ARL-TR-9889 • Mar 2024

A 99-Day Assessment of the Weather Research and Forecasting Model over the Southwest United States, Volume 2: Spatial Distribution of Verification Scores

by John W Raby, Leelinda Dawson, and Robert Dumais

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9889 ● MAR 2024

A 99-Day Assessment of the Weather Research and Forecasting Model over the Southwest United States, Volume 2: Spatial Distribution of Verification Scores

John W Raby, Leelinda Dawson, and Robert Dumais
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
March 2024	Technical Report	10/01/2021	9/30/2022

4. TITLE AND SUBTITLE
A 99-Day Assessment of the Weather Research and Forecasting Model over the Southwest United States, Volume 2 – Spatial Distribution of Verification Scores

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
John W Raby, Leelinda Dawson, and Robert Dumais

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-ID White Sands Missile Range, NM 88002	ARL-TR-9889

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID IDs: Leelinda Dawson, 0000-0003-4209-8459; Robert Dumais, 0000-0001-5038-6843

14. ABSTRACT
An assessment was conducted over a 99-day period during winter over complex terrain to evaluate the accuracy of forecasts produced by the Advanced Research version of the Weather Research and Forecasting model (WRF-ARW). The Army Weather Running Estimate–Nowcast Real-Time (WREN_RT) system is a scripted system that provides forecasts by executing WRF-ARW and its preprocessors used to produce the WRF-ARW forecasts for this study. WREN_RT aims to provide forecasts for ingestion into decision aids that produce knowledge products for Warfighters. These products include the 2-D distribution of weather phenomena that can impact Army missions and systems. The traditional categorical method of spatial verification was used on the WRF output and ground truth data to compute skill scores for a range of neighborhood sizes and thresholds. The scores were computed and aggregated over the entire model domain to show how they are affected by terrain features. For ground truth data, gridded observations from the UnRestricted Mesoscale Analysis were used. The results of the assessment showed that WRF scored very well over some terrain features and land surface types and not so well over others.

15. SUBJECT TERMS
spatial verification, gridded observations, forecast, threshold, decision aid, weather impacts, terrain, Military Information Sciences

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	111

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
John W Raby	(575) 678-2004

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

Preface

The content included in the Executive Summary, Introduction, and Sections 2.2, 3.1, and 3.2 can also be found in the same sections in Raby et al.[1] because the input data used for this report is the same as that used in the previous report. Though this report relies on the same input data, it presents the results of new research conducted per a suggestion that was made in the Future Work section of the previous report.

[1] Raby J, Cai H, Dawson L, Reen B. A 99-day assessment of the Weather Research and Forecasting Model over the southwest United States. DEVCOM Army Research Laboratory (US); 2021 July. Report No.: ARL-TR-9237.

Executive Summary

An assessment of the Advanced Research version of the Weather Research and Forecasting (WRF-ARW) model[1] was conducted over a winter-season 99-day study period to quantify the accuracy of forecasts produced over the complex terrain of the southwestern United States and northern Mexico. The study focused on near-surface meteorological variables and cloud cover. Weather Research and Forecasting–Chemistry (WRF-Chem) model[2] is a version of WRF-ARW that contains a code module forecasting chemical constituents in addition to the standard meteorological forecasts. This evaluation used outputs of WRF-Chem configured to only include dust forecasts beyond the standard WRF-ARW fields (without allowing dust to impact radiation); since dust is not evaluated in this study (and dust does not affect other fields) the model used in the study will generally be referred to as WRF-ARW, or more simply WRF. The model forecasts evaluated were produced using the Weather Running Estimate–Nowcast Real-Time System (WREN_RT),[3] which is a US Army Combat Capabilities Development Command Army Research Laboratory–scripted system that performs data gathering, executes WRF preprocessors, and runs WRF itself with the goal of providing high-resolution forecast information.

The WREN_RT provides the forecasts that are ingested into decision aids. The decision aids apply weather thresholds to locate areas in time and space that exceed the thresholds and indicate the possibility of significant impacts. To evaluate the accuracy of the forecasts at high resolutions, high-quality gridded observations are needed for ground truth to perform spatial verification. For this assessment, the gridded observations used were the UnRestricted Mesoscale Analysis (URMA) produced by the National Oceanic and Atmospheric Administration for verification of Numerical Weather Prediction models.[4] The assessment involved comparing forecasts produced by the WRF model with URMA gridded observations over two domains located over the southwestern United States and northern Mexico. This assessment has the benefit of a significantly larger set of input data compared with

[1] Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda M, Huang XY, Wang W, Powers JG. A description of the advanced research WRF version 3. University Corporation for Atmospheric Research; 2008. Report No.: NCAR/TN-475+STR.
[2] Grell GA, Peckham SE, Schmitz R, McKeen SA, Frost G, Skamarock WC, Eder B. Fully-coupled "online" chemistry within the WRF model. Atmos Environ. 2005;39(37):6957–6975.
[3] Reen BP, Dawson LP. The Weather Running Estimate–Nowcast Realtime (WREN_RT) system, version 1.03. Army Research Laboratory (US); 2018 Sep. Report No.: ARL-TR-8533. https://apps.dtic.mil/sti/pdfs/AD1060869.pdf.
[4] De Pondeca Manuel SFV, Manikin G, DiMego G, Benjamin S, Parrish D, Purser RJ, Wu WS, Horel J, Myrick D, Lin Y, et al. The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: current status and development. Weather Forecast. 2011;26(5):593–612.

previous assessments that were limited to periods of less than 30 days. The longer time period results in this study having statistically stronger skill scores and statistics. The results of the study show the skill and accuracy of forecasts produced by the WRF for ingesting into decision aids, and that these metrics vary depending on the location in the domain and the thresholds used for determining weather impacts.

1. Introduction

The Army requires weather-knowledge products for mission support. Weather systems can traverse multiple domains interacting with the varied terrain and topography features to produce unique conditions depending on location. These phenomena can range from large-scale areas of precipitation or dust storms extending across hundreds of kilometers occurring over a period of 24 h or less to erratic wind-flow patterns associated with dense urban environments that occur on spatial scales of less than 1 km and time scales of a few minutes to 1 h.

1.1 Army Numerical Weather Prediction (NWP) for Weather-Impacts Prediction

To address the need for the prediction of atmospheric conditions over multiple domains, the Army has developed new microscale meteorological models (Wang and Benson 2021) and modified existing NWP models that employ a range of grid sizes, initialization techniques, and parameterizations to simulate weather phenomena across multiple spatial and temporal scales. The Army Weather Running Estimate–Nowcast Real-Time System (WREN_RT; Reen and Dawson 2018) executes the Advanced Research version of the Weather Research and Forecasting (WRF-ARW; Skamarock et al. 2008) NWP model to provide the forecast grids, which can then be ingested into the decision aids. The decision aids apply thresholds to these forecasts to determine the spatial and temporal distribution of weather conditions that can impact the effectiveness of multidomain formations.

1.2 Evaluation of Army NWP Weather Forecasts

To evaluate the accuracy of the forecasts at high resolution, advanced methods of model verification are needed to verify such high-resolution output spatially as opposed to the more traditional methods that perform point-by-point comparisons with observational ground truth data coming from weather observations. This grid-to-point approach to verification cannot adequately assess the true skill of high-resolution forecasts (Ebert 2008).

Traditional grid-to-point methods use point observations to verify the accuracy of NWP models in predicting continuous meteorological variables by computing such statistics as mean error (ME) and root-mean-squared error (RMSE) to characterize model accuracy over the entire domain. When these techniques are applied to high-resolution models such as the WREN_RT, the results can give misleading error estimates when compared with lower-resolution models, which often score better when using these techniques. The issue is the inability of the verification technique

to evaluate the true skill of higher-resolution forecasts, which replicate mesoscale atmospheric features in a way that is more representative of the actual phenomenon owing to their use of a finer grid over smaller domains, higher-resolution land-surface input data and models, and better parameterization of subgrid physical processes (Jolliffe and Stephenson 2012).

In recent years, various nontraditional verification techniques were developed that apply different approaches to show the value of higher-resolution forecasts. In particular, spatial verification techniques have been developed that overcome the limitations of grid-to-point techniques, which score on the basis of the exact matching between point observations and the forecasts at those points. Fuzzy verification, also known as neighborhood verification, is a spatial technique using an approach that does not require exact matching and instead focuses on how well the atmospheric feature or object is replicated by the model—even if there is a spatial displacement of the feature. Ebert (2008) reviews a number of such methods. The goal is to determine the amount of displacement by using a range of sizes of neighborhoods of surrounding forecasts and observed grid points in the verification process. In this way, model performance as a function of spatial scale can be determined to allow selection of the scale required to have the desired accuracy. For this study, the categorical verification framework was used, but instead of computing domain averaged skill scores and statistics as a function of model lead time over the 99-day period as was done by Raby et al. (2021), the scores and statistics were computed independently at each grid square in both domains and were aggregated over all model lead times for the 99-day period. These spatial verification methods require gridded observations instead of point observations for ground truth.

1.3 UnRestricted Mesoscale Analysis (URMA) Gridded Observational Ground Truth Data for NWP Evaluation

Sources of gridded observations are few, particularly at the spatial scale needed for Army weather-knowledge products tailored for multidomain formations operating in regions with varied and complex terrain conditions. For this study, the gridded observations used were the UnRestricted Mesoscale Analysis (URMA) (De Pondeca et al. 2011). URMA is used by the National Oceanic and Atmospheric Agency (NOAA) National Weather Service (NWS) for verification of NWP models. The Real-Time Mesoscale Analysis (RTMA), in conjunction with URMA, provides real-time, 2-D meteorological gridded analysis products produced from NWP analyses and hourly point weather observations from the national networks of Météorologique Aviation Régulière (METAR) and mesonet sensors that are distributed over the continental United States (CONUS). Two-dimensional

RTMA/URMA was developed by National Centers for Environmental Prediction (NCEP) in collaboration with the Earth System Research Laboratory and the National Environmental, Satellite, and Data Information Service (De Pondeca et al. 2011). RTMA/URMA is produced on an hourly basis using a mesoscale analysis background field produced from the 3-km High-Resolution Rapid Refresh (HRRR) model and the 3-km North American Mesoscale model downscaled to the 2.5-km grid as a first-guess background field (Morris et al. 2020).

For the URMA products used for this study, HRRR v2 on a 3-km grid was used (Benjamin et al. 2016). To fill in gaps at the edges of the domain, the most recent forecasts from the Rapid Refresh (RAP) were used (Morris et al. 2020). The RAP (RAP v3 for this study) provides an hourly forecast on a 13-km grid over North America (Benjamin et al. 2016). The first guess field is then adjusted through a 2-D variational data assimilation technique (2DVAR) to analyze point weather observations from the national networks of METAR and mesonet sensors (De Pondeca et al. 2011). The first cycle of the analysis is the RTMA on a 2.5-km CONUS grid that is used for weather situational awareness, calibration, and aviation safety. URMA is produced by rerunning the RTMA on the same grid 6 h following the first cycle to enhance the number of point observations used for analysis to make it a better product for model verification/validation (Pondeca et al. 2015). For example, NOAA uses URMA gridded observations for verification and bias correction of the National Blend of Models used by NWS forecasters (Ruth et al. 2017). URMA also serves as the NWS Analysis of Record (UCAR 2015). A future development anticipated for the RTMA/URMA analysis system is the 3-D RTMA, which is planned to provide 3-D analysis fields with subhourly updates (Weygandt et al. 2019).

A number of studies have been conducted to compare the RTMA with observations. Morris et al. (2020) reviews the results from a few such studies and presents the results, which focused on performing an assessment of the RTMA to evaluate its value as an alternative source of weather observations for use by airports for current conditions affecting safety of flight. Their study consisted of running data-denial experiments for retrospective periods of time that involved generating RTMA output using specified ingest configurations. These configurations allowed the assimilation phase to be controlled to restrict the available observational data to create three distinct cases. The cases were 1) CONTROL case, which assimilated all expected observations considered to be a more typical or normal scenario, 2) EXP case, which denied access to observations coming from certain airports considered to be a rare but not unprecedented scenario, and 3) NODA case that denied access to all observations, which is the worst-case scenario. They determined the RTMA could be used as a substitute for airfield weather

observations under certain conditions, for only certain meteorological variables, and only at certain locations. This is the most complete assessment compared to any others investigated. The previous studies focused on evaluating the RTMA using independent analyses products and controlled data-denial experiments and not on providing a quantitative, grid-to-point verification over a longer, continuous period.

To address the lack of a quantitative evaluation of URMA, Raby et al. (2020) conducted an evaluation of the URMA during a continuous "winter" period from 11 Nov 2016 to 17 Feb 2017 over a large domain encompassing much of the western United States, northern Mexico, portions of the Gulf of Mexico, Sea of Cortez, and the eastern Pacific Ocean. This domain was also the outer nest region (d01) for the WRF simulations produced during the same time period for the subnests used in this study. The evaluation compared the URMA values for near-surface meteorological variables to point observations of the same variables using a traditional grid-to-point verification technique that generated continuous error statistics over the 99-day time period.

The results of the evaluation showed the URMA provided a good analytical product for use as ground truth with certain limitations. The limitations are attributable to 1) use of the grid-to-point verification technique for high-resolution forecasts and 2) use of point observations. This first limitation refers to the requirement for exact matching between the forecast value (in this case the URMA value) at the location of the point observation, which leads to double-penalty errors for the forecast object being slightly displaced in space from the observed object and gives no credit for a near-miss situation where the forecast (URMA) object, despite replicating the observed object quite well spatially, is displaced in location and/or time. The second limitation arises from two sources. One source is that the URMA product is generated from the same point observations that are being used for verification. The other source is the fact the verification was conducted only at the locations in the URMA grid where there were point observations and nowhere else, leaving areas where there is no verification. The combined effect of these limitations on the accuracy of the URMA error statistics generated from the evaluation is difficult to quantify, as well as their impact on this assessment of WRF. That said, with no other source of better ground truth and given the acceptance of URMA by NOAA as the analysis of record to be used for verification, this study does provide reasonable evidence of the performance of WRF based on a 99-day data set of simulation and URMA gridded observational data.

2. Design of the Assessment

2.1 Verification Approach

The approach used for this objective assessment was based on a suggestion from Raby et al. (2021). The authors of this report recommended that an assessment of the WRF be conducted over each of the two domains independently to focus on the impact of the domain location, size, and geography on model errors. The scores and error statistics were computed over each domain separately using the categorical verification framework. The scores and error statistics for each grid square were aggregated over the 99-day period to provide a 2-D distribution of the metrics over each domain. The scores and statistics computed were the Critical Success Index (CSI), and the frequency bias (FBIAS). Traditional categorical verification scores and statistics are computed by defining an event from both the forecast and the observation grids. The event is defined by applying a threshold over the entire domain as the basis for determining "hits" or "misses," which follows the established theoretical framework for evaluating deterministic binary forecasts. A CSI value of 1.0 indicates a perfect forecast. FBIAS is the ratio of the numbers of forecast events and the number of observed events. A value of 1.0 for FBIAS is optimal, while less than 1.0 shows an under-forecast tendency and greater than 1.0 shows an over-forecast tendency. This framework evaluates the forecast skill by counting the numbers of times the event was forecast—or not—and observed—or not—in a contingency table.

Although categorical scores and statistics have been widely used, they are not always reliable for assessing the skill of high-resolution forecasts due to their sensitivity to observed rate (Mittermaier et al. 2013). Raby (2016) determined that combining categorical scores and statistics with those computed using a fuzzy verification approach provides a more comprehensive assessment of model performance. To overcome the limited applicability of scores and statistics generated from small data sets for inferring information about the true accuracy of the model, Raby and Cai (2016) suggested using a more rigorous approach that requires the generation of larger data sets of forecast output and gridded observations so that more reliable statistical results can be obtained. This approach is intended to improve the validity of scores and statistics, particularly when observed event rates are low due to the use of very-high or very-low threshold values of interest to the Army for predicting impacts to systems and missions. For this reason, the decision was made to use WREN_RT for generating output from the WRF model, which was run daily producing 24 hourly forecasts from 1200 to 1100 UTC, and the hourly URMA gridded observations for the same hours over an extended time period. The period chosen was 11 Nov 2016 to 17 Feb 2017 because

there were no significant changes made to the WREN_RT software over this time period and because of the availability of URMA gridded observations produced using a single software version. This period contained 99 days that were characterized as having typical winter conditions for the southwestern United States and northern Mexico and coincides with the period of the URMA evaluation conducted by Raby et al. (2020).

2.2 Verification Domains

The two domains selected, shown in Fig. 1, were located over the southwestern United States and northern Mexico. These domains were also the middle (3-km) nest and the inner (1-km) nest for the WRF simulations produced during the 99-day time period.

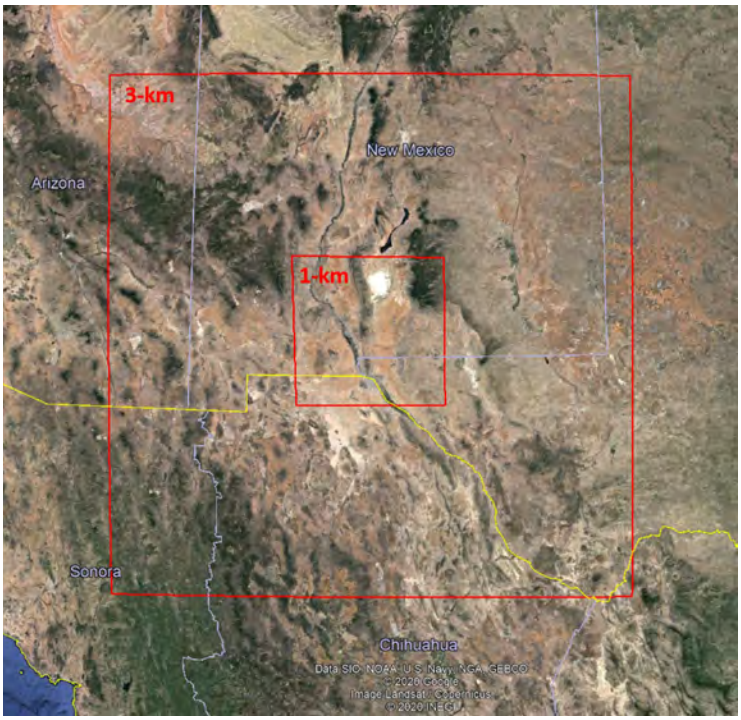

Fig. 1 Verification domains

The verification was conducted over the two domains, both characterized by a complex, mountain–desert–basin terrain landscape using hourly WRF forecasts and URMA gridded observations collected for the 99-day winter period.

Figures 2 and 3 show the 1-km and 3-km domains, respectively, with terrain elevation contours and geographic points of interest. The contour interval is 200 m.

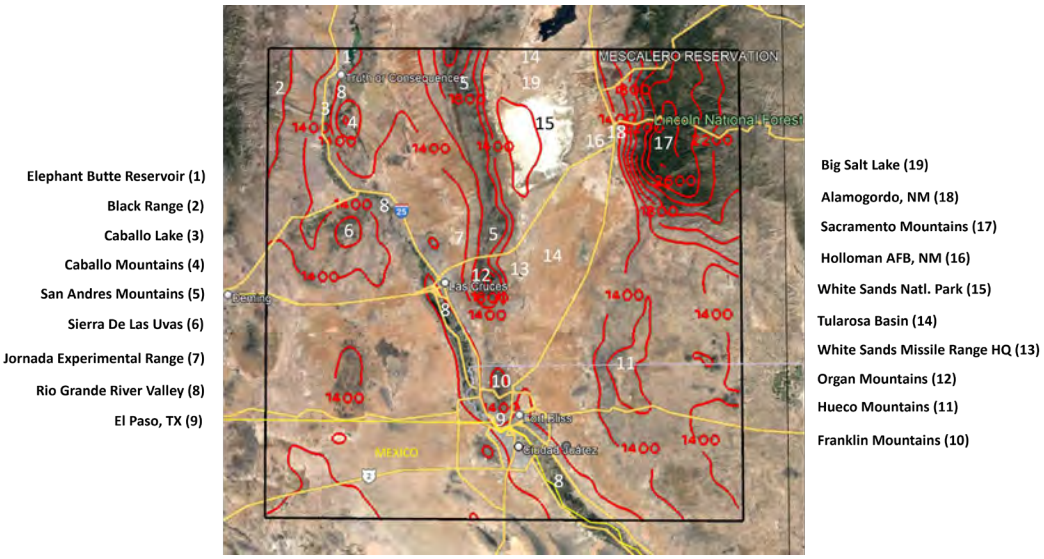

Fig. 2 1-km WRF verification domain

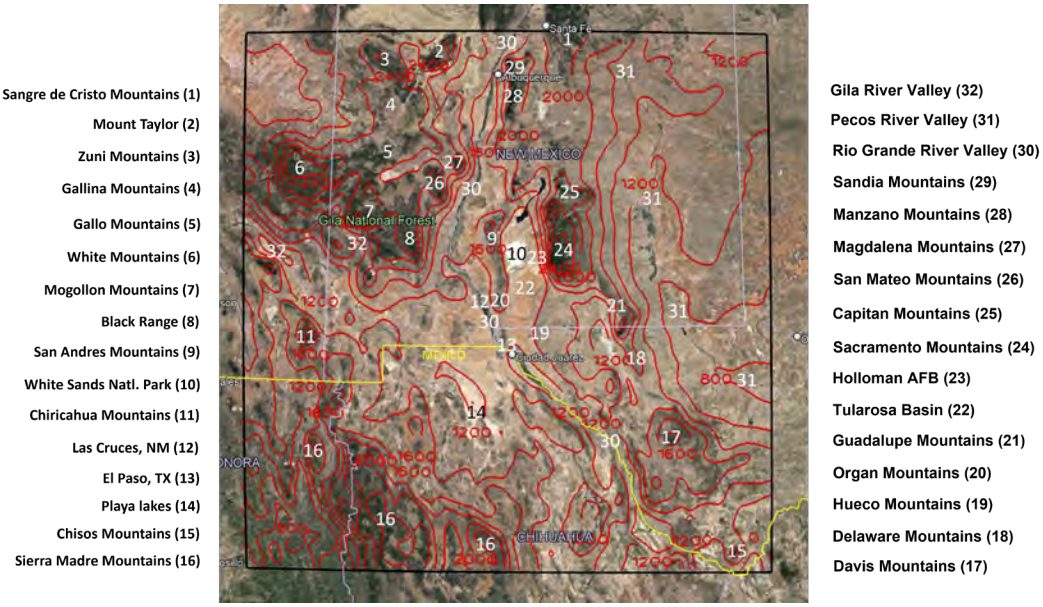

Fig. 3 3-km WRF verification domain

To better understand the patterns and signatures observed in the scores and error statistics attributable to sharp spatial contrasts in land surface characteristics, the distributions of these features from the Noah land surface model (LSM) used by the 1-and 3-km WRF were analyzed. Figure 4 shows the distribution of the simulated vegetation type for the 1- and 3-km WRF. Figure 5 shows the distribution of skin temperature for the 1- and 3-km WRF. The skin temperature features, being dynamic in response to other model fields, represent a sample taken from one forecast valid time. Figure 6 shows the distribution of land versus water surface

type for the 1- and 3-km WRF. Figure 7 shows the distribution of soil type for the 1- and 3-km WRF. Figure 8 shows the distribution of soil type for the URMA gridded observations. Figure 9 shows the distribution of vegetation type for the URMA gridded observations.

