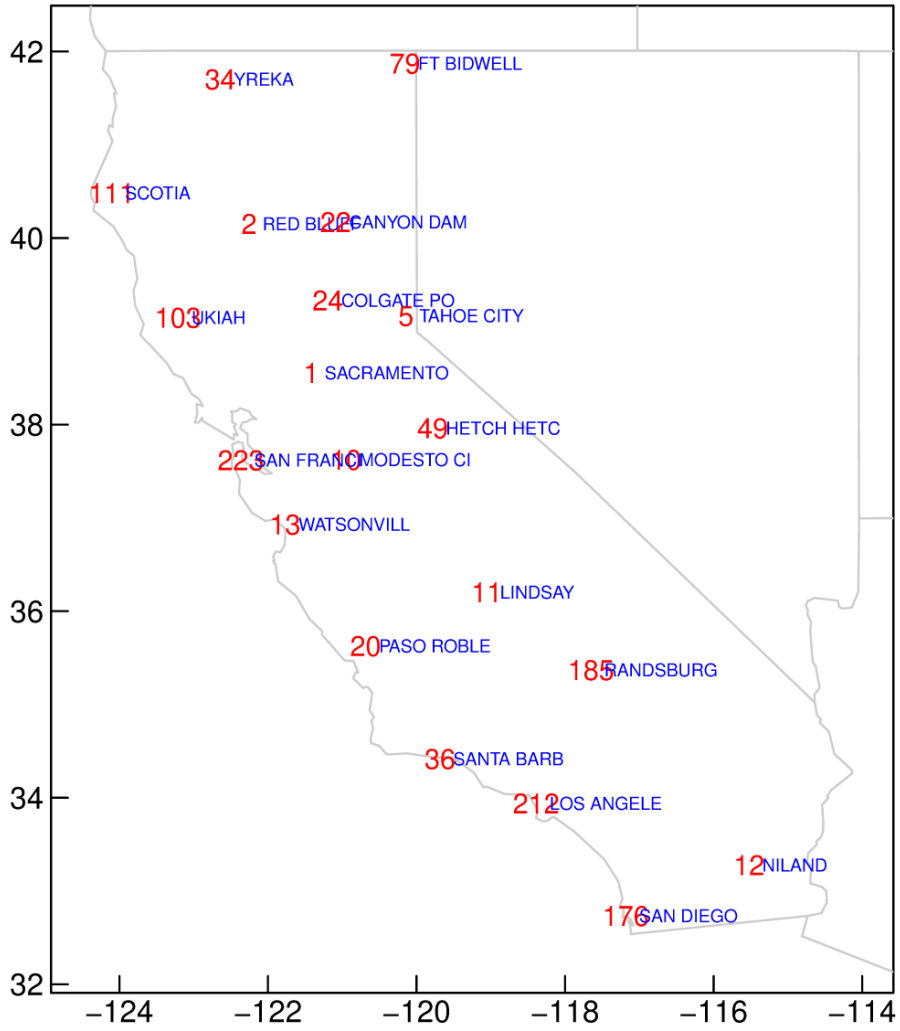

Figure 1. Locations of the 20 stations used in the cross-validation analysis. Red numbers are the rank of the station in data completeness amongst all California stations. Station names are shown in blue.