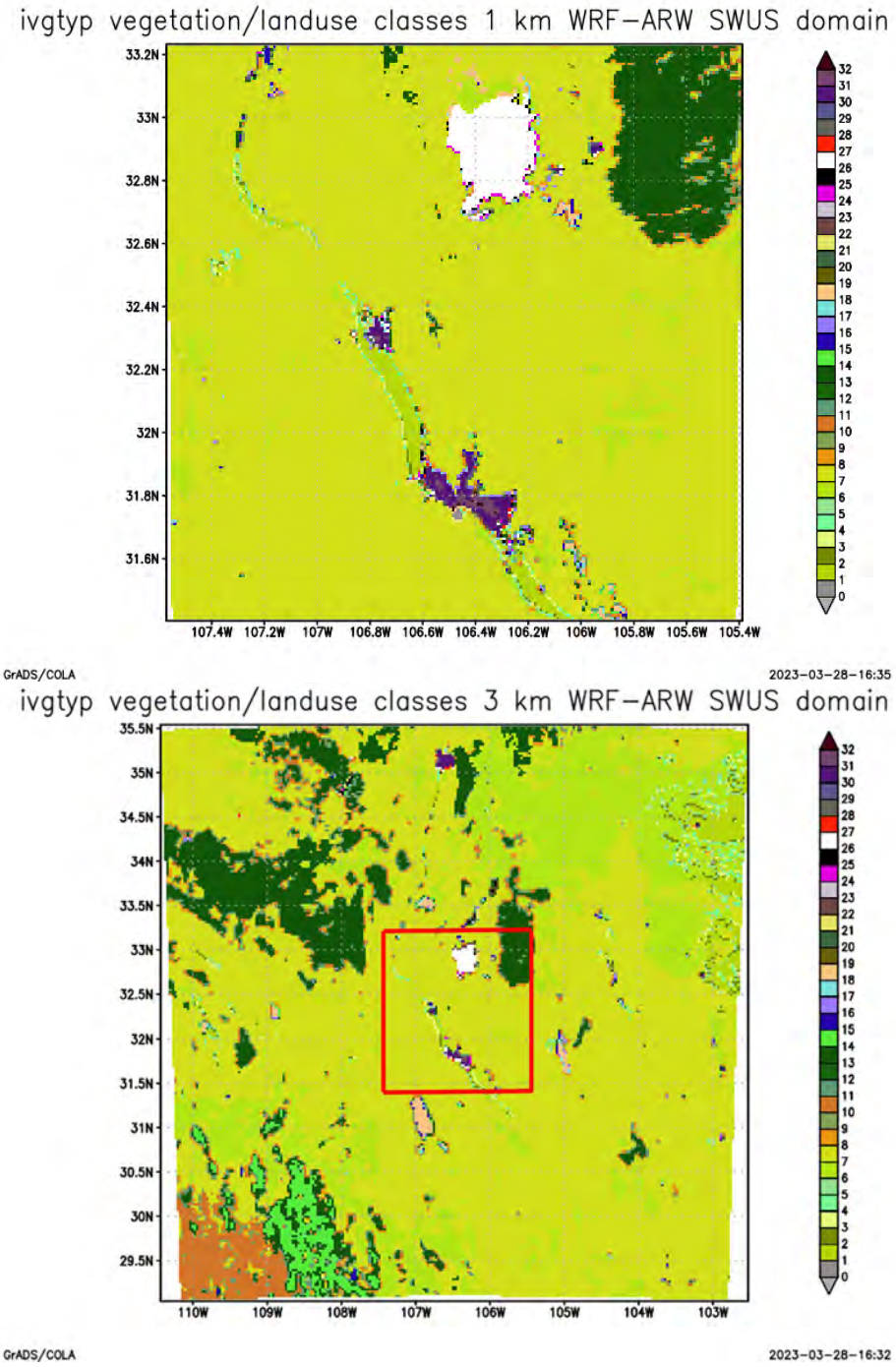

Fig. 4 Vegetation type from Noah LSM for the 1- and 3-km WRF

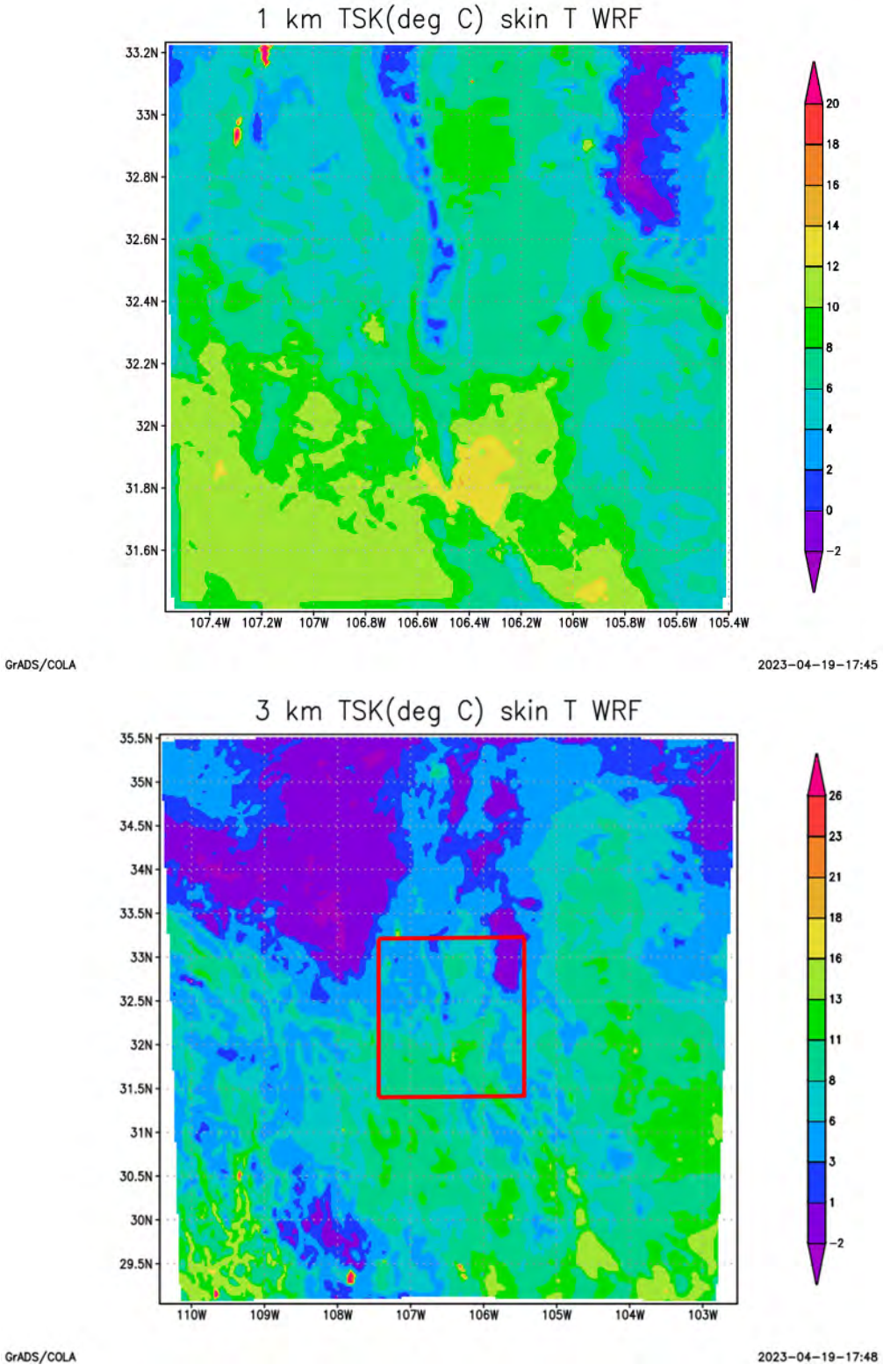

Fig. 5 Skin temperature for 1- and 3-km WRF from Noah LSM

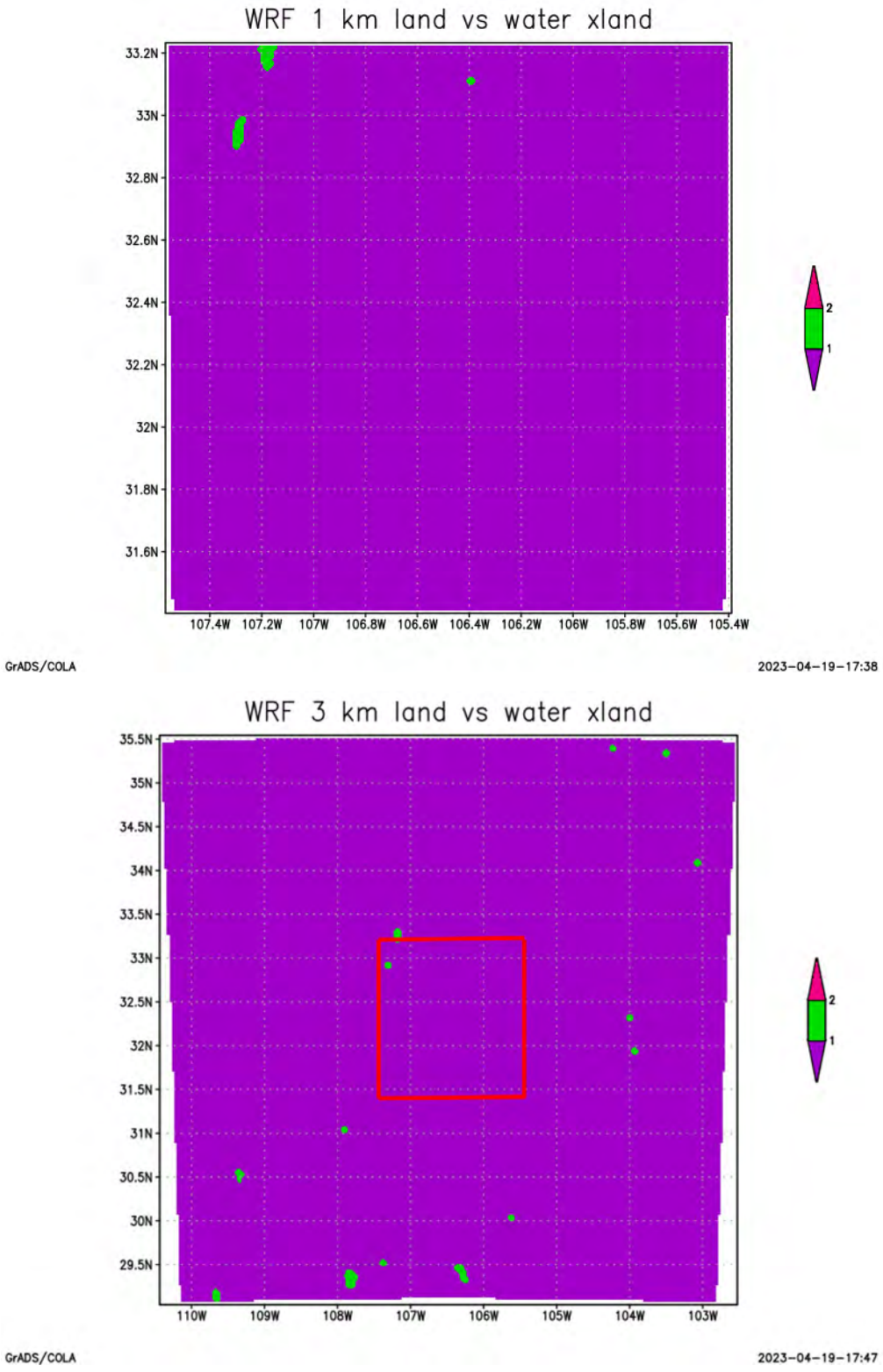

Fig. 6 Land vs. water type for 1- and 3-km WRF from Noah LSM

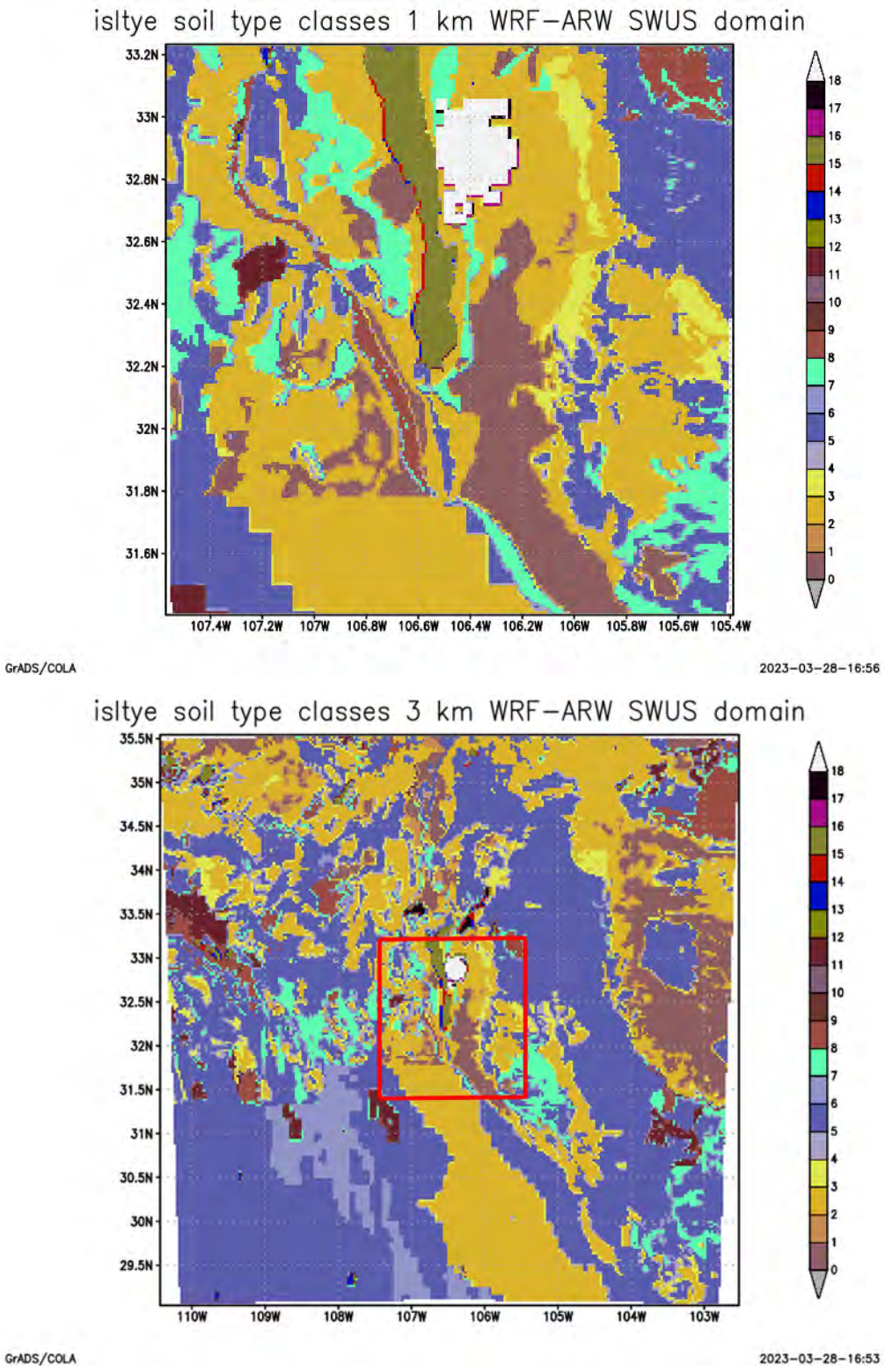

Fig. 7 Soil type for 1- and 3-km WRF from Noah LSM

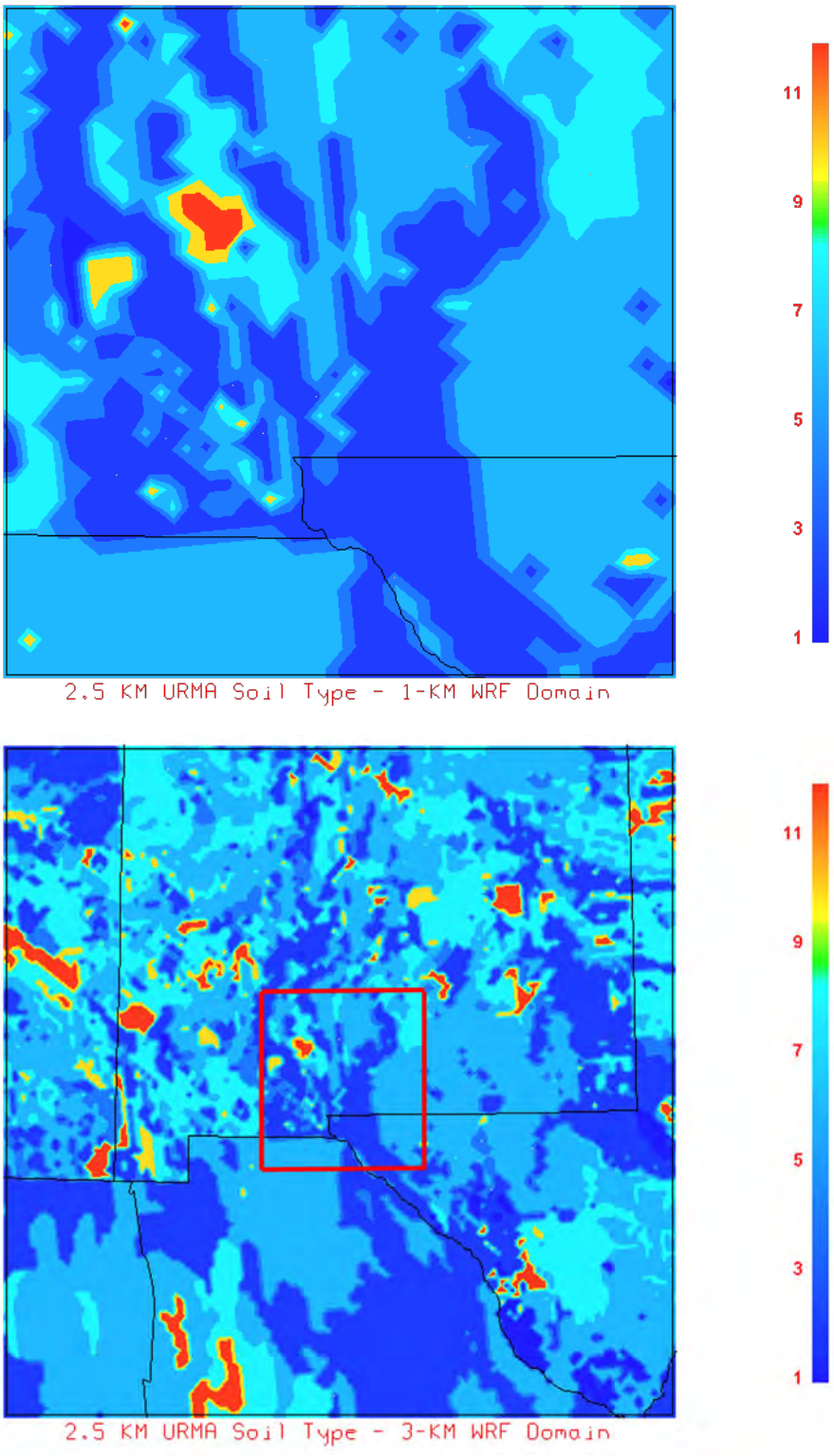

Fig. 8 Soil type for the 2.5-km URMA observations over the 1- and 3-km WRF domains

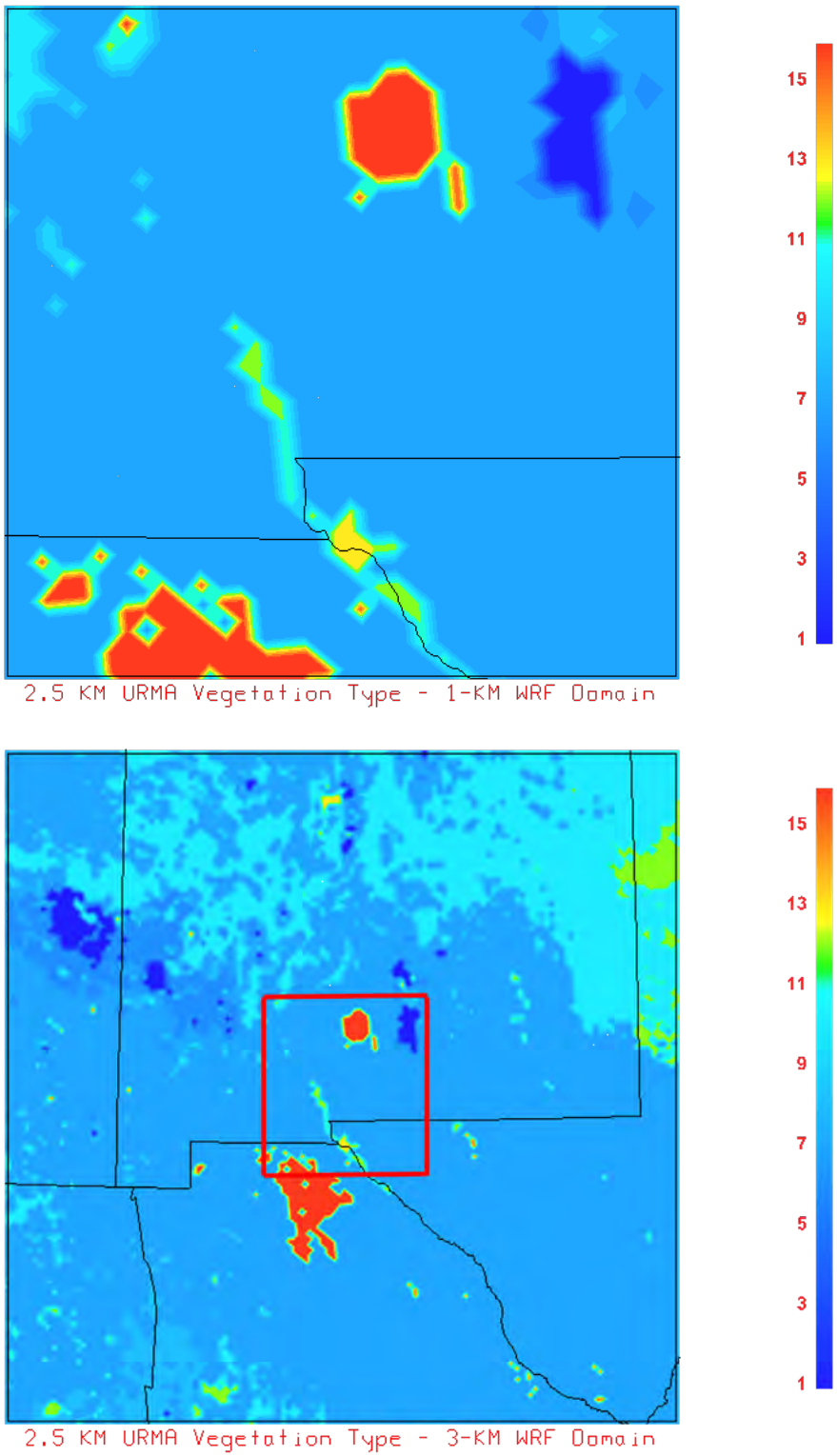

Fig. 9 Vegetation type for the 2.5-km URMA observations over the 1- and 3-km domains

2.3 Areas That Present Challenges for Verification

There are various challenges that impact the quality of the assessment for high-resolution mesoscale NWP. The first is a lack of sufficient point observations in certain areas of the resolved domain, particularly over features such as the Carrizozo Lava Fields/Malpais and White Sands National Park (WSNP) dunes. In the western United States, having sufficient point observations for ground truth presents a challenge for verification as well as for model initialization. Over a western US location such as Dugway Proving Ground, observations over the Great Salt playa have been shown to be of potentially great benefit during the MATERHORN campaign (Massey et al. 2017). Along regions where there are international borders, observation density can also show great variation (such as in Mexico, where observation density is much worse than in the United States). In very high and rugged mountain ranges and over small, isolated inland water bodies, observations also tend to be quite scarce. Finally, observation representativeness is also an issue of consideration. For example, if observations are taken over or within land use and roughness elements not resolved by the model, this will not reflect the model output very well (Stauffer 2012).

The URMA gridded observations are dependent on having good representation of near-surface meteorological conditions that can only come from point measurements. The mountain–desert–basin geography makes it difficult to place and properly site weather sensors as well as presents challenges for high-resolution NWP models when attempting to resolve small-scale land surface features in complex terrains. The URMA background analysis field from HRRR, which is adjusted to the point observations, is itself adversely impacted by these challenges. Figure 10 shows a typical distribution of the locations of point observations in the 3-km WRF domain (d02). Figure 11 shows a typical distribution of the locations of point observations in the 1-km WRF domain (d03).

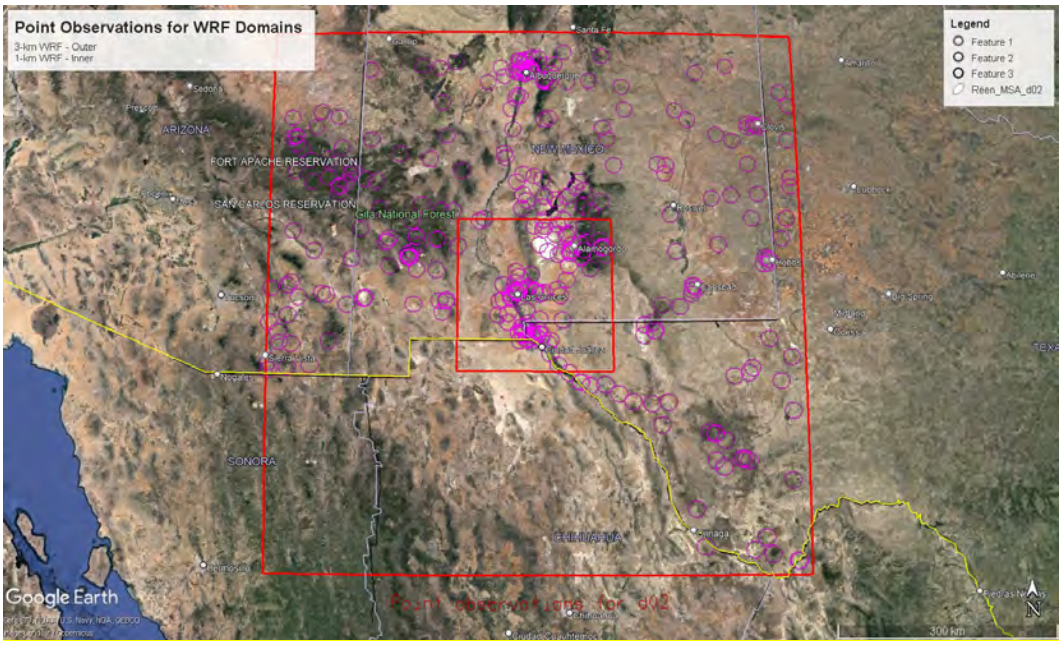

Fig. 10 Point observations for 3-km WRF domain

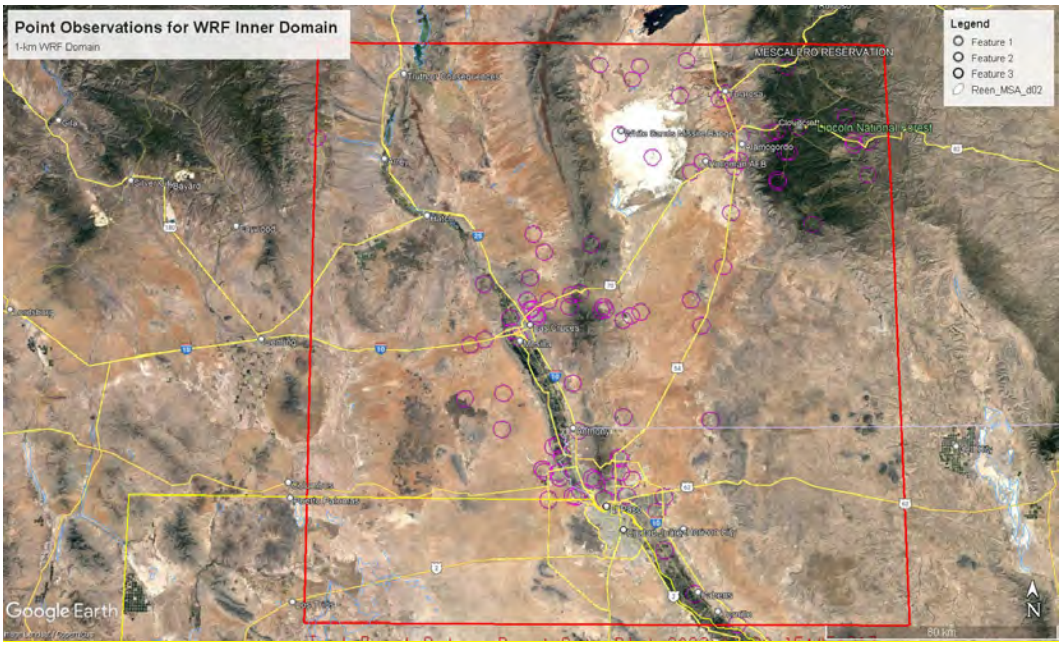

Fig. 11 Point observations for 1-km WRF domain

The distribution over the 3-km domain shows multiple areas that have sparse coverage of observations. Most notable of these are the southern third of the domain especially over Mexico, the northwest corner of the domain near and west of the Zuni Mountains, and the eastern third of the domain over eastern New Mexico and west Texas. Over the 1-km domain, areas of sparse observations are located over

the eastern and western thirds of the domain especially the southwest, southeast, and northwest corners.

In the two domains used for this study, the features that manifest these challenges are the white sand dune fields of the WSNP, the small inland lakes (namely Caballo Lake and Elephant Butte Reservoir), and high-elevation mountainous areas such as the Black Range, San Andres Mountains, and Sacramento Mountains. These features are simulated in varying degrees in the Noah LSM for the 1-km WRF and to a lesser extent in the 3-km WRF in Figs. 4–7.

The dune fields of the WSNP had no point observations during the 99-day period with only two nearby mesonet stations in the areas outside of the largest dune fields that provided observations. Being such a distinct land surface feature compared to surrounding areas in the Tularosa Basin, the contrast it presents in terms of the local physical and thermal properties and their fluxes is likely to impact the evolution of the lower planetary boundary layer (PBL) in a unique way over the diurnal period. Gunn et al. (2021) conducted a study in the dune field that involved making measurements of the wind speed profile using LIDAR over the course of the field campaign. They concluded that the diurnal change in atmospheric stability was more pronounced than that predicted by Monin-Obukhov similarity theory under conditions of strong periodic solar radiation and resultant non-steady state stratification of the PBL causing stronger afternoon surface winds. They documented this phenomenon at 45 other dune fields worldwide (Gunn et al. 2021).

Another possible source of uncertainty may be due to the quality of the land surface characterization of white sands from the Noah LSM used by the WRF. In the Noah land surface physics option of WRF, many parameters of the land use and soil are set in static tables named VEGPARM.TBL and SOILPARM.TBL while soil moisture is typically initialized from the external model (typically of poorer spatial resolution) being used for initial and lateral boundary conditions (usually GFS, NAM or HRRR). Methods such as spinning up a preforecast period of the model to develop soil moisture at the right model spatial scales or using high resolution soil moisture products assimilated elsewhere (Case et al. 2016) are some ways to alleviate the soil moisture issue, but they require extra steps and care. These factors point out that having no observational data over an area with a unique land surface very distinct from the surrounding areas could contribute to errors in the URMA observations as well as to the WRF simulation results for this area. Another feature having a very unique land surface that has no observations is the Carrizozo Lava Malpais. This area lies just north of the northern boundary of the 1-km WRF domain.

There are three significant bodies of water that are in the two domains. These are the Elephant Butte Reservoir, Caballo Lake, and Big Salt Lake. Figure 2 shows their locations in the 1-km WRF domain. The accuracy of the water surface temperature values used when initializing the WRF and HRRR models is critical to establishing the skin temperature values used as input to both models. The source of this data for the WRF is the real-time, global, sea surface temperature (RTG SST) data produced by NOAA. For small inland lakes, the values for the water surface temperature are interpolated from the global grid. However, this data is derived from in situ measurements and from satellite sensors and has its own uncertainty. The result is that interpolated values to marginally resolved small lakes may contribute to errors in the URMA observations and the WRF simulations.

Though there are several areas where the coverage of point observations is sparse, of particular concern is the lack of sufficient observations over the Sacramento Mountains in the northeast corner of the domain. This area has the highest elevations in the inner domain and complex terrain that presents challenges for the WRF as well as the URMA observations. The drastic change in elevation from the Tularosa Basin to the peaks of the mountains in this area is not completely resolved by the models, especially for the steepest slopes.

3. Generation of Assessment Data

3.1 Model Evaluation Tools

The software used to perform the scores and error statistic calculations was the Model Evaluation Tools (MET) (Jensen et al. 2020). The MET was developed at the National Center for Atmospheric Research (NCAR) through grants from the United States National Science Foundation (NSF), NOAA, the United States Air Force (USAF), and the United States Department of Energy (DOE). NCAR is sponsored by the NSF. The MET Series-Analysis tool enabled the aggregation of the error statistics for each grid square over all 99 days. Visualizations of the 2-D distribution of the scores and error statistics from Series-Analysis were produced using MATLAB.

3.2 Assessment Data

The URMA gridded observations used for this study were collected from the real-time repository operated by NCEP (NOAA 2017).

The forecasts were created with WREN_RT using WRF-ARW V3.8 and the WRF Pre-Processing System V3.8.1. Nested 9-, 3-, and 1-km horizontal grid spacing

domains centered just south of the White Sands Missile Range (WSMR), New Mexico, (Fig. 12) were executed for each day with 57 vertical full levels. The number of grid points in the three domains are 9 km: 279 × 279, 3 km: 241 × 241, and 1 km: 205 × 205. The 3-km domain covers about 12.4 times as much area as the 1-km domain, and thus the 1- and 3-km domain overlap for only 8% of the area covered by the 3-km domain. Each day a 3-h data-assimilation preforecast (0900–1200 UTC) preceded a 24-h forecast from 1200–1200 UTC. (This study uses the 0–23 h forecast within this period.)

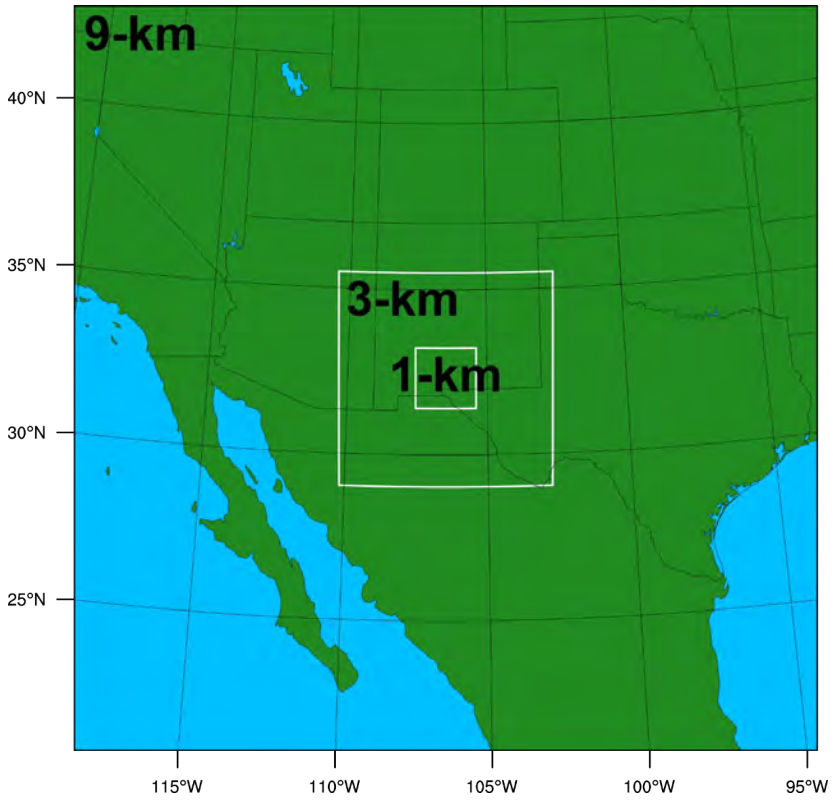

Fig. 12 Area covered by the 9-, 3-, and 1-km WRF domains

Initial conditions were created by using Obsgrid (NCAR 2016) to perform multiscan Cressman analyses with observations using 0.5° Global Forecast System model output as the first guess field. Observations were obtained from NCEP's Meteorological Assimilation Data Ingest System (MADIS; madis.noaa.gov). Specifically, the MADIS surface, maritime, radiosonde, profiler, and Aircraft Communications, Addressing, and Reporting System (ACARS) data sets were used. In additional to being used in the initial conditions analysis, these observations were also applied in observation nudging data assimilation (Reen 2016) during the preforecast from 0900 to 1200 UTC (the nudging terms ramp down in the following hour after 1200 UTC, but no observations valid after 1200 UTC are nudged toward). Observation nudging of wind, potential

temperature, and water vapor mixing ratio is applied with a weighting of 6×10^{-4} s^{-1}. The base horizontal radius of influence for the 9-, 3-, and 1-km domains are 120, 45, and 20 km, respectively, while the actual radius of influence increases linearly with decreasing pressure to twice this value at 500 hPa and is half this value at the surface. Observations are nudged in a 3-h time window centered on the valid time of the observation with linearly decreasing temporal weight in the outer half of the time window (for surface observations the time window is two-thirds as large).

The planetary boundary layer scheme used was the Mellor-Yamada Nakanishi Niino (MYNN) Level 2.5 scheme (with the MYNN surface-layer scheme) (Nakanishi and Niino 2006). Microphysics were parameterized using the Thompson aerosol-aware scheme (Thompson and Eidhammer 2014). The Grell–Freitas ensemble cumulus parameterization was used (Grell and Freitas 2014). For radiation, the RRTMG (rapid radiative transfer model for general circulation models) shortwave and longwave schemes were employed (Iacono et al. 2008). The Unified Noah LSM was used to simulate the land surface (Tewari et al. 2004). The simulations use WRF-Chem with dust-only enabled using the Air Force Weather Agency's dust scheme (WRF namelist settings chem_opt = 401, dust_opt = 3); however, dust forecasts are not evaluated in this report (Jones et al. 2010, 2012).

3.3 Verification Data Preprocessing

Preprocessing tasks as described in Raby et al. (2021) were required before both the URMA gridded observations and WRF forecasts for all 99 case-study days could be ingested into the MET Series-Analysis tool to produce 2-D distributions of the error statistics, as shown in Fig. 13. The scripts were developed and implemented in Python to make the preprocessing and postprocessing tasks easier and more efficient resulting in the generation of verification data that is better organized compared to running the tool on its own. The 24 hourly URMA gridded observations in GRIB2 format from the evaluation study described in Raby et al. (2020) were used as observation input into the MET Series-Analysis tool. Next, the 24 hourly WRF forecasts were postprocessed using Unified Post Processor (UPP) developed by NCEP (NCEP 2020) and a Python script, *rename_upp.py*, was used to rename each hourly, postprocessed WRF forecast output file for all 99 case-study days to a standard filename convention. Then, the renamed, postprocessed WRF forecasts were used as forecast input into the MET Series-Analysis tool. Both the URMA gridded observation files and the postprocessed WRF forecasts were ingested into the MET Series-Analysis tool using another Python wrapper script, *run_met_tools.py*, that extends and performs the automation of the run and data processes associated with several MET tools used by the US Army Combat

Capabilities Development Command (DEVCOM) Army Research Laboratory. (ARL) The process of this script is comparable to the one described in Dawson et al. (2016). However, since this was the first-time using the MET Series-Analysis tool, the script was modified to perform an additional step of automatically creating the forecast and observation lists, which are text files that list the directory locations and names of the forecast and observation files for all the 99 case-study days as shown in Fig. 13. For each MET Series-Analysis run, two file lists are automatically created by the script, before running MET Series-Analysis, one for the listing of the forecast files with the prefix of *series_analysis_forecast_file_list* and another one for the listing of the observation files with the prefix of *series_analysis_obs_file_list*. The automated creation of forecast and observation file lists before running the MET Series-Analysis tool allowed the quick generation of the scores and error statistics data for the study described in this report considering the large amount of both forecast and observation input data.

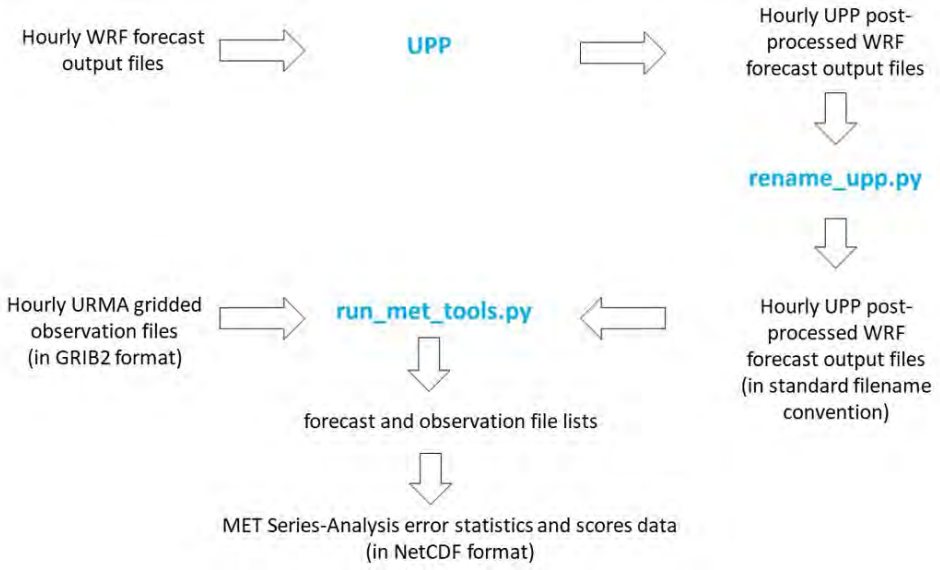

Fig. 13 Generation of verification-data flow diagram using the MET Series-Analysis tool

3.4 MET Series-Analysis Processing

The MET Series-Analysis tool ingests the URMA gridded observations and the postprocessed WRF forecasts so that matched pairs of forecast and observed values for all the variables at each lead time can be processed over the 99-day period. Because the URMA data is on a CONUS domain with 2.5-km grid spacing, Series-Analysis regridded it to create domains with grids that matched the 1- and 3-km grids of the WRF output to achieve the grid matching necessary for computing the forecast-observation differences and error statistics. While the regridding successfully enabled the matching of grids and the computation of the error

statistics and scores, it also introduced some error into the results. Regridding from the native 2.5-km grid of the URMA gridded observations to the 1-km grid of the WRF necessarily resulted in the addition of "new" data at the grid points located between the URMA grid points. Likewise, when regridding from the URMA grid to the 3-km grid, WRF would likely result in some degradation cause by the downscaling from 2.5-km to 3-km, but the difference of only 0.5 km in the grid spacing would lead to only a modest amount of error compared to upscaling from a 2.5-km to a 1-km grid. The differences in the impacts of the upscaling and the downscaling can be seen in Fig. 14. This figure shows graphics of the 2-m above ground level (AGL) temperature from the URMA observations on the 2.5-km grid (center image), on a 1-km grid (left image) and on a 3-km grid (right image) over the common domain of the 1-km WRF.

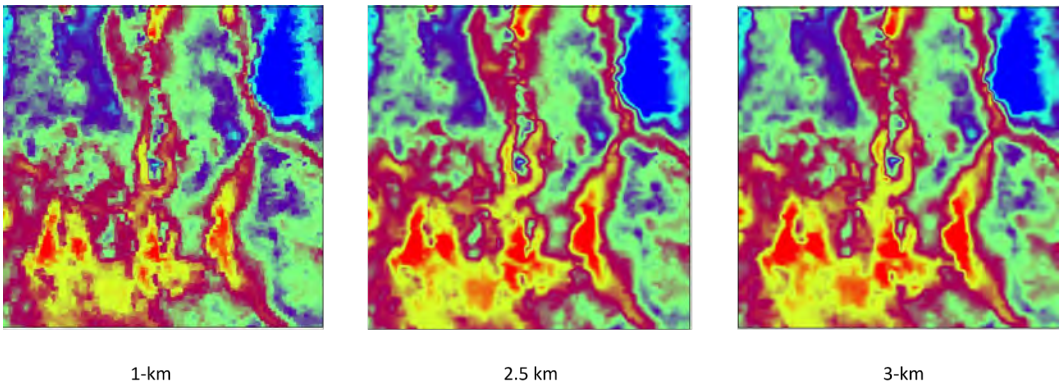

Fig. 14 2-m AGL temperatures from URMA on native 2.5 km-grid after regridding to a 1-km grid and a 3-km grid

Note the grainy appearance of the temperature features at the 1-km grid spacing compared to the smoother appearance at 2.5-km spacing. The details of the 1-km features imply higher resolution of the features that may mislead users into thinking that these features are more accurate representations of the true temperature features. Note the slight increase in smoothing seen in the 3-km spacing image compared to the 2.5-km image. The features seen at 3-km, though with less detail, are not significantly different from those at the native 2.5-km grid.

MET Series-Analysis applied the thresholds and other user-specified settings and computed the contingency-table statistics and skill scores for each grid square in both domains independently. In addition, the traditional continuous error statistics, ME and root-mean-squared error RMSE, were computed for each forecast-observation grid pair in each domain.

The output of MET Series-Analysis consists of a NetCDF formatted data file, which contains the scores and error statistics aggregated for each grid square over the 99-day period. The MET Plot-Data-Plane software loads the NetCDF files and

generates plots of the statistics (Halley Gotway et al. 2021). Plots of the statistics for both domains were generated for meteorological variables at the 2- and 10-m AGL and cloud cover variables listed in Table 1.

Table 1 Near-surface meteorological and cloud-cover variables and threshold values used for the assessment

Variable name/units	Abbreviation	Level (AGL)	Threshold values
Temperature (degrees Kelvin [K])	TMP	2 m	GE 280, GE 290
Dew-point temperature (degrees K)	DPT	2 m	GE 265, GE 280
U wind component (m/s)	UGRD	10 m	GE 0, GE 8
V wind component (m/s)	VGRD	10 m	GE 0, GE 8
Wind speed (m/s)	WIND	10 m	GE 3, GE 8
Specific humidity (kg/kg)	SPFH	2 m	GE 0.004, GE 0.006

The Table 1 abbreviations for the variable names are used in lieu of the names throughout the remainder of this report. Threshold values and other logical statements used in this study are defined by the following acronyms:

- "greater than or equal to" logical statement (GE)
- "greater than" logical statement (GT)
- "less than or equal to" logical statement (LE)
- "less than" logical statement (LT)

The thresholds selected for this study were chosen to generate the scores and statistics over the entire domain with little or reduced instances of low observed rate and forecast rate to avoid the inclusion of "missing" data when computing scores. This ensured maximum coverage of scores within each domain to optimize analysis with respect to terrain features. Thus, most of the thresholds used in this study did not have operational significance regarding their potential impact on aviation safety.

For the analysis of the output from MET Series-Analysis, plots of the CSI and FBIAS scores and the ME and RMSE error statistics aggregated over all model lead times for all 99 days and both WRF nested grids were generated to show the 2-D distribution over the domain. Overlays showing the terrain elevation contours were added to show the relationship between the scores and the terrain. The contour interval is 200 m. The readers are referred to the MET User's Guide for the formulas used for computing these statistics (Halley Gotway et al. 2021).

4. Analysis of Assessment Data

The graphics showing the CSI, FBIAS, ME, and RMSE for all variables and the analysis for the 1-km WRF domain are presented in Section 4.1, followed by those for the 3-km WRF domain in Section 4.2.

4.1 1-km WRF Domain Analysis of Data

4.1.1 TMP

The analysis for the 1-km WRF will make use of specific references to the geographic features shown in Fig. 2. Figure 15 shows the 2-m AGL TMP scores for GE 280 K, and Fig. 16 shows the scores for GE 290 K for the 99-day period. Figure 17 shows the TMP ME and RMSE statistics for the same period.