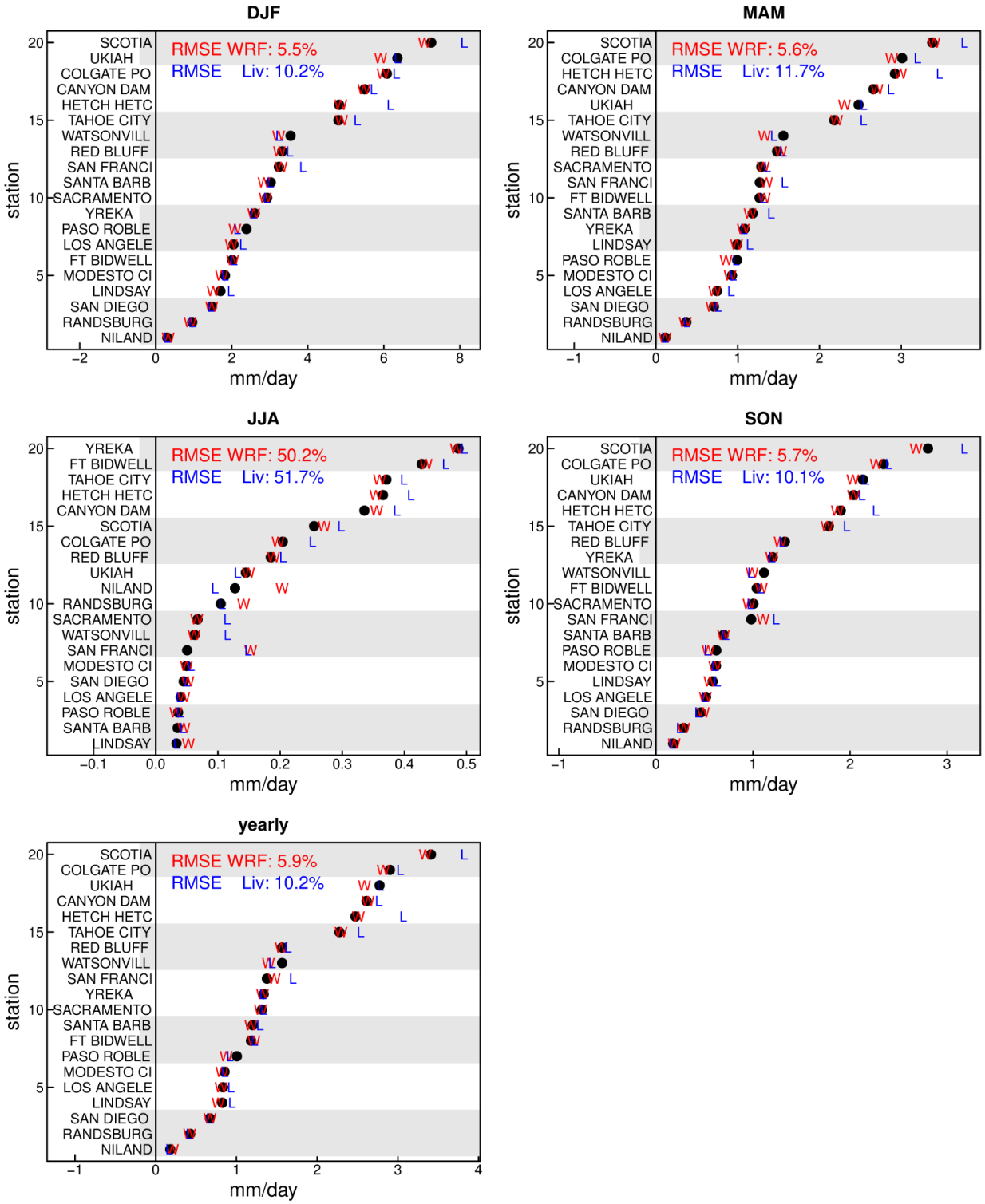

Figure 2. Mean daily precipitation by season at the 20 cross-validation stations from the original station observations (black dots), cross-validated Livneh results (blue L), and cross-validated WRF results (red W). In each panel stations are sorted from wettest to driest. The indicated RMSE values are calculated as percent errors across the 20 stations.