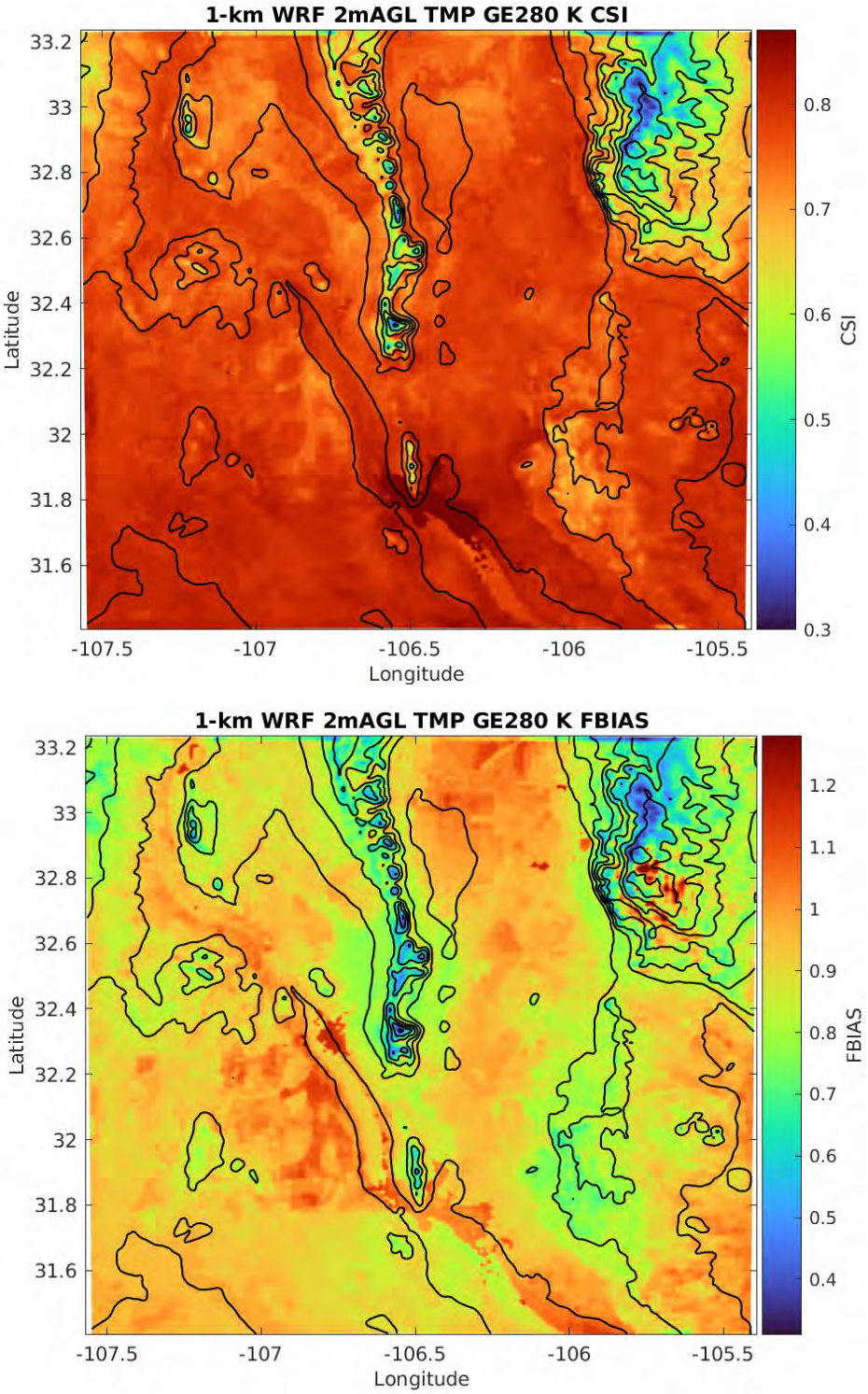

Fig. 15 CSI and FBIAS for 1-km WRF for TMP GE 280 K

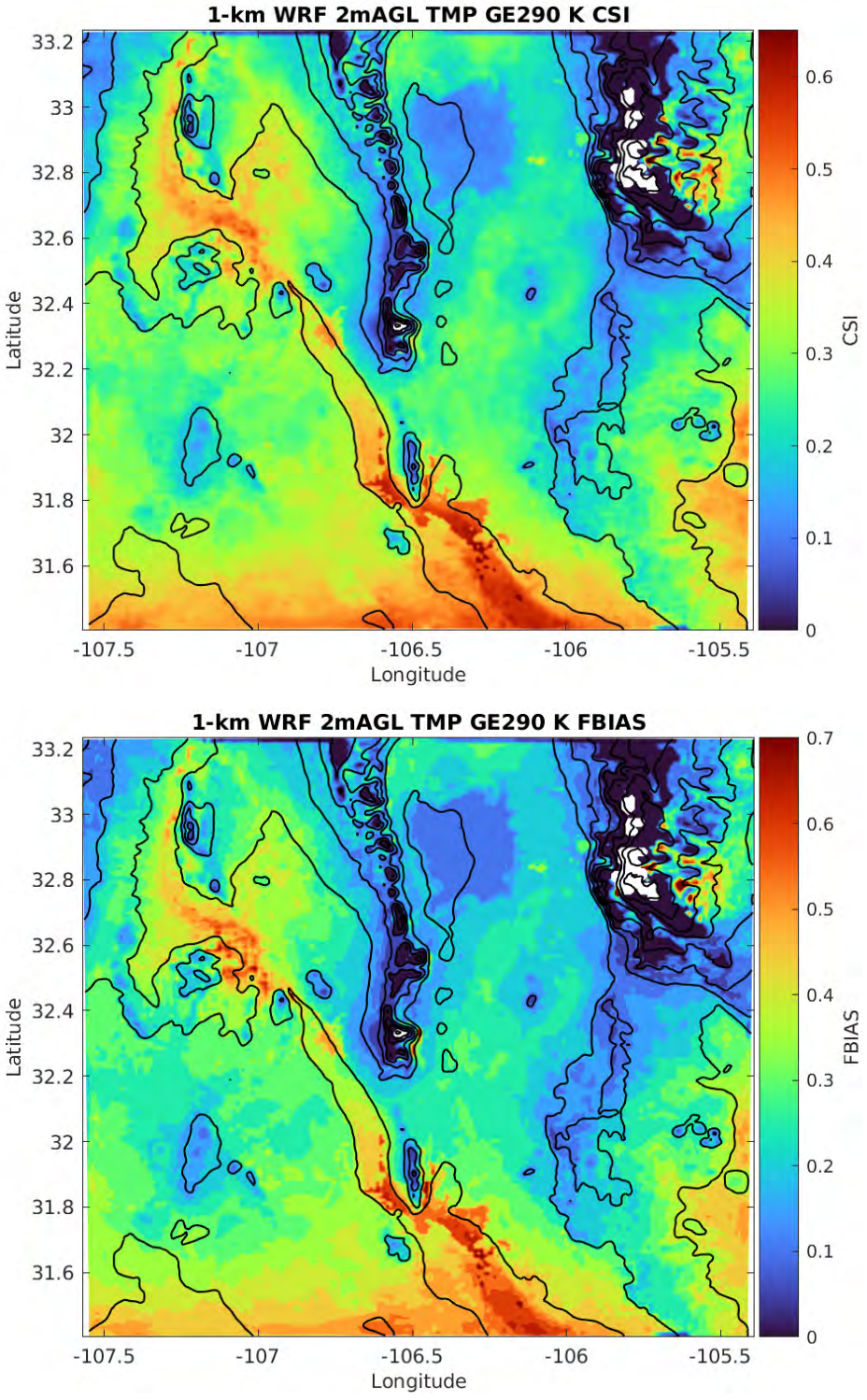

Fig. 16 CSI and FBIAS for 1-km WRF for TMP GE 290 K

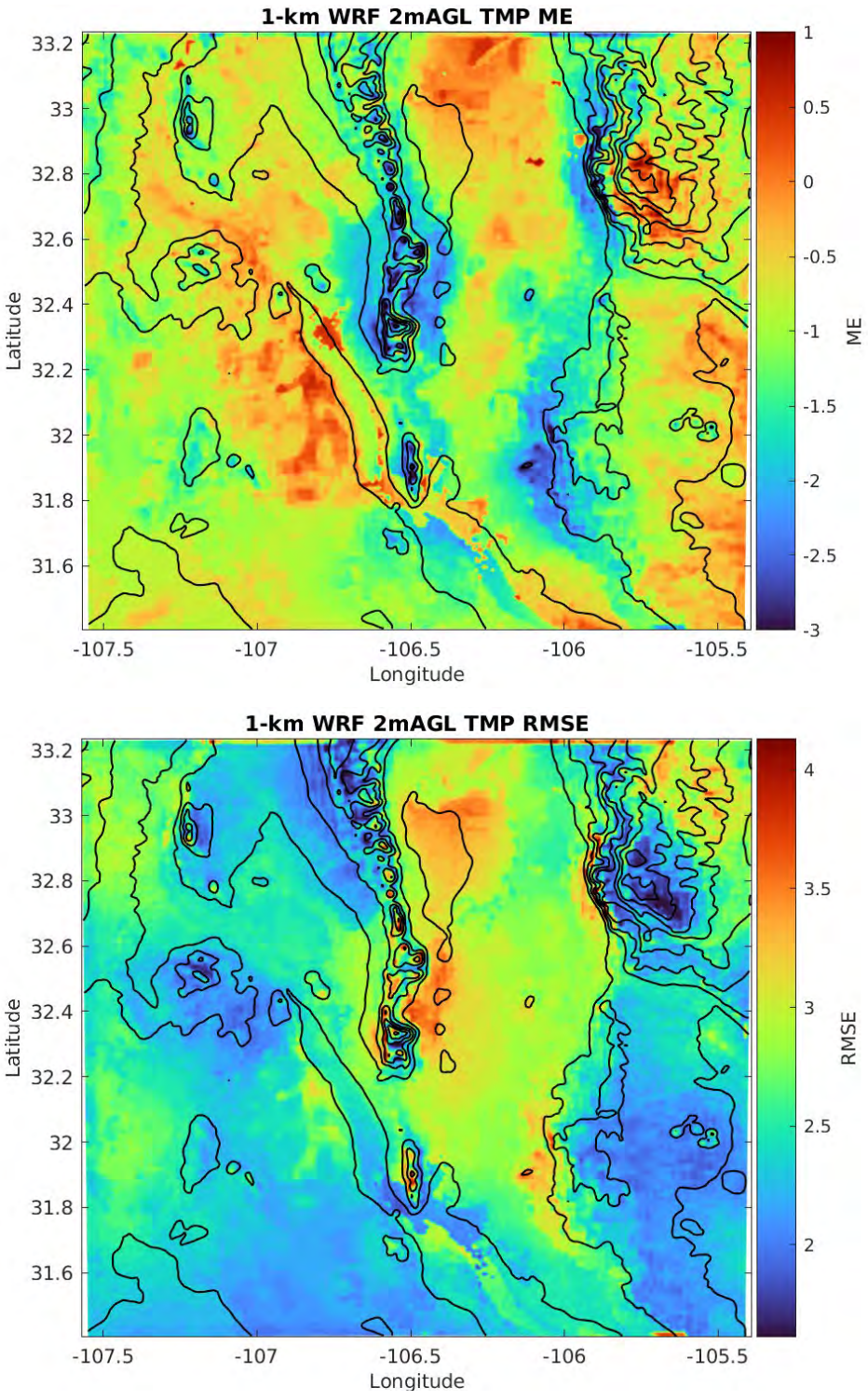

Fig. 17 ME and RMSE for 1-km WRF for TMP

From Fig. 15, the distribution of CSI for TMP GE 280 shows good skill (dark orange and red areas) in the valley and mesa areas and lower elevation mountains and lower skill in higher elevation mountainous areas. The pattern of FBIAS is somewhat in agreement with the distribution of areas having a value at or near 1.0 (good skill) in the lower elevation areas. In the higher elevation mountainous areas, generally there is a shift to lower FBIAS values indicating an under-forecast tendency, but there are exceptions to this in the Sacramento Mountains (northeast corner of the domain) where there are small, isolated areas with an over-forecasting tendency juxtaposed with small areas of under-forecasting bias. This mixed pattern may be due to the complexity of the land surface characteristics in this area.

However, there are areas with lower elevation that do not have good skill. These areas have the green and yellow colors indicating modest to slight under-forecast tendency, which occupy a large portion of the domain. There are objects with strongly contrasting forecast tendencies located in the area along and just west of the valley of the Rio Grande River extending from El Paso, Texas, north to just west of the Organ Mountains. The valley shows a slight under-forecast tendency (yellow color) while the higher elevation mesa area to the west has a slight over-forecast tendency (red color). There is an anomalous, small area of stronger positive bias indicated by a dark orange color located on the east side of the valley and west of the Organ Mountains at the point corresponding with the location of Las Cruces, New Mexico. Just to the east of this area over the Jornada Experimental Range (JER) is a contrasting area of stronger negative bias indicated by a light green color that runs north along the western slopes of the San Andres Mountains. This contrast in bias, though expected based on increasing elevation from west to east, may also be associated with a contrast in land surface characteristics between the urbanized area near Las Cruces and the semi-desert grassland of the JER. There is evidence of a similar contrast in the forecast bias between the area surrounding El Paso, Texas, and the nearby Franklin Mountains.

The geometric and linear nature of some of these features in southern New Mexico suggests that the Interstate Highway 10 corridor between Las Cruces and El Paso has a unique impact on the FBIAS that distinguishes it from the non-paved areas on either side of the corridor. The corridor has almost perfect FBIAS values (shown by a light red color) in contrast to the areas either side of the corridor, which show an under-forecast tendency (shown by a yellow color). Some additional small areas with anomalous stronger positive bias are in stark contrast to the surrounding areas of slight negative bias in the northwest corner of the domain at the location of the Elephant Butte Reservoir near Truth or Consequences, New Mexico. In the absence of a large contrast in elevation, it appears possible that this anomalous over-forecast tendency may be caused by the change in land surface type to water surface. Two

more anomalies of strong positive bias are in the northeastern corner of the domain at the locations of Alamogordo, New Mexico, and Holloman Air Force Base just west of the Sacramento Mountains. These are the small dark orange areas situated inside lighter orange and yellow areas of slightly negative bias, which are most likely associated with the change in land surface type going from urban to less developed land surface.

From Fig. 16, (TMP GE 290 K), the CSI scores overall are lower with the FBIAS values showing a decided tendency for under-forecasting TMP GE 290 K. As was the case for the lower threshold, the higher scores are in the areas of lower elevation, but in this case even lower elevation mountains and some of the valley areas share some of the lowest scores along with the higher elevation mountains. The highest elevation mountains have the poorest performance as indicated by the areas of dark blue color in the graphics. These are the San Andres Mountains in the northern central portion the domain and the Sacramento Mountains in the northeast corner of the domain. As is the case at the lower threshold, there are small, isolated areas in the Sacramento Mountains that have higher scores juxtaposed with the lowest scores, which appear to be associated with the complex land surface characteristics in this area. The white areas are the result of missing data where no events occurred either forecast or observed. One feature that stands out particularly well at this higher threshold is the area of blue color (light blue CSI approximately 0.15 and darker blue FBIAS approximately 0.1) situated within a larger area of slightly higher scores (green shades) in the north central part of the domain almost coinciding with the closed elevation contour located just east of the San Andres Mountains and west of the Sacramento Mountains. This area is the location of the WSNP where the dunes of white sand dominate the terrain very distinctly from the surrounding terrain. The sands are clearly visible as the white area in Fig. 3 and the outline of the sand dune area resembles the CSI and FBIAS objects described previously. This illustrates the impact of land surface characteristics on model performance. The areas of strong contrast in the scores between Las Cruces, New Mexico, and the JER and near El Paso, Texas, as well as the other small bias anomalies, possibly attributable to contrasting land surface characteristics, are also evident at this higher threshold value.

The distribution of ME and RMSE is shown in Fig. 17. Where ME is positive (dark red and orange areas) the WRF is over-forecasting TMP and where ME is negative (light orange, yellow, green, purple, blue areas) the WRF is under-forecasting TMP. There is a mixed pattern of forecast tendency. The Organ and San Andres Mountains and surrounding lower elevations have a negative bias while some of the highest elevation areas of the Sacramento Mountains have a positive bias. For the lower elevation areas, the bias is the least in general, but the pattern is also

mixed. The areas of strong contrast in the scores between Las Cruces, New Mexico, and the JER and near El Paso, Texas, as well as the other small bias anomalies, possibly attributable to contrasting land surface characteristics, are also evident in the distribution of ME, which agrees with the previous analyses of FBIAS. For RMSE, the Tularosa basin stands out from the rest of the domain as a decidedly uniform area of lower forecast accuracy ranging from light green, yellow, to orange with a value of approximately 2.8 to 3.5 K with the higher values (orange colors) located along the eastern slopes of the San Andres Mountains at the southern end and farther north over an area coinciding almost exactly with the terrain dominated by the white sand dunes. The highest elevation areas in the Sacramento Mountains, San Andres Mountains, and other mountainous areas have best forecast accuracy corresponding with the lowest RMSE values (1.7–2.3 K) as indicated by the light and dark blue colors.

4.1.2 DPT

The graphics showing the CSI, FBIAS, ME, and RMSE for 2-m-AGL DPT for the 1-km WRF are presented in Figs. 18–20. Figure 18 shows the scores for DPT GE 265 K and Fig. 19 shows the scores for DPT GE 280 K for the 99-day period. Figure 20 shows the ME and RMSE statistics for the same period.

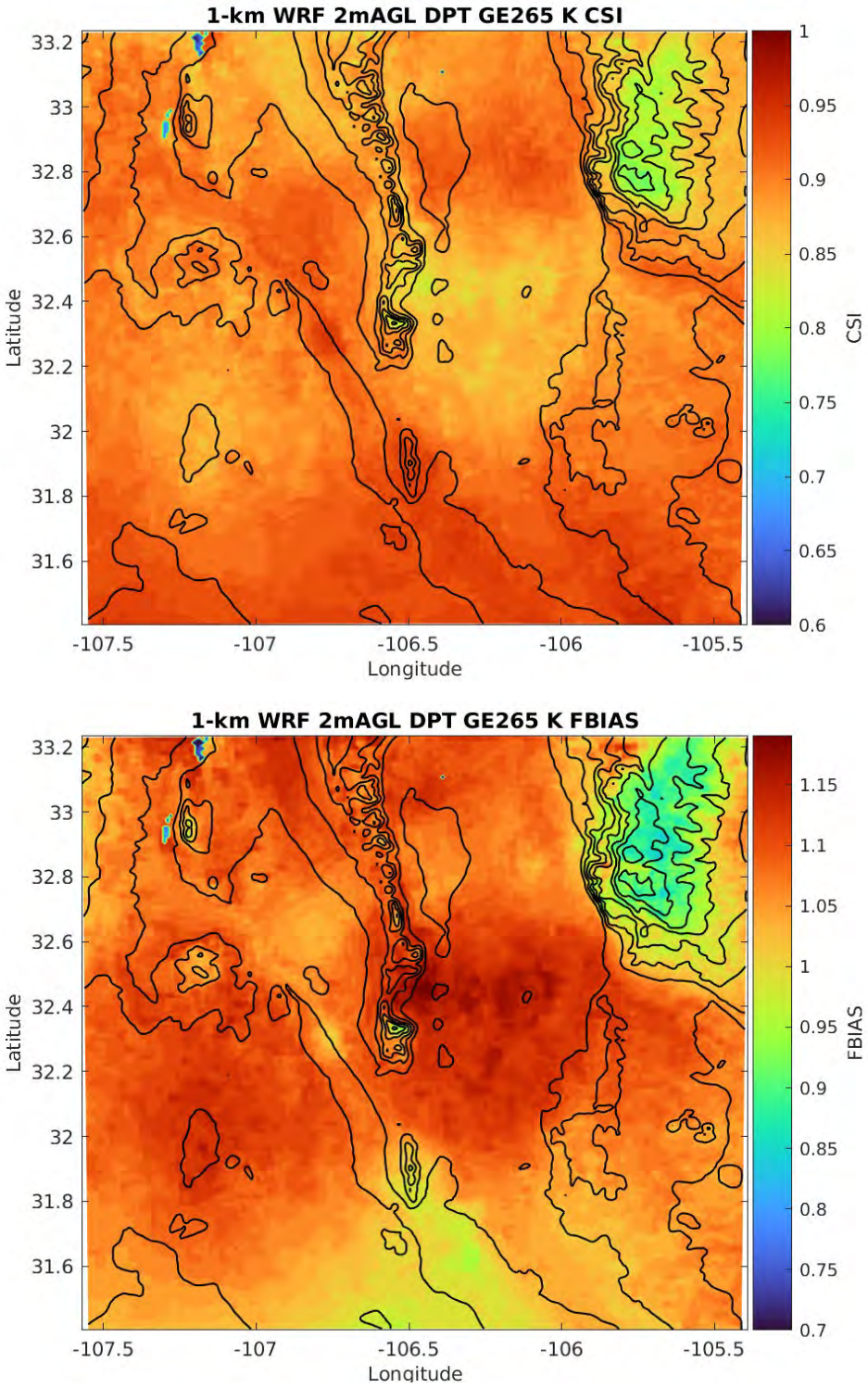

Fig. 18 CSI and FBIAS for 1-km WRF for DPT GE 265 K

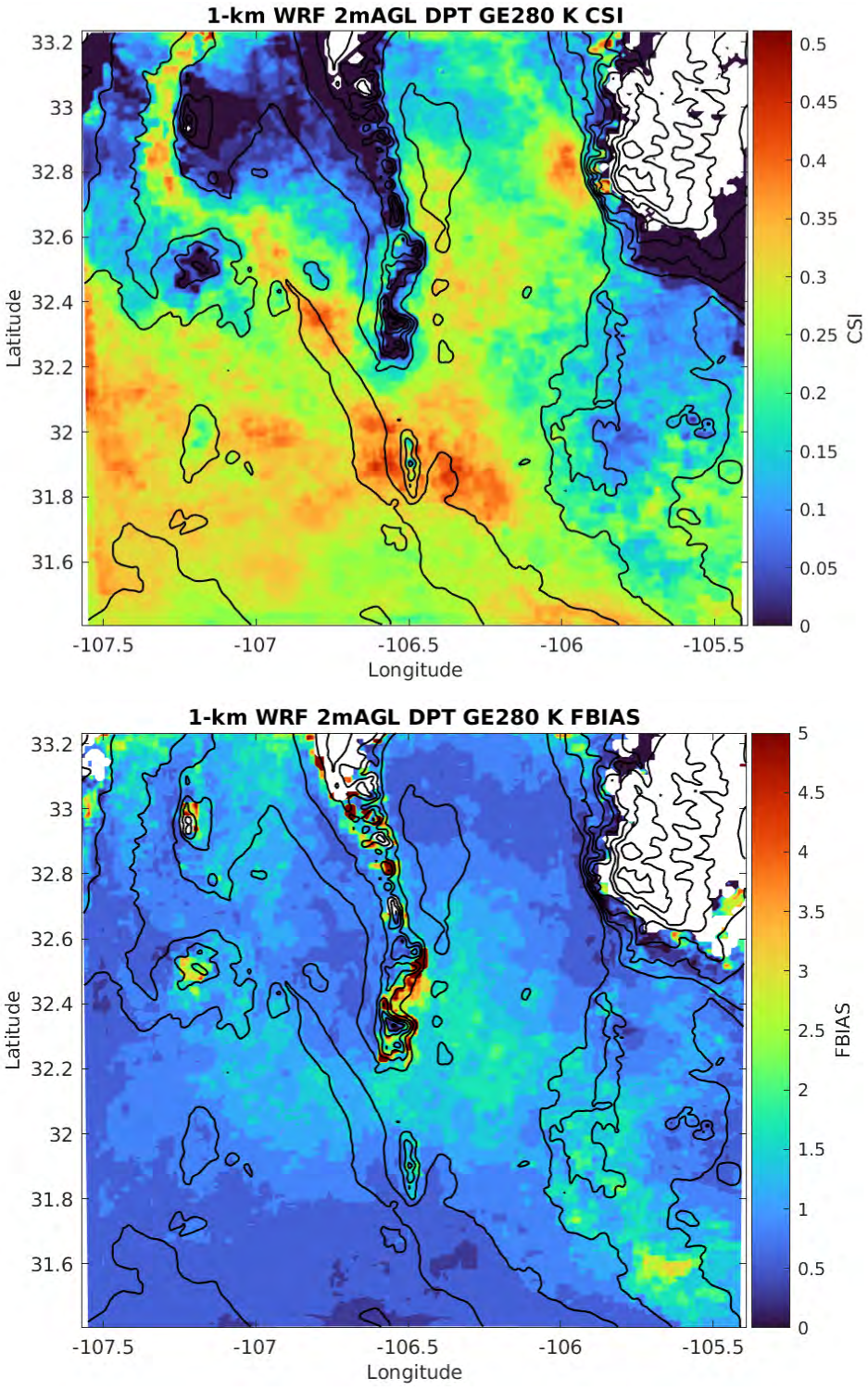

Fig. 19 CSI and FBIAS for 1-km WRF for DPT GE 280 K

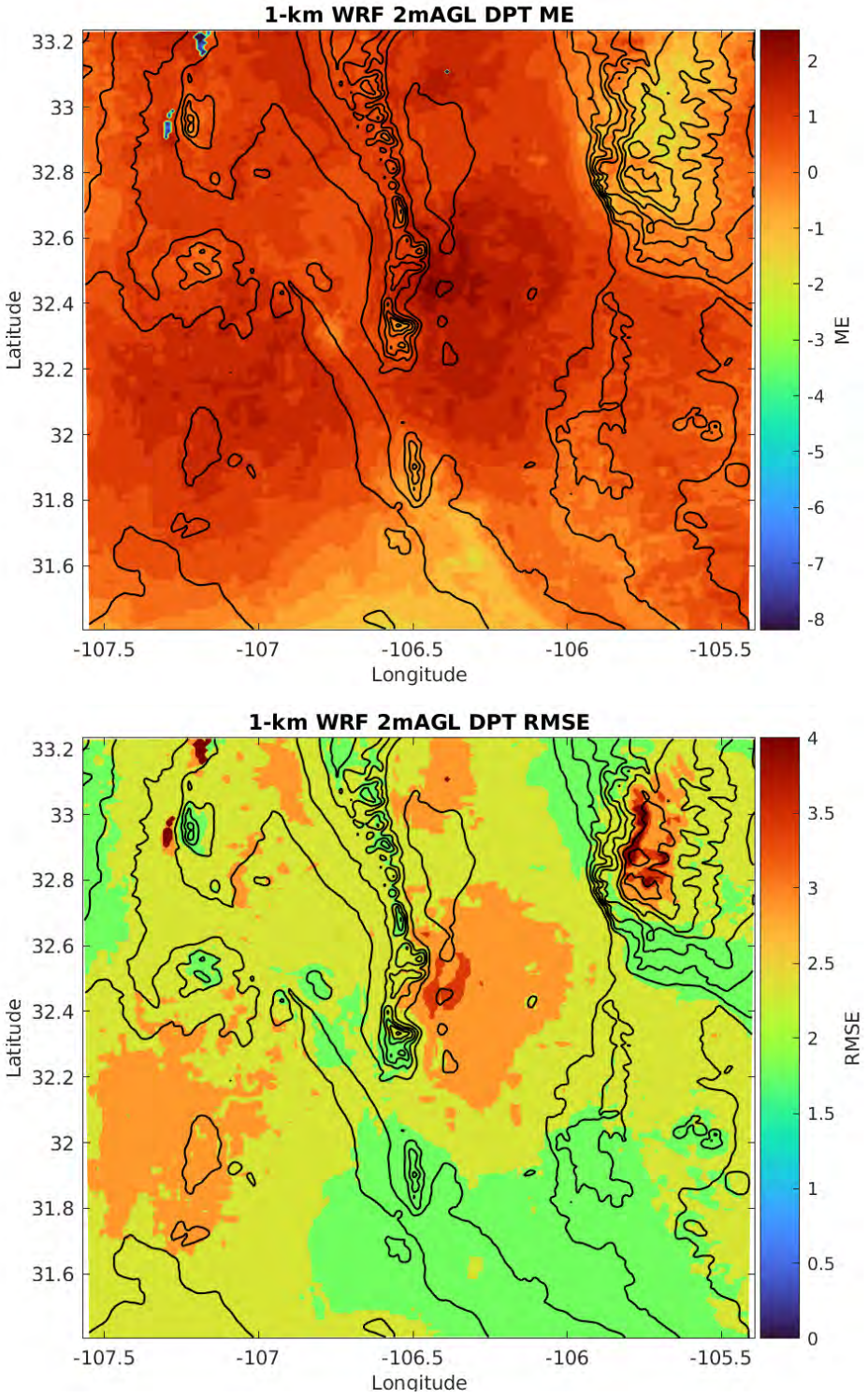

Fig. 20 ME and RMSE for 1-km WRF for DPT

From Fig. 18, the distribution of CSI for DPT GE 265 shows a general pattern of very good skill indicated by the orange areas in the valley and mesa areas and lower elevation mountains and lower skill, indicated by the light yellow and green color, in higher elevation mountainous areas such as the Sacramento Mountains. The best skill appears to be associated with the lowest elevation areas located in the southern third of the domain in Mexico and west Texas including the Franklin Mountains and the valley of the Rio Grande River.

The pattern of the FBIAS is generally in agreement with the distribution of yellow areas having a value at or near 1.0 in the lower elevation areas in Texas and Mexico. The areas with a negative forecast bias (FBIAS LT 1) are confined to the Sacramento Mountains and the highest peaks of the Organ Mountains. There are only weak features associated with contrasts in bias indicative of sharp contrasts in land surface characteristics at some of the same locations identified in the FBIAS for TMP. Three FBIAS objects contrast strongly in color from the surrounding areas at the Elephant Butte Reservoir near the northwest corner of the domain, Caballo Lake farther to the south in the northwest corner and another very small object located just north of the White Sands dune fields in the Tularosa Basin associated with Big Salt Lake. These lakes have a distinct under forecast bias.

From Fig. 19, (DPT GE 280 K), the CSI scores overall are lower compared to the lower threshold. Notably, the areas with higher skill indicated by the dark orange color are located near the urbanized areas of Las Cruces, New Mexico, El Paso, Texas, and Alamogordo, New Mexico, but the scores for these areas are not as good as those for the lower threshold. The areas with the least skill are the highest elevation mountains.

The FBIAS scores show large areas of near-perfect FBIAS values near 1.0 indicated by the lighter blue color over large areas of the domain generally associated with the valley and mesa areas and lower elevation mountains. The areas with an under-forecast tendency indicated by the darker blue color are in lower elevation areas in the Tularosa Basin between the WSNP and the Sacramento Mountains and in the low elevation areas of western New Mexico, west Texas near and including the Rio Grande Valley and in Mexico. The highest elevation areas of the Sacramento Mountains have missing data as indicated by the white color. The missing data at this location and in the extreme northern part of the San Andres Mountains was caused by the non-occurrence of events in the forecast and observations where the threshold was exceeded. There are no valid scores for CSI and FBIAS in these two areas. In the San Andres Mountains, outside of the extreme northern part with missing data, there are small, isolated areas with stronger over-forecast tendency, which may be associated with small-scale differences in land surface characteristics in the complex terrain in this area. There are similar, smaller instances of this

occurring in the Sierra De Las Uvas Mountains northwest of Las Cruces and the Caballo Mountains just south of Truth or Consequences, New Mexico, as well as in the lower elevation portions of the Sacramento Mountains. There are no features that show the strong contrast in bias indicative of sharp contrasts in land surface characteristics like those identified at the lower threshold and in the FBIAS for TMP.

The distribution of ME and RMSE is shown in Fig. 20. Where ME is positive (dark red and magenta areas) such as in the Tularosa Basin and western New Mexico the WRF is over-forecasting DPT. Where ME is negative (light yellow, green, and blue areas) the WRF is under-forecasting DPT. There is a mixed pattern of forecast tendency. The areas with lighter orange covering large portions of the domain have little to no bias. This includes the mesa areas and lower elevation mountains and even the San Andres Mountains. However, other low elevation areas have an over-forecast tendency such as the Tularosa Basin east of the southern end of the San Andres Mountains and southwestern New Mexico. Areas with an under-forecast tendency are the Rio Grande Valley in Texas and in the higher elevation areas of the Sacramento Mountains. There are weak signatures at some of the same locations where strong contrasts in TMP bias appear to be associated with contrasts in land surface characteristics. These are at Las Cruces, New Mexico, El Paso, Texas, Holloman Air Force Base (AFB), New Mexico, and the three lakes near Truth or Consequences, New Mexico, and north of the White Sands dune field.

The range of RMSE for DPT is very large with areas of 1.9 indicated by the green color in southern New Mexico near the Rio Grande River near Las Cruces, New Mexico, and near El Paso, Texas, to values in excess of 4.0 in the northwest corner of the domain north and south of Truth or Consequences, New Mexico, and in the Sacramento Mountains. Other than these areas, most of the domain is fairly uniform at approximately 2.2–2.6 K RMSE indicating no strong association with terrain elevation. There is a strong signature near the locations of Las Cruces where the low RMSE values in green are surrounded by higher values delineating a contrast in land surface characteristics from semi-desert grassland to urban. This suggests that the WRF errors are lowest over urban areas compared to rural areas. The isolated dark orange areas in the northwest corner of the domain show the adverse impact on RMSE values from the water surfaces of Elephant Butte Reservoir and Caballo Lake north and south of Truth or Consequences, New Mexico, respectively. There is a single pixel of high RMSE located in the north central part of the domain just north of the WSNP, which is the Big Salt Lake, a small lake occupying 743 acres in an area of playa lakes located on WSMR.

4.1.3 SPFH

The graphics showing the CSI, FBIAS, ME and RMSE for 2-m-AGL SPFH for the 1-km WRF for both thresholds are presented in Figs. 21–23. Figure 21 shows scores for SPFH GE .004 Kg/Kg and Fig. 22 shows scores for SPFH GE .006 Kg/Kg for the 99-day period. Figure 23 shows ME and RMSE statistics for the same period.

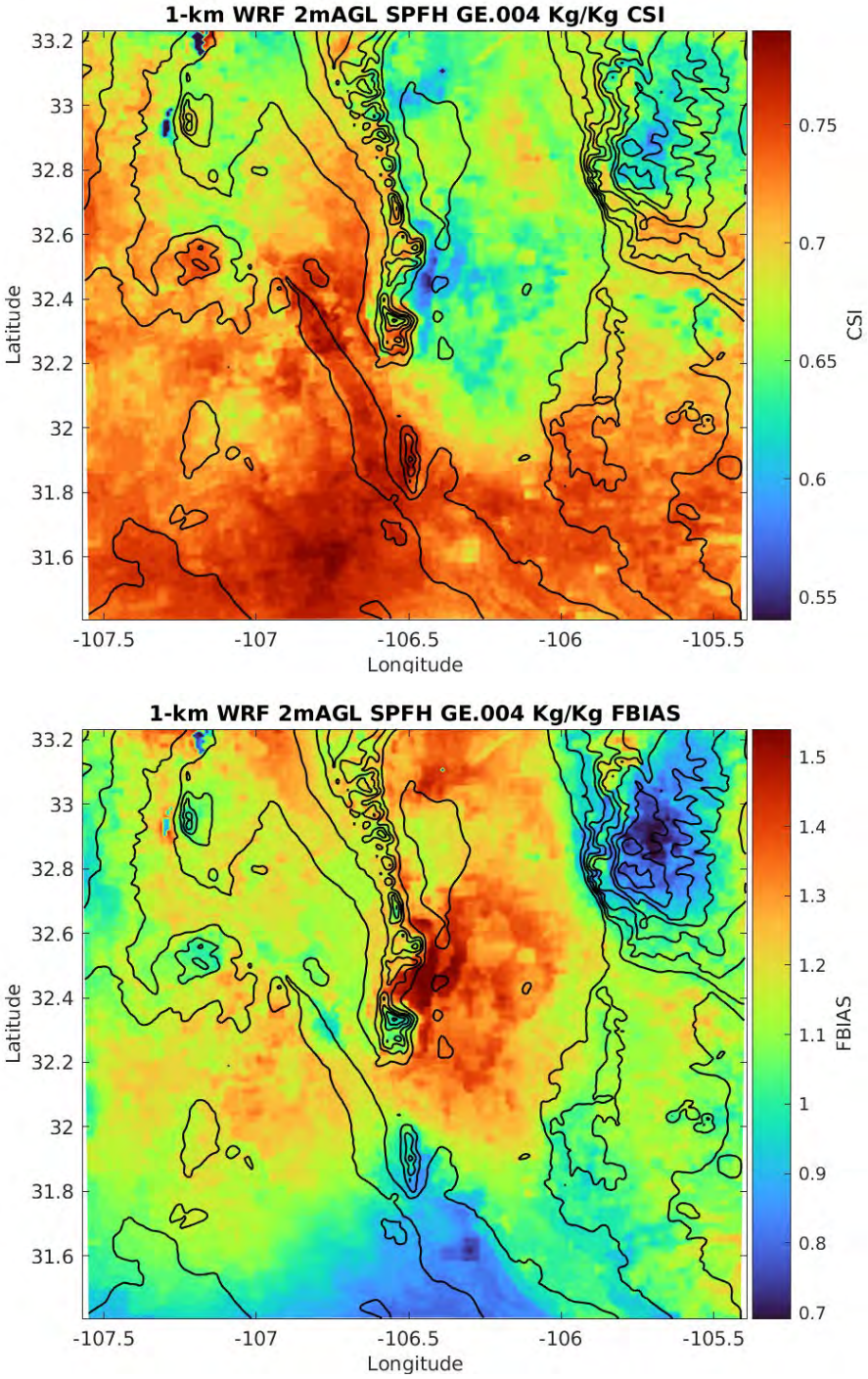

Fig. 21 CSI and FBIAS for 1-km WRF for SPFH GE .004 Kg/Kg

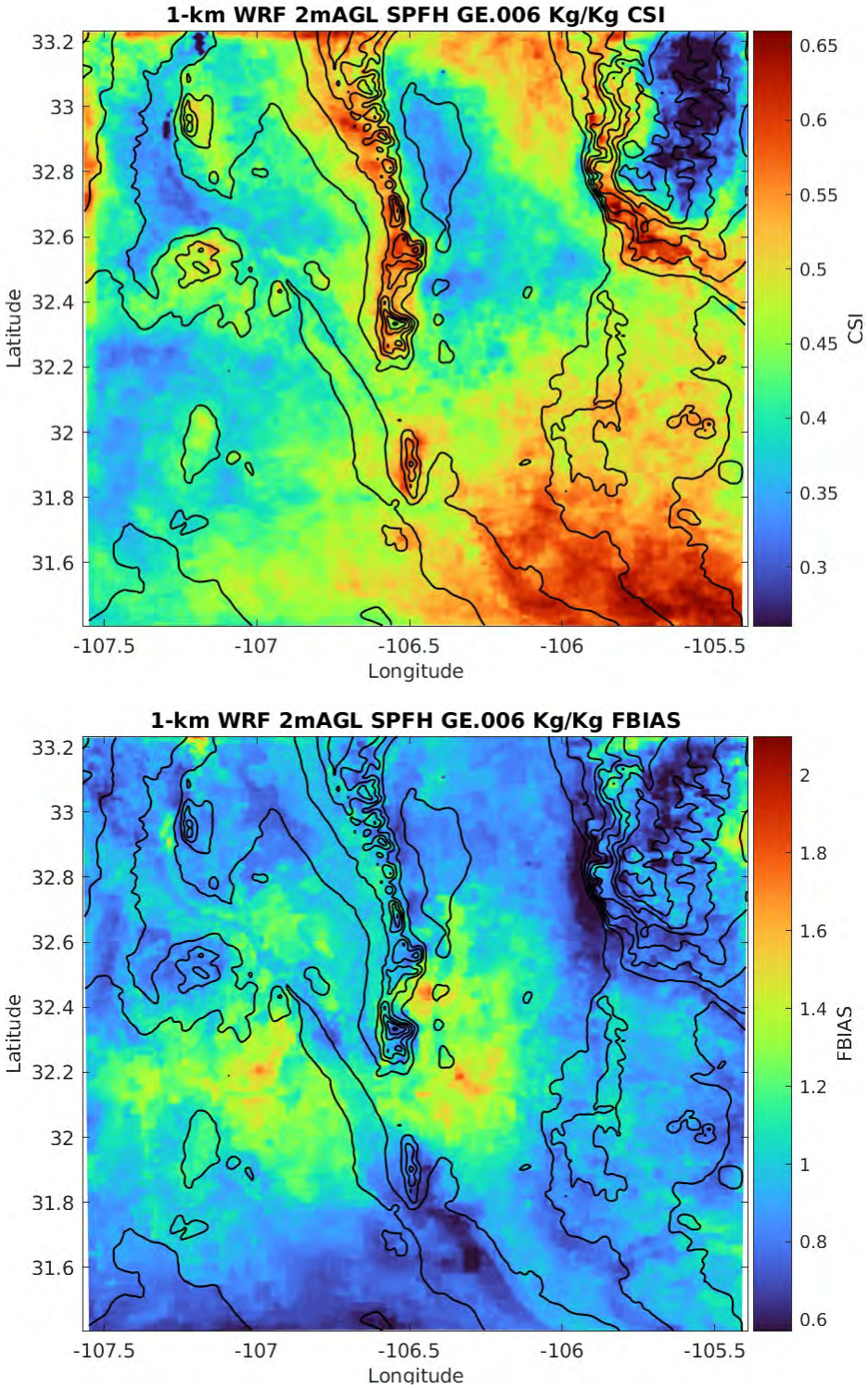

Fig. 22 CSI and FBIAS for 1-km WRF for SPFH GE .006 Kg/Kg

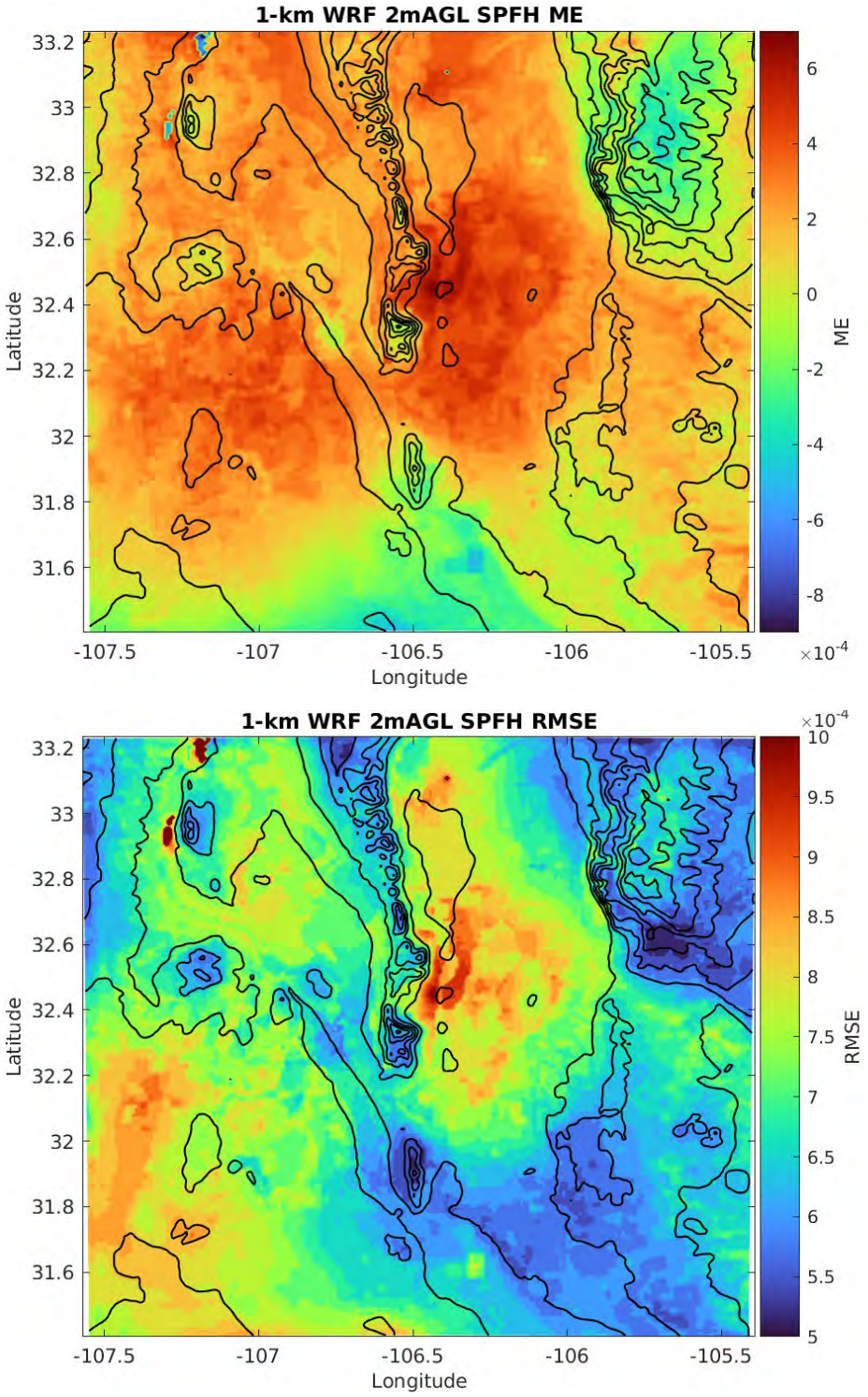

Fig. 23 ME and RMSE for 1-km WRF for SPFH

From Fig. 21, the distribution of CSI for SPFH GE .004 Kg/Kg shows fair skill (dark orange areas) in the lower elevation mesa and mountainous areas of Northern Mexico, west Texas and in the Rio Grande River valley extending north into New Mexico. The same level of skill is also seen for areas north of Las Cruces, along the western slopes of the San Andres Mountains, the Sierra de Las Uvas Mountains, and even including some parts of the Organ Mountains. In contrast, large portions of the Tularosa Basin, areas west of the northern San Andres Mountains, and the Sacramento Mountains have lower CSI scores as indicated by the blue, green, and yellow colors. The lowest scores occur in the highest elevations of the Sacramento Mountains as well as in the lowest elevations of the Tularosa Basin near the WSNP and just east of the southern end of the San Andres Mountains. There is no clear pattern in forecast skill associated with elevation.

The pattern of FBIAS is somewhat in agreement with the distribution of green and light blue areas having a value at or near 1.0 (good skill) in the same lower elevation areas in Mexico and west Texas but are not as extensive in areal coverage. There are some small, isolated areas of good skill elsewhere in the Organ Mountains, Sierra de Las Uvas Mountains, the Black Range Mountains, and parts of the Sacramento Mountains. The highest elevations of the Sacramento Mountains have stronger under-forecast bias (dark blue color) as does a small area in the southernmost part of the Rio Grande Valley extending west into Mexico. Large areas in the Tularosa Basin and in the western half of the domain have an over-forecast bias indicated by the dark yellow and orange colors. The strongest positive bias values (dark orange color) are located on the eastern slopes at the southern end of the San Andres Mountains. Some anomalous areas (green color) with very little bias (~1.0) contrast with the surrounding higher over-forecast bias values similar to those seen for TMP and to a lesser extent DPT at the locations of Las Cruces, New Mexico (green), Holloman AFB (green), New Mexico and possibly El Paso, Texas (light blue). Although, the light blue area is combined with a larger area of low bias to the south making it difficult to see the linear features and geometric shapes that are characteristic signatures of the impact of land surface features of urbanized areas on the bias.

From Fig. 22 (SPFH GE .006 Kg/Kg), the CSI scores overall are lower. The highest scores, though not considered good scores, are indicated by the dark orange color, and are in lower elevation areas of west Texas and the northern part of the Tularosa Basin and in some higher elevation areas such as the western foothills of the Sacramento Mountains, San Andres Mountains, Organ Mountains, Franklin Mountains, and the Black Range. The poorest scores, indicated by the blue colors, are found in the highest elevation parts of the Sacramento Mountains as well as the lowest elevation parts of the Tularosa Basin near WSNP and large areas in the lower

elevation areas in the west part of the domain. The distribution of FBIAS shows that large areas have values of approximately 1.0 (no bias) as indicated by a range of light green to light blue colors in both the lower and higher elevation areas across the entire domain.

Elsewhere, in the highest elevations and western slopes of the Sacramento Mountains, portions of northwest New Mexico, the southern lower Rio Grande Valley extending west into Mexico there is an under-forecast bias indicated by the dark blue colors. The areas of over-forecast bias shown by orange and yellow colors are in the Tularosa Basin east of the San Andres and Organ Mountains extending south across the lower part of WSMR and Otero Mesa and west to the Rio Grande Valley. Another area of positive bias lies west of the Rio Grande Valley across the mesa areas of southwestern New Mexico. Again, at this higher threshold, there is no clear and consistent association of forecast skill with elevation. There are similar areas with blue color having an under-forecast tendency, which contrast with the surrounding areas that have almost no bias (~1.0) like those seen for TMP and to a lesser extent DPT at the same cities as those for the lower threshold, but the signatures are less distinct than those at the lower threshold.

The distribution of ME and RMSE is shown in Fig. 23. Where ME is positive (dark and light orange and yellow areas) the WRF is over-forecasting SPFH and where ME is negative (greens and blue areas) the WRF is under-forecasting SPFH. Good values of ME (~0) are found in many areas across the domain ranging in elevation from the valleys and mesas to the lower elevation mountains and including parts of the San Andres Mountains. There is also a mixed pattern of under-forecast tendency occurring in the highest elevations of the Sacramento and Organ Mountains as well as the lower elevation areas in Texas along the Rio Grande Valley extending east into west Texas and west into Mexico. The areas with an over-forecast tendency seem to be confined to the lower elevation areas, particularly in the Tularosa Basin and in New Mexico west of the Rio Grande Valley. The same anomalous areas with strong contrasts with surrounding areas as those mentioned SPFH GE .006 are present at the locations of Las Cruces, New Mexico, (yellow), Holloman AFB (yellow), New Mexico, and possibly El Paso, Texas (light green). Although, the light green area is combined with a larger area of negative bias to the south making it difficult to see the linear features and geometric shapes that are characteristic signatures of the impact of urbanized areas on the bias.