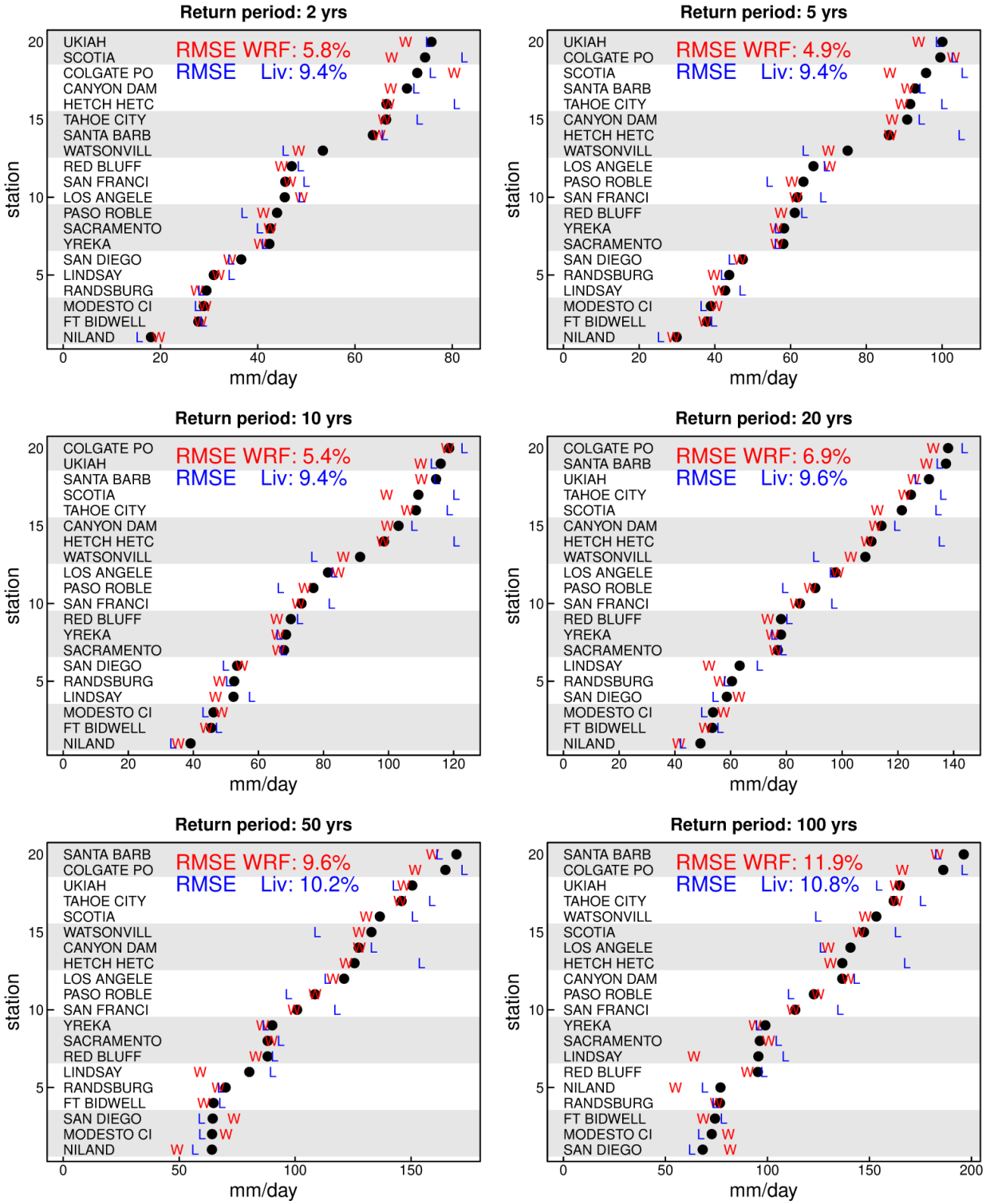

Figure 3. Return values (mm/day) of daily precipitation from the 20 cross-validated stations for different return periods from 2 to 100 years. Values from the original station data are shown as black dots, cross-validated Livneh values as the blue L, and cross-validated WRF results as the red W. The indicated RMSE values are calculated as percent errors across the 20 stations.