The range of the RMSE values across the domain is large with areas of approximately 5 indicated by dark blue in the higher elevation areas of the domain as well as surrounding lower elevation areas such as the Rio Grande River valley and west Texas. Higher values of RMSE are distributed in the Tularosa Basin and in the lower elevation areas of New Mexico west of the San Andres Mountains. The

highest RMSE values are in the northwest corner of the domain and in the north central part of the domain in three isolated lake areas. These areas show the impact of water surfaces on the RMSE values. In the northwest corner, the water surfaces are Elephant Butte Reservoir and Caballo Lake, which were previously identified in the distribution of RMSE for DPT. The single pixel in the north central part of the domain north of WSNP is also attributable to the contrasting water surface of Big Salt Lake, which was also noted in the distribution of DPT RMSE.

4.1.4 WIND

The graphics showing the CSI, FBIAS, ME and RMSE for 10-m-AGL WIND for the 1-km WRF are presented in Figs. 24–26. Figure 24 shows the scores for WIND GE 3 m/s and Fig. 25 shows the scores for WIND GE 8 m/s for the 99-day period. Figure 26 shows the ME and RMSE statistics for the same period.

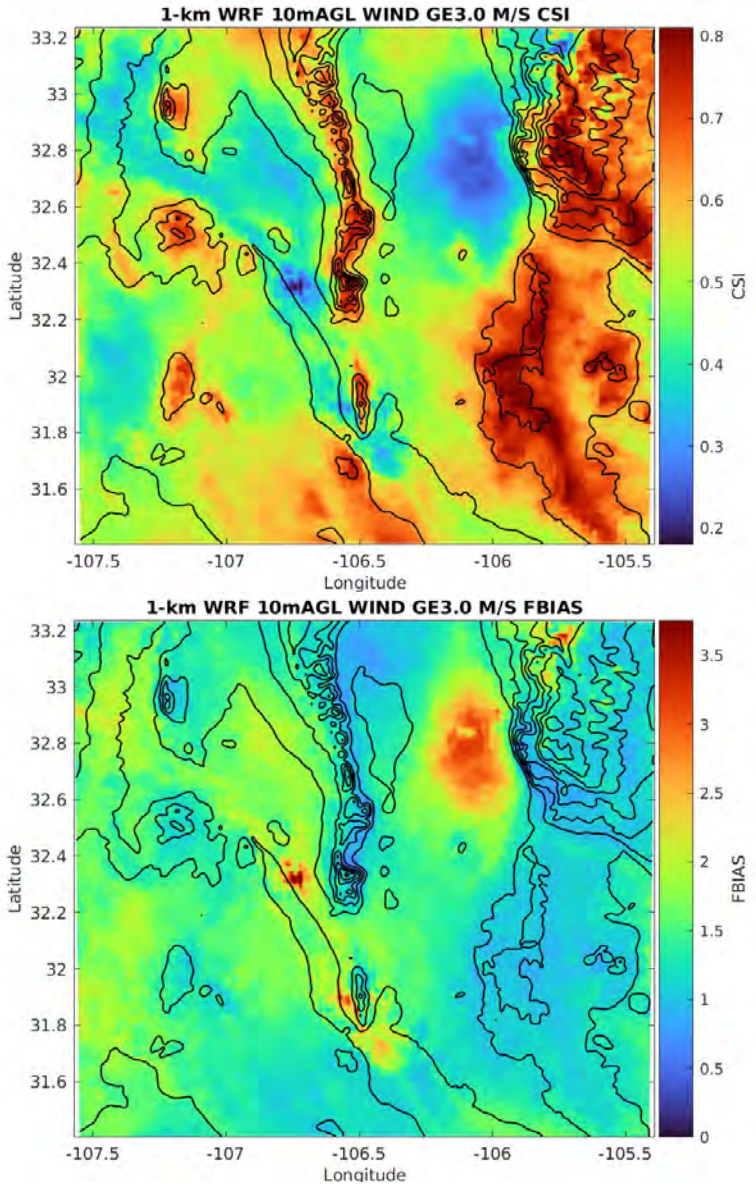

Fig. 24 CSI and FBIAS for 1-km WRF for WIND GE 3 m/s

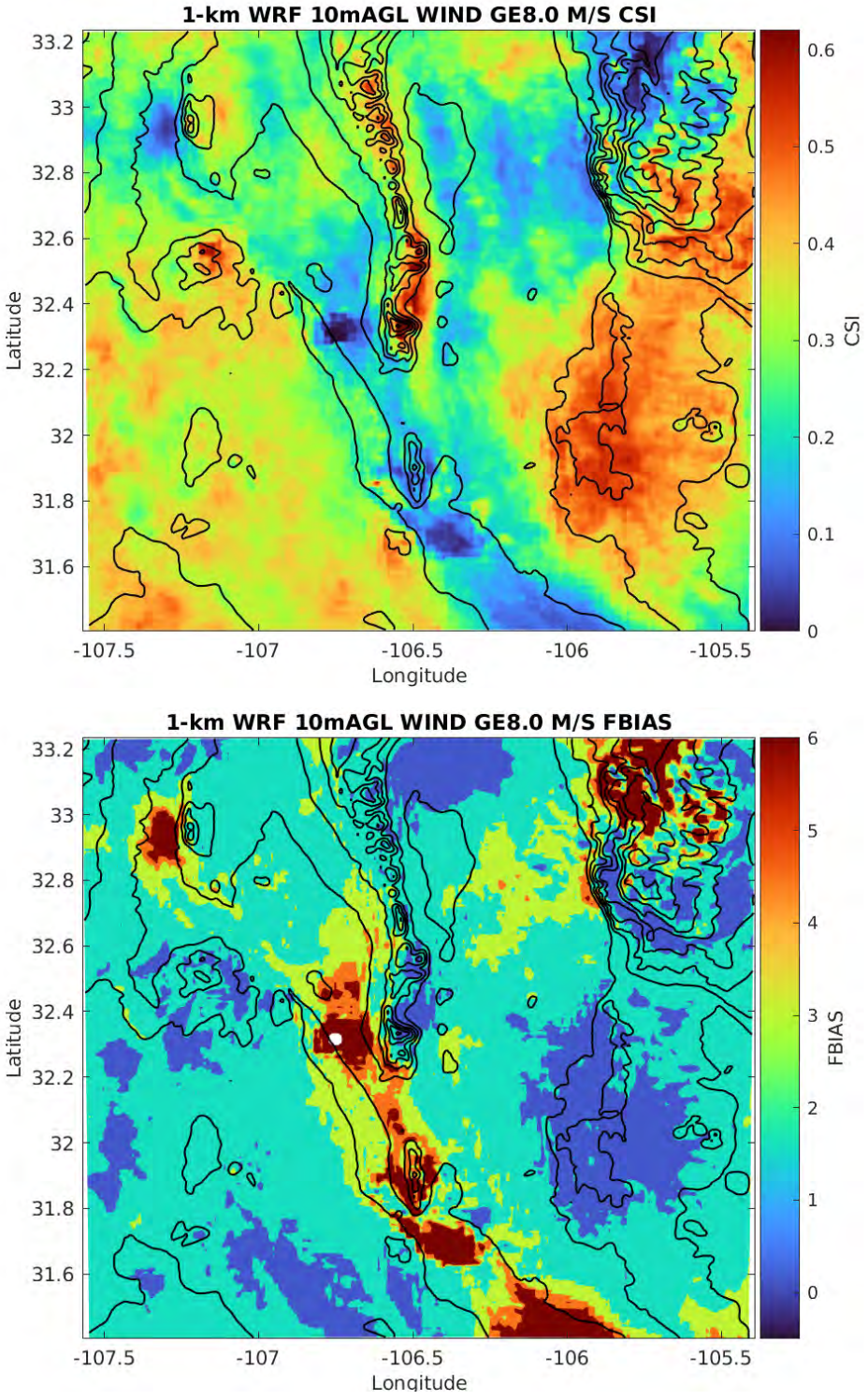

Fig. 25 CSI and FBIAS for 1-km WRF for WIND GE 8 m/s

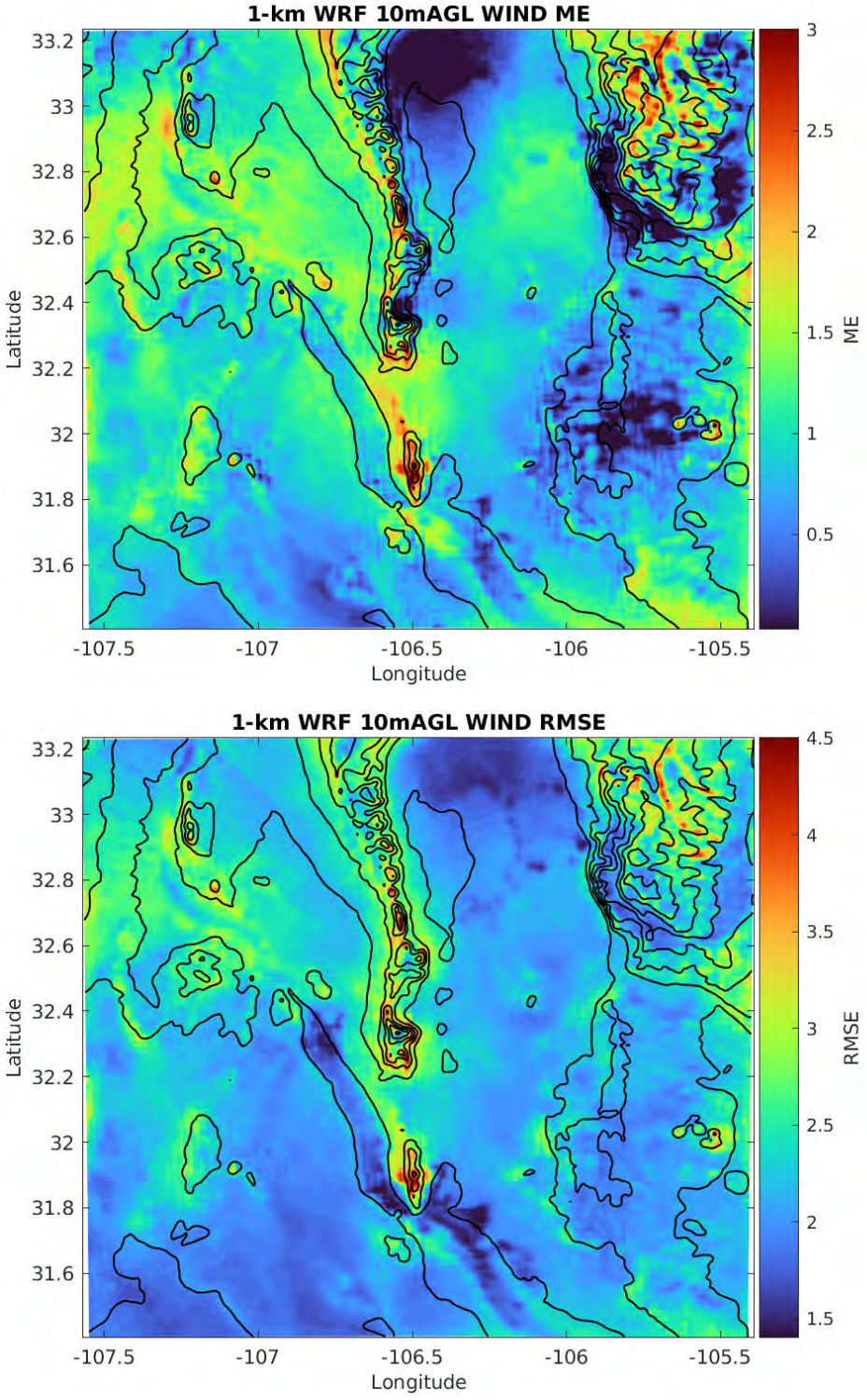

Fig. 26 ME and RMSE for 1-km WRF for WIND

From Fig. 24, the distribution of CSI for WIND GE 3.0 m/s shows very limited areas of fair skill (dark orange areas) in the highest elevation parts of the Sacramento Mountains, Organ Mountains, and Hueco Mountains east of El Paso, Texas. The scores appear to decrease with decreasing elevation in general over the domain with the lowest scores in the Tularosa Basin just west of the Sacramento Mountains indicated by the blue-colored area. The pattern of FBIAS is somewhat in agreement with the distribution of areas having a value at or near 1.0 (good skill with dark green color) in the higher elevation areas, although there are notable exceptions found in lower elevation areas of the Tularosa Basin and Mexico and New Mexico west of the Rio Grande River Valley. Areas with modest under-forecast tendency include small areas of blue color located north of the WSNP, the western foothills of the Sacramento Mountains, and the eastern foothills of the San Andres Mountains. There are no areas of strong under-forecast tendency. There are small, isolated areas of significant over-forecast bias at Las Cruces, New Mexico, and El Paso, Texas, which contrast sharply with the surrounding lesser bias indicating a possible impact from the urban land surface characteristics. A larger area with over-forecast tendency is the yellow and orange area located just west of the Sacramento Mountains. Most of the areas with an over-forecast tendency are lower in elevation with an exception in the higher elevation area north of the Sacramento Mountains.

Figure 25 shows the CSI and FBIAS for WIND GE 8.0 m/s. The CSI scores overall are lower than those for the previous threshold. The highest scores indicated by dark orange color, though not considered good scores, occur generally in the highest elevation areas such as the Hueco Mountains in southern New Mexico and west Texas, the southern part of the of the Sacramento Mountains, the San Andres Mountains, and the Organ Mountains. The lowest scores (blue and green colors) are generally in the lower elevation areas such as the Tularosa Basin and the Rio Grande Valley in New Mexico and Texas, but an exception to this is the northern portion of the Sacramento Mountains.

The FBIAS values over the domain show an over-forecast tendency ranging from the blue-colored areas, which have the least bias (approximately 1.0), to isolated areas of anomalously high values having green, yellow, and orange colors, which contrast sharply with surrounding areas of significantly lower positive bias. The prevailing bias values over the domain are generally between 1 and 10 with the anomalous area at Las Cruces, New Mexico, reaching a value of 78 and to a lesser extent the area near El Paso, Texas. These extreme values may be attributable to a sharp change in land surface characteristics associated with an urban setting.

The distribution of ME and RMSE is shown in Fig. 26. ME is positive over the entire domain, which indicates that the WRF is over-forecasting WIND to a modest

degree that varies depending on location. There is a mixed pattern of positive forecast bias. The higher elevation mountains such as the Sacramento Mountains, the San Andres Mountains, the Organ Mountains, and the Franklin Mountains have distributions of the bias characterized by small, isolated pockets of varying magnitude throughout (green, yellow, orange colors) with bias values generally between 1.0 and 3.0 m/s with the lowest values indicated by the green color. The lower elevation and mesa and valley areas generally have bias values between 0.5 and 1.5, but these values also occur in some higher elevation areas such as the Hueco Mountains east of El Paso, Texas, in the southeast corner of the domain straddling the border between Texas and New Mexico as indicated by the blue areas. Conversely, there are also lower elevation areas that have higher bias values such as the yellow-colored areas between El Paso and Las Cruces and large areas west of the San Andres Mountains and Rio Grande Valley in the western part of the domain.

For RMSE, large areas of the domain have reasonable values in the range between 1.4 and 3.0. The higher values, ranging from 3.0 to 5.1 indicated by the orange shades, are dispersed in spotty fashion over the highest elevation mountains such as the northern part of the Sacramento Mountains, the San Andres Mountains, the Caballo Mountains, and the Franklin Mountains. However, some of these same mountain areas also have small areas of lower RMSE values indicated by the blue areas. The lower elevation mesa and valley areas tend to have lower RMSE values indicated by the blue colors. Noteworthy among these areas is the Rio Grande River valley from Las Cruces, New Mexico, southeast through El Paso, Texas, into west Texas. The impacts on the ME errors from contrasts in land surface characteristics were not present for WIND for the urban locations identified previously for the other variables. For RMSE, however, the signatures for Las Cruces, New Mexico and El Paso, Texas, were distinguishable.

The scores and statistics for the remaining variables UGRD and VGRD all present similar patterns in terms of higher scores with lower thresholds and lower scores with higher thresholds. Since there are no operational thresholds for these variables, their scores will not be presented here but are presented in the Appendix.

4.1.5 Analysis of Impact of Land Surface Features on Scores and Error Statistics

The mixed pattern of the FBIAS for TMP GE 280, and TMP GE 290 in the Sacramento Mountains (Figs. 15 and 16) appears as small, isolated areas with over- and under-forecast tendencies. In this area, it is possible that small-scale differences in land surface characteristics may play a significant role in explaining the observed pattern. The patterns of variability in the vegetation type and the skin temperature

in Figs. 4 and 5 appear to be similar to the patterns present here. The size and granularity of the smaller vegetation type objects resembles some aspects of the patterns seen in the FBIAS, but the placement and shapes do not always match. Likewise, for skin temperature, the large object assigned the value range of 0 to –3 (purple) resembles the white area (0 FBIAS) in the FBIAS for TMP GE 290. More investigation is needed to confirm the linkage between the vegetation type and skin temperature and the FBIAS patterns in the Sacramento Mountains.

Another feature of interest that appears to be associated with land surface differences are the sand dunes of the WSNP. The signature of the sand dunes is very evident in the CSI and FBIAS for TMP GE 290 in Fig. 16 and in the RMSE in Fig. 17. The area of the dunes closely resembles the corresponding objects in the vegetation and soil type distributions in Figs. 4 and 7. The close association of these modeled objects with the same object in the FBIAS and RMSE fields suggests a close relationship that may have resulted in a net adverse impact on the skill of the WRF since the surrounding area shows better skill and less bias than the skill and bias within the area of the dunes. More investigation is needed to confirm this as there may be a potential source of bias from the URMA ground truth data if the observed TMP was analyzed with a positive bias relative to the WRF forecast over this area that would produce the same outcome.

The anomalous areas characterized by geometric and linear features with strong positive bias noted in the analysis of the FBIAS for TMP GE 280 in Fig. 15 at the locations of Las Cruces, New Mexico; El Paso, Texas; Alamogordo, New Mexico; Holloman AFB, New Mexico; and Elephant Butte Reservoir, New Mexico, also appear to be the result of the impact of land surface characteristics. These areas closely resemble areas in the vegetation type and skin temperature in Figs. 4 and 5. In the skin temperature field, there are light yellow objects situated at the locations of the previously mentioned cities except El Paso where the object has a darker yellow color. These objects are red in the vegetation type. In the same fields, there is a red-colored object at the location of the Elephant Butte Reservoir. Figure 6, which shows the land versus water characteristics, has water surface objects (green) associated with Elephant Butte Reservoir, Caballo Lake, and Big Salt Lake but no objects associated with urban areas. For the urban areas, the net impact of the land surface characteristics from vegetation type and skin temperature is an apparent adverse impact on the forecast TMP at the lower threshold as these areas showed a stronger over-forecast tendency relative to the surrounding areas. At the higher threshold, the net impact was favorable in terms of the bias. In terms of ME, the net effect is favorable as these areas showed less bias. For water surface area (lakes), the net impact from all four land surface characteristics also resulted in an adverse impact on the TMP forecast as this area had a distinct over-forecast bias relative to

the surrounding areas. More investigation will be needed to determine which surface type or combination of types has impacted the forecast or if there is bias coming from the URMA gridded observations or a possible combination of both bias sources.

For DPT GE 265 (Fig. 18) and ME (Fig. 20), there are some weak signatures in some of the urban and lake areas where there are strong contrasts in land use characteristics such as Las Cruces, New Mexico; El Paso, Texas; Holloman AFB, New Mexico; Caballo Lake, New Mexico; and Elephant Butte Reservoir, New Mexico. At the higher threshold (Fig. 19), the signatures were weaker. However, in the RMSE distribution (Fig. 20), there were stronger signatures associated with urban land surface at Las Cruces, New Mexico, and El Paso, Texas. There are very noteworthy signatures at the previously mentioned lakes as well as the Big Salt Lake in New Mexico. None of the distributions of scores and statistics for DPT have objects that flagged the location of the white sand dunes in the Tularosa Basin as was seen for TMP. The resemblance of the DPT objects associated with urban locations to the corresponding objects in the land surface types in Figs. 18 and 19 was not as close as that for TMP. On the other hand, the resemblance of the DPT objects for the lakes to the land surface types was closer.

For the urban areas, the net impact of the land surface characteristics from vegetation type and skin temperature is an apparent favorable impact on the forecast DPT as these areas showed a reduced over-forecast tendency and lower RMSE relative to the surrounding areas. For the lakes, the net impact from all four land surface characteristics was mixed. In terms of the FBIAS, the result was a favorable impact on the DPT forecast as these areas had a reduced over-forecast bias relative to the surrounding areas. In contrast, for ME, the result went from an over-forecast tendency (in the case of FBIAS) to a stronger, more adverse under-forecast tendency relative to the surrounding areas that have a prevailing over-forecast tendency. For RMSE, the result was an apparent unfavorable impact as the magnitude of the error increased significantly relative to the surrounding areas. More investigation will be needed to determine which surface type or combination of types has impacted the forecast or if there is bias coming from the URMA gridded observations or a possible combination of both bias sources.

For SPFH GE .004 (Fig. 21) and ME (Fig. 23), there are some signatures in some of the urban areas where there are strong contrasts in land use characteristics such as Las Cruces, New Mexico; El Paso, Texas; and Holloman AFB, New Mexico. At the higher threshold (Fig. 22), the same signatures are present but are weaker. In the RMSE distribution (Fig. 23), there were strong signatures associated with Elephant Butte Reservoir, Caballo Lake, and the Big Salt Lake in New Mexico. None of the scores and statistics for SPFH have objects that flagged the location of

the white sand dunes in the Tularosa Basin as was seen for TMP. The resemblance of the SPFH objects associated with urban locations to the corresponding objects in the vegetation type in Fig. 4 was not as close as that for TMP. On the other hand, the resemblance of the SPFH objects for the lakes to the land surface types was closer. At the lower threshold, for the urban areas, the net impact of the land surface characteristics from vegetation type and skin temperature is an apparent favorable impact on the forecast SPFH as these areas showed a reduced over-forecast tendency relative to the surrounding areas. At the higher threshold, the net impact was an apparent unfavorable one as these areas showed an increased under-forecast tendency relative to the surrounding areas. For the lakes, the net impact from all four land surface characteristics in terms of bias for both thresholds was adverse. The lakes had an increased over-forecast tendency relative to the surrounding areas. In terms of the RMSE, the result was an apparent adverse impact as the magnitude of the error increased significantly relative to the surrounding areas. More investigation will be needed to determine which surface type or combination of types has impacted the forecast or if there is bias coming from the URMA gridded observations or a possible combination of both bias sources.

For WIND, the signatures indicative of changes in land surface characteristics were present in the FBIAS, though not as distinct as with TMP. At both thresholds objects attributable to the sharp change from rural to urban land surface types were located at Las Cruces, New Mexico, El Paso, Texas, and Holloman AFB, New Mexico. However, at El Paso the signature was difficult to separate from the potential impact on FBIAS from the nearby Franklin Mountains. For two of the urban areas the net effect from land surface characteristics from vegetation type and skin temperature in terms of FBIAS is an apparent adverse impact on the forecast WIND as these areas showed an increased over-forecast tendency relative to the surrounding areas at both thresholds. However, at Holloman AFB, the impact was favorable. In this case, its location within a much larger area of relatively higher over-forecast tendency (not attributable to land surface changes) was distinct from the situations at Las Cruces and El Paso where the surrounding area had a relatively lower over-forecast tendency. For ME, the same objects were present, but much more difficult to discern compared to those in the RMSE field. The resemblance of the WIND objects, associated with urban locations, in the FBIAS and RMSE fields to the corresponding objects in the land surface types in Figs. 4 and 5 were not as close as that for TMP. In terms of the RMSE, the result for all three locations was an apparent favorable impact as the magnitude of the error decreased relative to the surrounding areas. More investigation will be needed to determine which surface type or combination of types has impacted the forecast or if there is bias coming from the URMA gridded observations or a possible combination of both bias sources. None of the scores and statistics for WIND have objects that flagged the

location of the white sand dunes in the Tularosa Basin or the three lake areas as was seen for TMP.

4.2 3-km WRF Domain Analysis of Data

The graphics in this section show the CSI, FBIAS, ME, and RMSE scores and statistics for all variables for both thresholds and the analysis for the 3-km WRF domain. The outline of the 1-km WRF domain is shown for reference. The analysis for the 3-km WRF makes use of specific references to the geographic features shown in Fig. 3.

4.2.1 TMP

The graphics showing the scores for TMP for both thresholds are presented in Figs. 27–29. Figure 27 shows the 2-m-AGL TMP scores for GE 280 K and Fig. 28 shows the scores for TMP GE 290 K for the 99-day period. Figure 29 shows the ME and RMSE statistics for the same period.

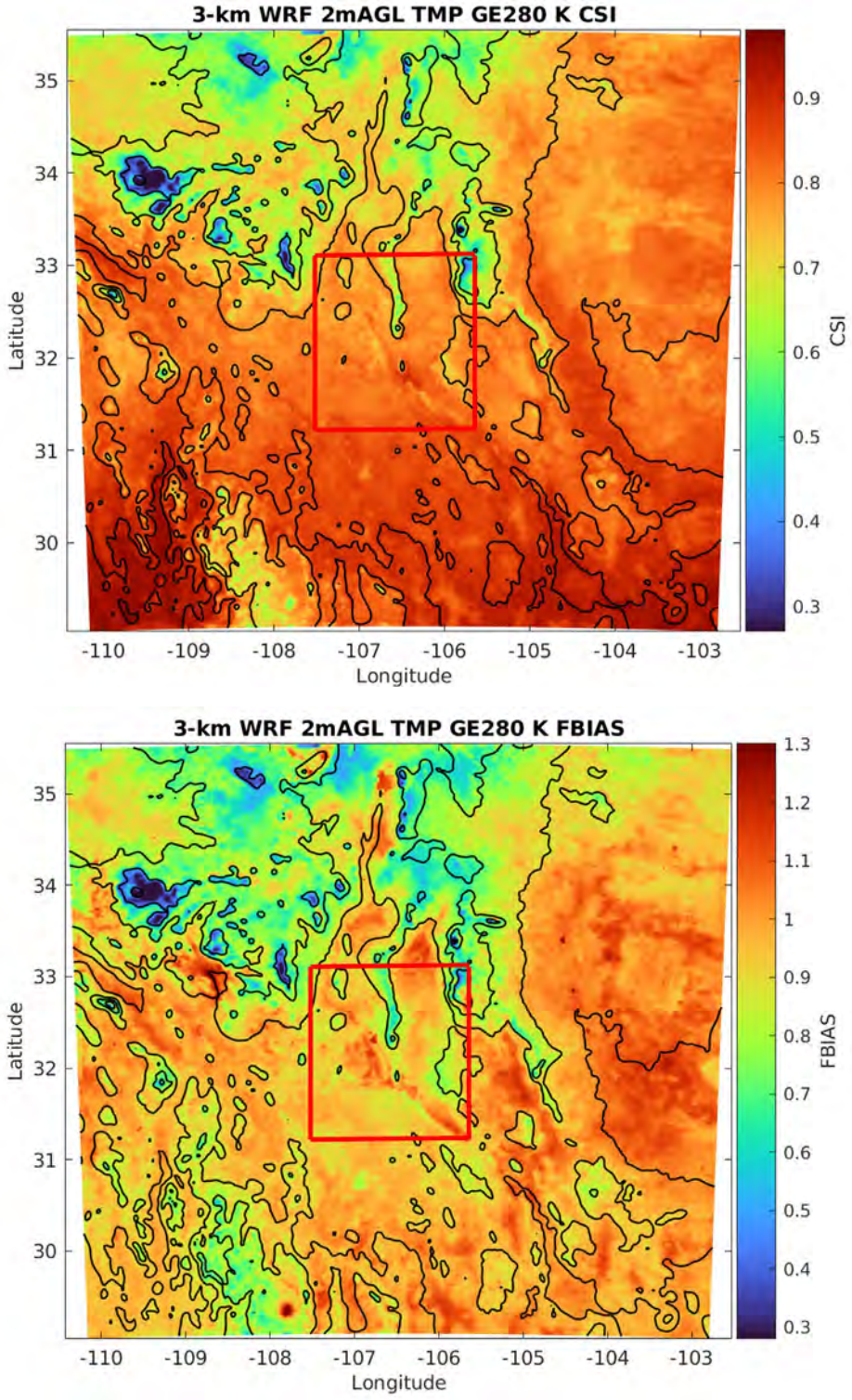

Fig. 27 CSI and FBIAS for 3-km WRF for TMP GE 280 K

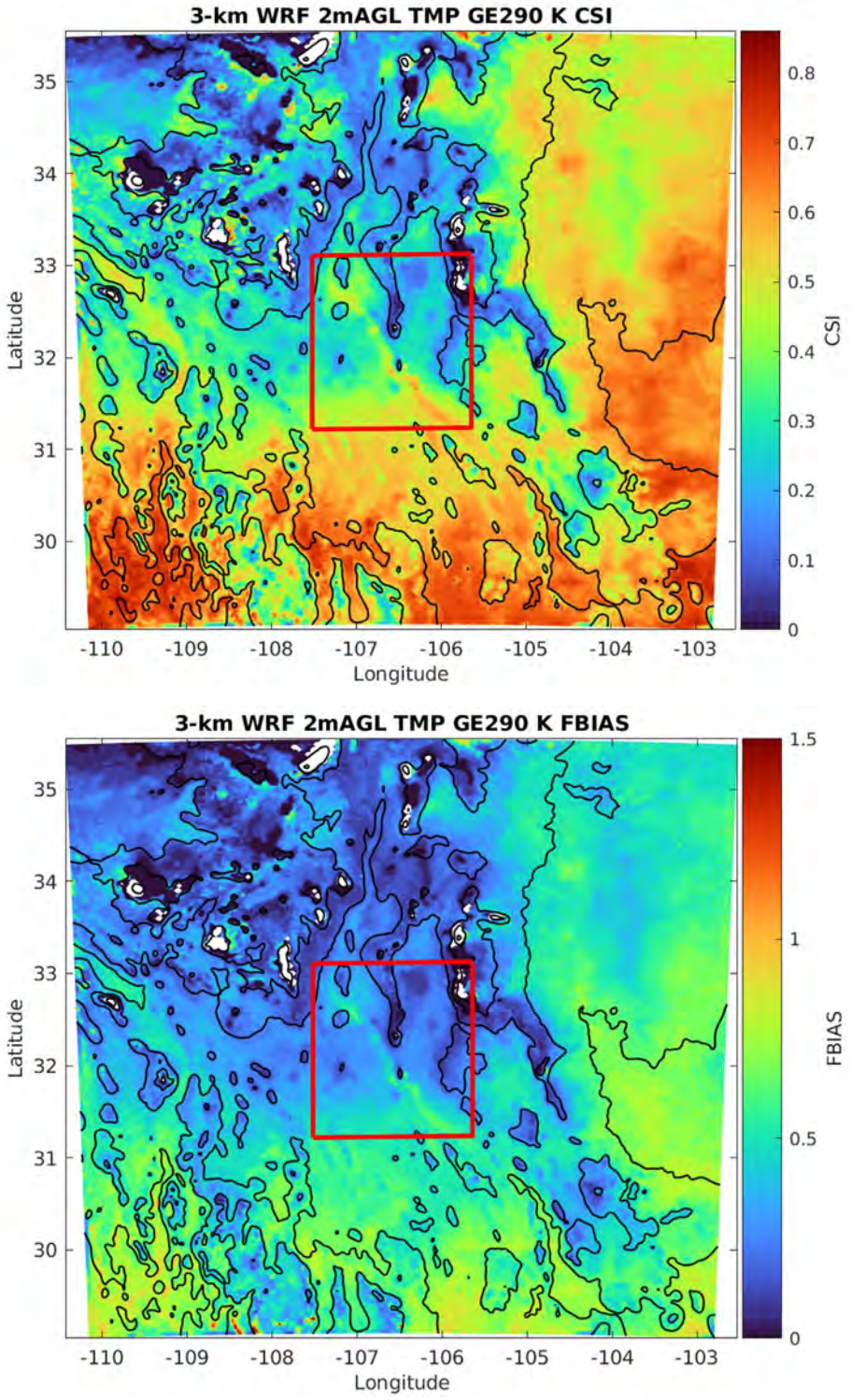

Fig. 28 CSI and FBIAS for 3-km WRF for TMP GE 290 K

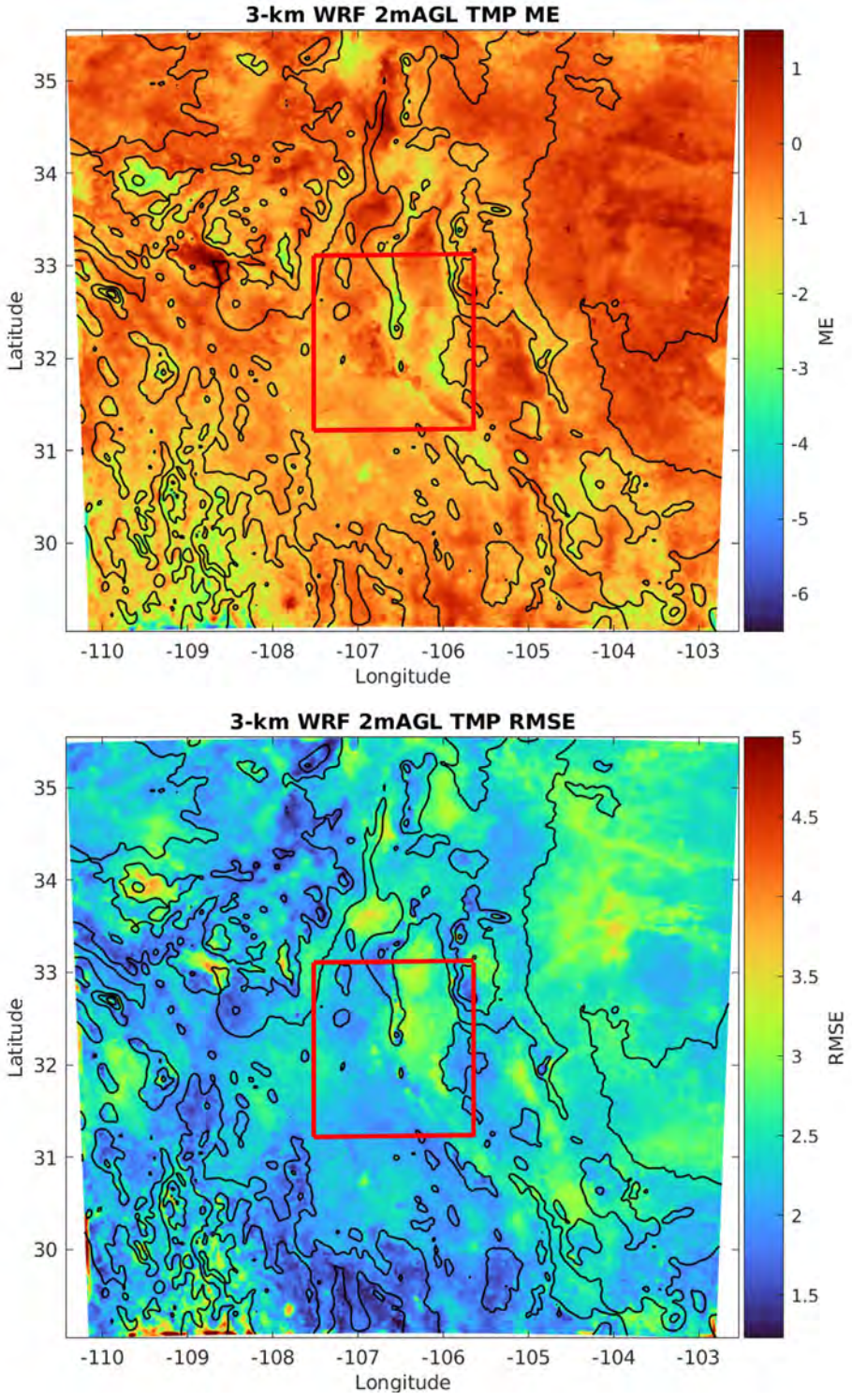

Fig. 29 ME and RMSE for 3-km WRF for TMP

From Fig. 27, the distribution of CSI for TMP GE 280 shows very good skill (orange shades) in the lower elevation areas in the southern part of the domain in Texas and Mexico . The eastern foothills of the Sacramento Mountains bordering on the Pecos River Valley in New Mexico as well as the Gila River Valley in Arizona also show good skill in these lower elevation areas. The areas with lower skill are primarily in the northwest corner of the domain where the higher elevations are, particularly in the mountainous areas of New Mexico and Arizona. Another area of lower scores is the Sierra Madre Mountains in northern Mexico. The lowest scores are in the highest elevations of the White Mountains in Arizona, and the Black Range and Sacramento Mountains in New Mexico. The pattern of FBIAS over the domain shows a similar trend with the minimum bias values at or near 1.0 (light orange) over the same lower elevation areas. There is under-forecast tendency (LT 1.0) indicated by the yellow, green, and blue colors, over areas with higher elevation across the domain with the mountainous areas having the largest values. The areas with significant over-forecast tendency (GT 1.0) indicated by the darkest orange color are very isolated and limited in areal extent. There are no obvious indications of linear or geometric patterns associated with contrasts in land surface characteristics like those found in the distribution of FBIAS for the 1-km WRF.

From Fig. 28, (TMP GE 290 K), the CSI scores overall are lower with the FBIAS values showing a clear tendency for under-forecasting TMP GE 290 K over most of the domain. As was the case for the lower threshold, the higher CSI scores (darkest orange color) are in the areas of lower elevation, but in this case even lower elevation mountains and some of the valley areas share some of the lowest scores along with the higher elevation mountains. The highest elevation mountains have the poorest performance as indicated by the areas of dark blue color. The most notable examples are the White Mountains, the Sacramento Mountains, and the Zuni Mountains. One feature that stands out particularly well at this higher threshold is the area of blue color (CSI approximately 0.25 and FBIAS approximately 0.1) situated within a larger area of slightly higher scores (green color for CSI and light blue for FBIAS) in the Tularosa Basin west of the Sacramento Mountains, which is the location of the WSNP and the white sands dune fields. This feature with unique land surface characteristics was also identified in the scores for the 1-km WRF and illustrates the locally adverse impact of the white sands on the CSI and bias. There are no other significant anomalies in the FBIAS and CSI fields that might be attributable to strong contrasts in land surface characteristics.

The distribution of TMP ME and RMSE is shown in Fig. 29. The areas of the lighter orange colors in the ME field indicate those areas with minimal bias that are generally located over lower elevation mesas and valleys across the domain, but a

notable exception is the low elevation areas located in Mexico west of the Sierra Madre Mountains where there is an under-forecast tendency indicated by the yellow and green colors. Numerous higher elevation mountain areas also have the same colors and negative bias tendency such as the White Mountains, the Black Range, the San Andres Mountains, the Hueco Mountains, the Franklin Mountains in El Paso, Texas, the Chiricahua Mountains, the Zuni Mountains, the Chisos Mountains, and the Davis Mountains. However, an obvious exception is the Sacramento Mountains that have minimal areas of green and yellow colors and have significant areas with the red colors of minimal bias. Areas of significant over-forecasting indicated by the darker red and magenta colors are limited to the foothills of the Mogollon Mountains near the Gila River Valley and the foothills of the Manzano Mountains near the Rio Grande River Valley.

For RMSE, there is a mixed pattern for the distribution of the larger magnitude values. The areas of green color (approximately 2.5 to 3.0 K) cover large areas of the lower elevation plains of eastern New Mexico extending south into west Texas as well as the Rio Grande River Valley. However, significant parts of the Sacramento Mountains, the Guadalupe Mountains, the Delaware Mountains, the Black Range, the Mogollon Mountains, and the Zuni Mountains also have the same green color. The highest errors indicated by the yellow color are in the highest elevations of the White Mountains as well as the Gila River Valley in New Mexico and near Truth or Consequences, New Mexico, in the Rio Grande Valley. The distribution of the lowest values of RMSE, indicated by the blue shades, is also mixed with some areas in high elevation mountains and others in lower elevation areas. There are no significant anomalies in the ME and RMSE fields that might be attributable to strong contrasts in land surface characteristics.

4.2.2 DPT

The graphics showing the CSI, FBIAS, ME, and RMSE for 2-m-AGL DPT for the 3-km WRF for both thresholds are presented in Figs. 30–32. Figure 30 shows the scores for DPT GE 265 K and Fig. 31 shows the scores for DPT GE 280 K for each model lead time for the 99-day period. Figure 32 shows the ME and RMSE statistics for the same period.

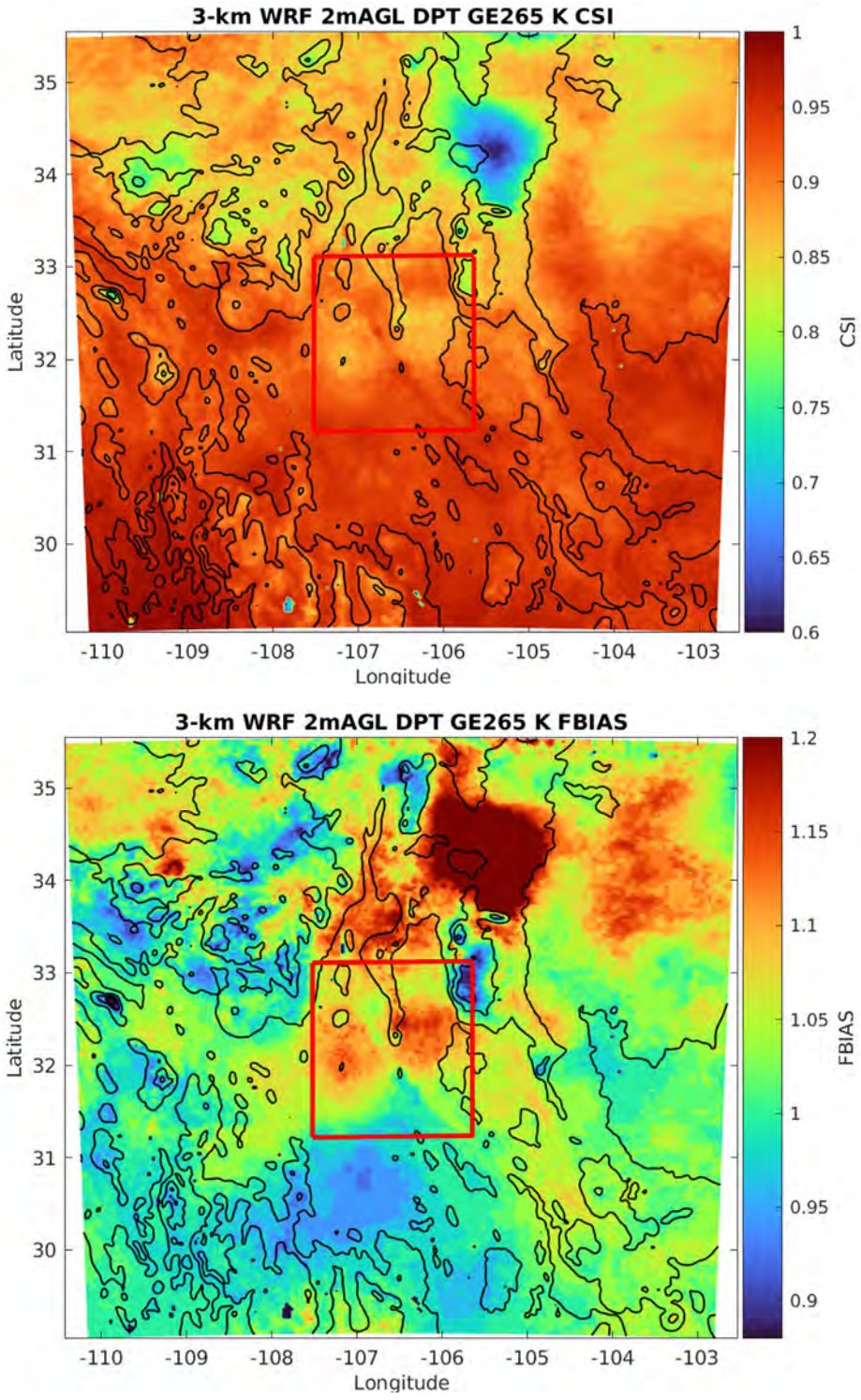

Fig. 30 CSI and FBIAS for 3-km WRF for DPT GE 265 K

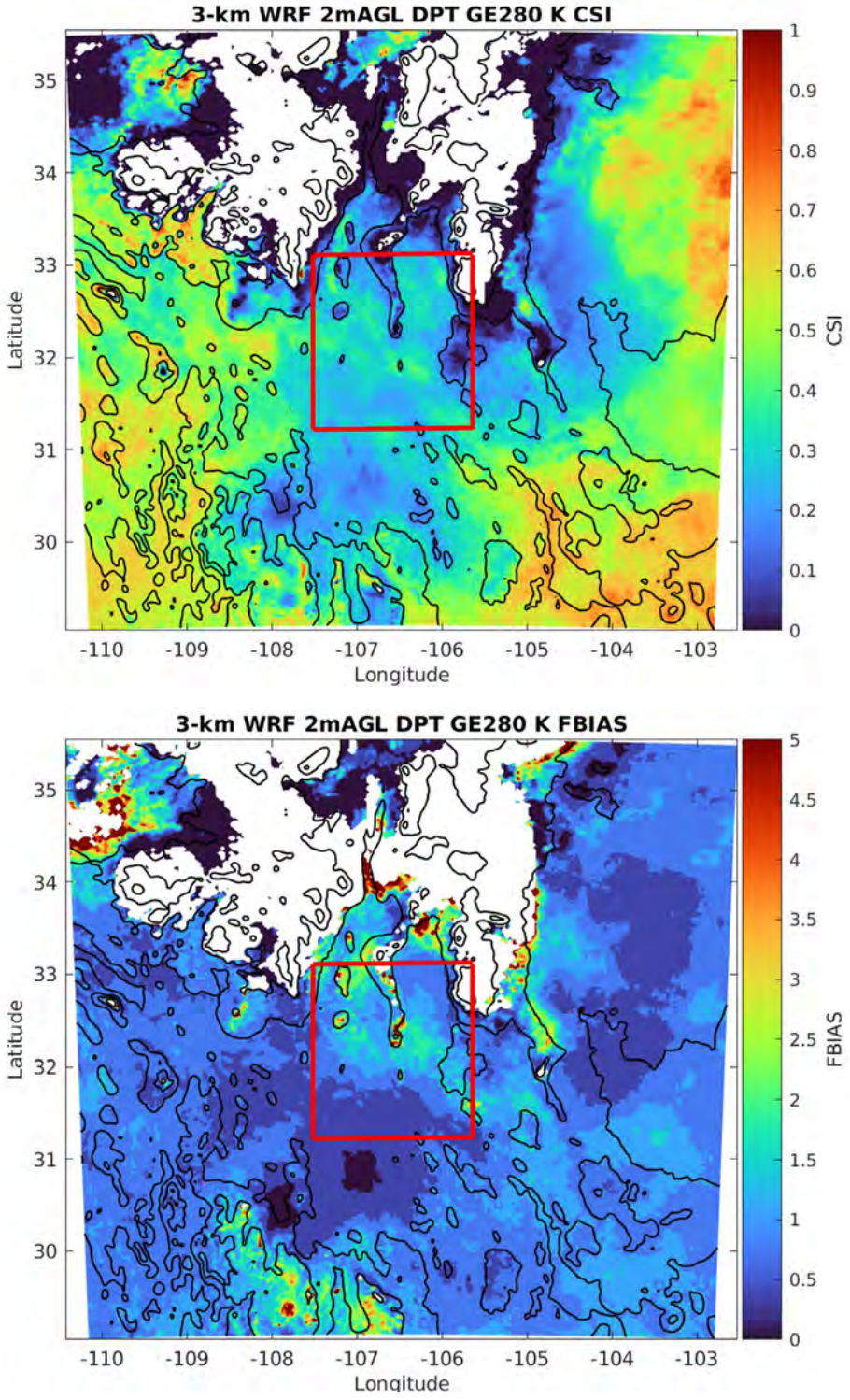

Fig. 31 CSI and FBIAS for 3-km WRF for DPT GE 280 K

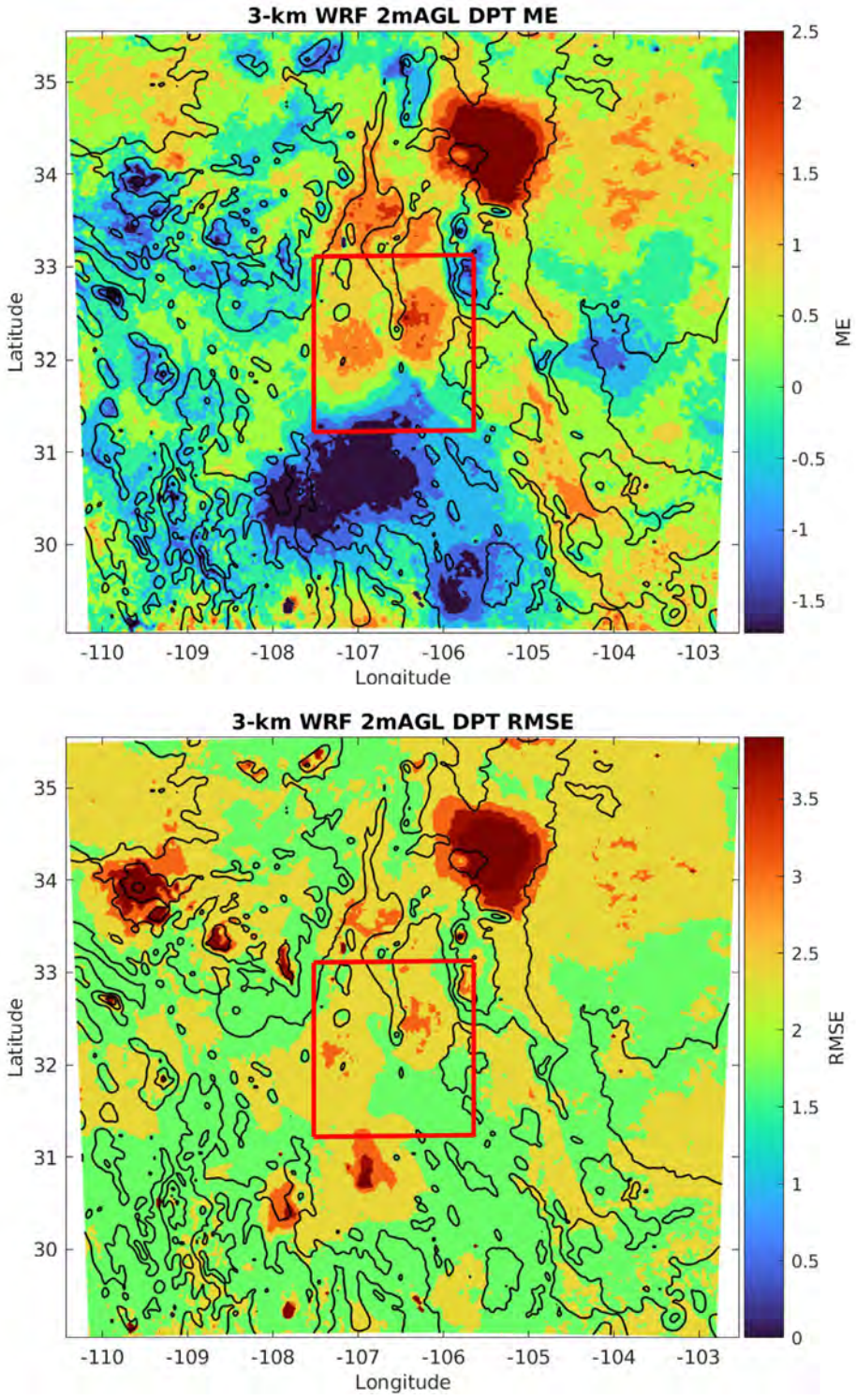

Fig. 32 ME and RMSE for 3-km WRF for DPT

From Fig. 30, the distribution of CSI for DPT GE 265 shows a general pattern of very good skill indicated by the orange shaded areas in the valley and mesa areas and lower elevation mountains in the southern two-thirds of the domain. The areas with lower skill are in the northern third of the domain as indicated by the yellow, green, purple, and gray colors in higher elevation mountainous areas such as the Sacramento Mountains, the White Mountains, the Mogollon Mountains, and the Black Range. The best skill appears to be associated with the lowest elevation areas located west of the Sierra Madre Mountains in Mexico. The lowest skill, indicated by the blue-colored object located north of the Capitan Mountains in the high elevation mesa area west of the Pecos River Valley.

The pattern of the FBIAS is generally in agreement with the distribution of yellow, green, and light orange areas having a value at or near 1.0 in the lower elevation areas in the southern two-thirds of the domain. The areas with a significant over-forecast bias, indicated by the dark orange color, are in the southern half of the Tularosa Basin, the mesa area west of the Rio Grande Valley in southern New Mexico, and in the Rio Grande Valley between the Magdalena Mountains and the Sacramento Mountains. A conspicuous object with the highest FBIAS values indicated by the darkest orange color is in the area north of the Capitan Mountains in the high elevation mesa area west of the Pecos River Valley. The lowest values of FBIAS, indicated by the dark blue color, are limited to some of the highest elevation mountainous areas including the Sacramento Mountains, the Sandia Mountains, the Manzano Mountains, Mount Taylor, and the White Mountains. There are no features that show the strong contrast in skill and bias indicative of sharp contrasts in land surface characteristics.

From Fig. 31, (DPT GE 280 K), the CSI scores overall are lower compared to the lower threshold. The only areas considered as having very good to good skill (dark orange color) are isolated very small areas located in the Black Range, west of the Zuni Mountains in the northwest corner of the domain and the foothills of the White Mountains. Lower scores indicated by the lighter orange, yellow, and green colors are distributed over lower elevation areas such as the plains of eastern New Mexico, the north Texas Panhandle, west Texas to include some lower elevation mountains, northern Mexico, southwestern New Mexico, and eastern Arizona including the Chiricahua Mountains. The lowest scores indicated by the blue shades cover areas of low and high elevation from the Sierra Madre Mountains extending eastward across the lower elevation areas of Mexico, New Mexico, and west Texas. The highest elevation mountains of Arizona and New Mexico have a white color indicative of “missing data” caused by the non-occurrence of events in the forecast and observations where the threshold was exceeded and thus do not have valid CSI and FBIAS scores.

The FBIAS scores show large areas of near-perfect FBIAS values near 1.0 indicated by the lighter blue color over large areas of the domain generally associated with the lower elevation valley and mesa areas and some lower elevation mountains in Mexico, southern New Mexico, and the plains of eastern New Mexico. The areas with significant under-forecast tendency indicated by the dark blue color are limited in areal extent mostly in the northern part of the domain in lower elevation areas northeast of the White Mountains, west of the Sandia Mountains in the Rio Grande River Valley, west of the Pecos River valley and in the playa lakes area of Mexico. Areas with significant over-forecasting tendency, indicated by the orange, yellow, and green colors, are limited in extent and confined to very small, isolated areas of lower elevation in the extreme northwest corner of the domain. These areas include the Rio Grande Valley east of the San Mateo Mountains, north of the WSNP between the San Andres Mountains and the Sacramento Mountains, and the eastern slopes of the Sacramento Mountains.

The distribution of ME and RMSE is shown in Fig. 32. ME is positive (orange, yellow, and light green colors) indicating an over-forecast tendency over large portions of the domain, mainly over the lower elevation mesas and valleys, but also over some mountainous areas too. The dark orange-colored and most conspicuous object in the domain has the highest positive bias value of approximately 2.5 K. Where ME is negative (darker green and blue shades), the WRF is under-forecasting DPT. Often these areas are in the higher elevation mountains such as the White Mountains and the Sacramento Mountains, but there is a significant exception to this pattern in the lower elevation areas including the playas or dry lake beds in Mexico west of the Rio Grande River Valley. The areas having the least bias (approximately 0) as indicated by the green colors are scattered across the domain in high and low elevation areas. Examples of the low elevation areas are the Rio Grande Valley between Texas and Mexico and the southeast corner of New Mexico. Examples of the high elevation areas are portions of the Sacramento Mountains, Mogollon Mountains, and the Manzano Mountains.

For RMSE, the lowest error magnitudes, as indicated by the green color, are in low elevation areas such as the Rio Grande Valley between Texas and Mexico and the lower elevation Davis Mountains. The largest magnitude values indicated by the orange and yellow colors frequently occur in areas of higher elevation such as the White Mountains, the Black Range, and the Mogollon Mountains. The most conspicuous object of high RMSE is the large area north of the Capitan Mountains in the high elevation mesa area west of the Pecos River Valley. It is noteworthy that the higher elevation areas of the nearby Sacramento Mountains do not have the same coverage of high values of RMSE as the large object located to the north. Some other areas of lower elevation also have the higher RMSE values in the lower

elevation and playas of northern Mexico west of the Rio Grande Valley. There appear to be other factors besides elevation that are adversely impacting DPT forecasts. One such possibility is bias in the URMA gridded observations. More investigation is needed to confirm and quantify this potential source of bias.

4.2.3 SPFH

Graphics showing the CSI, FBIAS, ME, and RMSE for 2-m-AGL SPFH for the 3-km WRF for both thresholds are presented in Figs. 33–35. Figure 33 shows the scores for SPFH GE .004 Kg/Kg and Fig. 34 shows the scores for SPFH GE .006 Kg/Kg for the 99-day period. Figure 35 shows the ME and RMSE statistics for the same period.

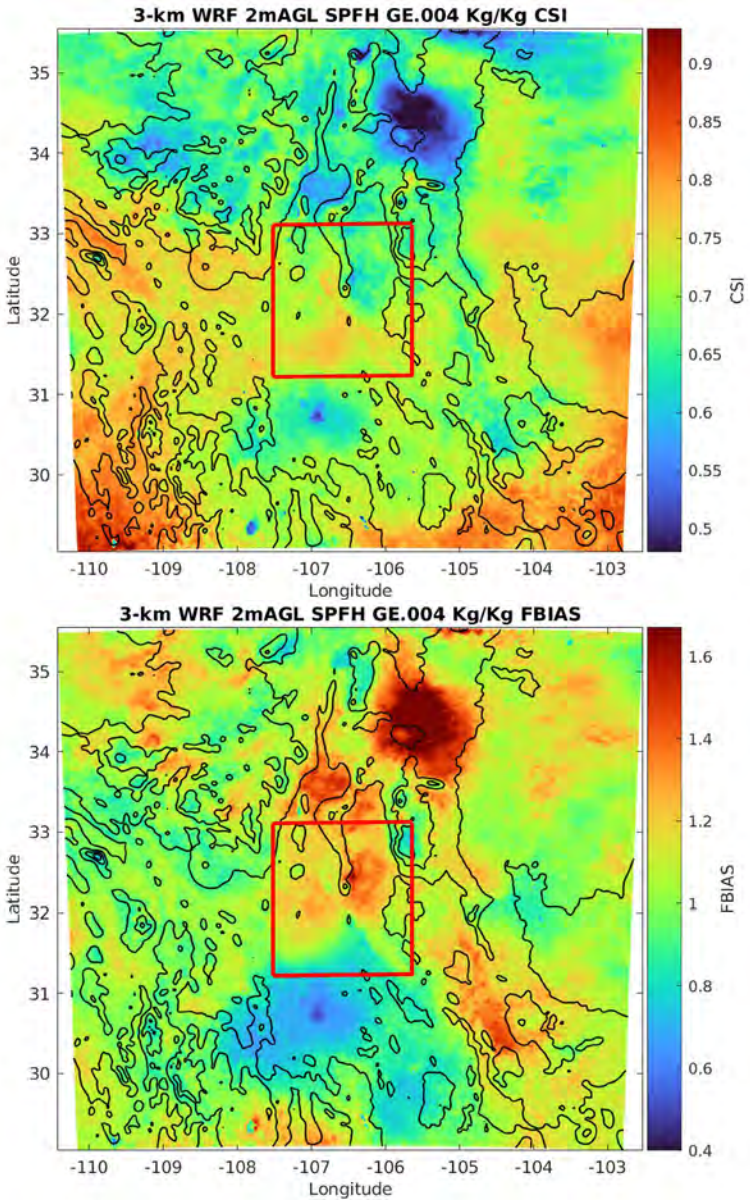

Fig. 33 CSI and FBIAS for 3-km WRF for SPFH GE .004 Kg/Kg

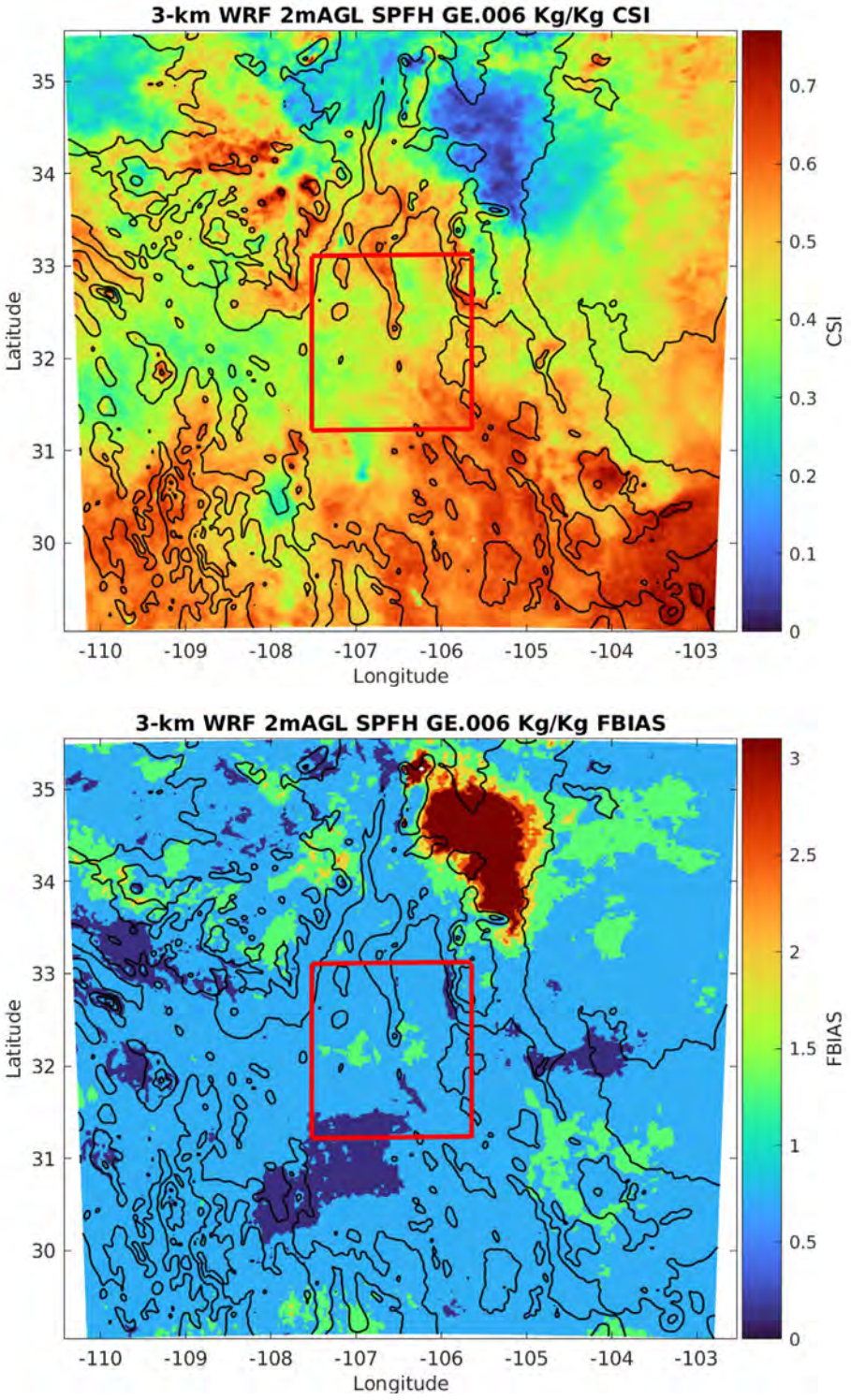

Fig. 34 CSI and FBIAS for 3-km WRF for SPFH GE .006 Kg/Kg

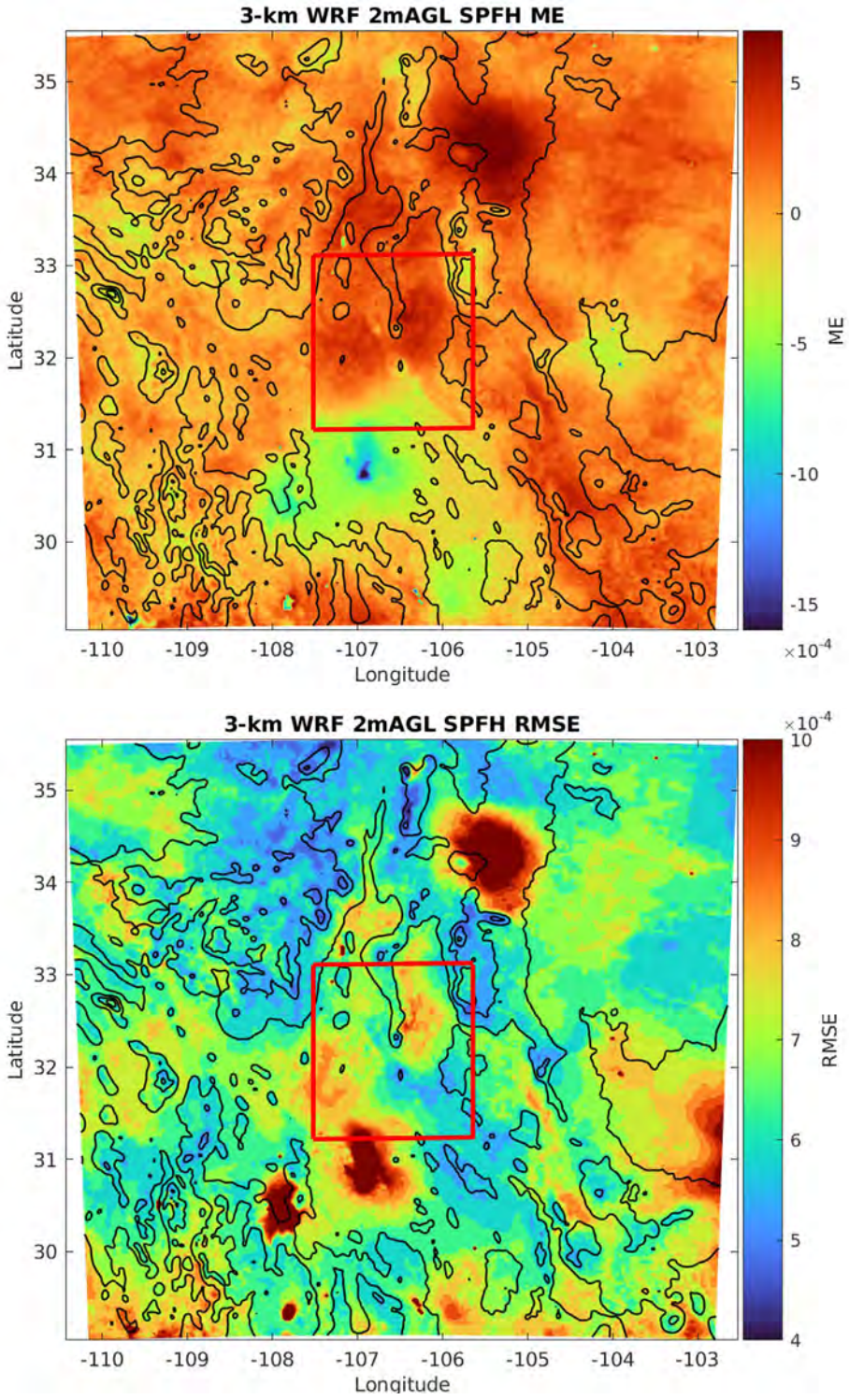

Fig. 35 ME and RMSE for 3-km WRF for SPFH

From Fig. 33, the distribution of CSI for SPFH GE .004 Kg/Kg shows good skill (darker orange areas) in the lower elevation areas of northern Mexico in the southwest corner of the domain. The lowest skill is found in areas of higher elevation such as the White Mountains, the Mogollon Mountains, and the Sandia Mountains, but there is not a clear relationship between elevation and CSI. For example, lower scores indicated by the blue color are found in the lower elevation mesa and valley areas such as the northern part of the Rio Grande Valley, the Tularosa Basin, as well as the playa areas in northern Mexico northeast of the Sierra Madre Mountains. Another interesting feature in the CSI field is the conspicuous object located north of the Capitan Mountains that has the low CSI values. This object has also been found in most of the scores and error statistics for DPT, which is not surprising since DPT and SPFH are both measures of atmospheric moisture. The contrast between this object and the higher elevation areas in the Sacramento Mountains that have better scores is evidence that other factors besides elevation are influencing the CSI scores.

The distribution of good values of FBIAS (approximately 1.0), indicated by the light green color, is across the domain mostly in the lower elevations. There is a larger, more contiguous area of low bias located in the Pecos River Valley and the plains of eastern New Mexico and the Texas Panhandle. The lowest values of FBIAS (blue color), indicative of an under-forecast tendency, are distributed over a considerable area in Mexico west of the Rio Grande Valley in generally lower elevation mesa areas east of the Sierra Madre Mountains. There are other, smaller isolated green areas of under-forecast bias in mountainous areas such as the Sierra Madre Mountains, the White Mountains, and the Sacramento Mountains, but there are also isolated areas in the Pecos River Valley, and the Rio Grande Valley south of El Paso, Texas, and near Albuquerque, New Mexico. The highest values of FBIAS, associated with an over-forecast tendency (orange shades) are located in the same area north of the Sacramento Mountains, noted previously as having the lowest CSI scores. Other areas with this tendency are in the Tularosa Basin just east of the San Andres Mountains, the Rio Grande Valley east of the San Mateo and Magdalena Mountains, and west of the Davis Mountains. There is no strong, consistent association of forecast skill with elevation evident at this threshold.

From Fig. 34, (SPFH GE .006 Kg/Kg), the CSI scores overall are lower. The highest scores, considered only fair scores, are indicated by the orange colors, and are in lower elevation mesa and mountainous areas of west Texas and Mexico and in some higher elevation mountainous areas in New Mexico such as the western foothills of the Sacramento Mountains, the San Andres Mountains, the Organ Mountains, the Black Range, the Gallo Mountains, and the San Mateo Mountains. The lowest scores, indicated by the green and blue colors, are also found in high elevation

mountainous areas such as the White Mountains, the Sacramento Mountains, the high elevation mesa area north of the Sacramento Mountains, and the Sandia Mountains. Conversely, the lowest scores are also found in lower elevation areas such as the Pecos River Valley northeast of the Sacramento Mountains, the extreme northwest corner of the domain, and the Rio Grande River Valley north of Albuquerque, New Mexico. There is no evidence of a strong, consistent association of forecast skill with elevation evident at this higher threshold.

The range of FBIAS values across the domain shows that there are areas of stronger over- and under-forecasting tendencies at this higher threshold. The best values at or near 1.0, indicated by the green colors, are found in the lower elevation mesa and mountainous areas in Mexico in the southern foothills of the Sierra Madre Mountains, west Texas, and the northern Pecos Valley in New Mexico but are also found in higher elevation locations such as the White Mountains, San Mateo Mountains, and the Magdalena Mountains. There are instances where low bias values (blue shades) are found at higher elevations in parts of the Mogollon Mountains, the Sacramento Mountains, and the San Andres Mountains. The highest over-forecast bias values, indicated by the yellow and orange colors, are distributed over parts of the White Mountains, Magdalena Mountains, and the Gallina Mountains. The largest contiguous area with high FBIAS value is the dark orange-colored object located north of the Capitan Mountains, which has also been identified as a dominant feature in the analysis of the DPT scores. The areas where the under-forecast tendency dominates, indicated by the blue colors, are found in the lower elevation playa lakes area in Mexico, the eastern foothills of the Sierra Madre Mountains, the southern facing slopes of the White Mountains, the Pecos River Valley where it crosses the New Mexico–Texas border, and the Rio Grande Valley near El Paso, Texas. In contrast, the same low values of FBIAS are found in isolated, but high elevation parts of the White Mountains, the Chiricahua Mountains, the Guadalupe Mountains, and Mount Taylor. While stronger over-forecast tendency seems to be restricted to higher elevation terrain, it is not always present over all higher elevation mountains. The same inconsistency is apparent for areas of under-forecast tendency. There are no features that show the strong contrast in skill and bias indicative of sharp contrasts in land surface characteristics.

The distribution of ME and RMSE is shown in Fig 35. ME is positive (orange colors) indicating an over-forecast tendency over large portions of the domain, mainly over the lower elevation mesas and valleys, but also over some higher elevation areas too. For example, there are red areas in lower elevation mesa areas in southern New Mexico and west Texas, but in the higher elevation mesa located north of the Capitan Mountains there is a conspicuous object of high positive ME. Negative values of ME, indicated by yellow, green, and blue colors, are found in

the lower elevation mesa areas in a large area of Mexico lying west of the Rio Grande Valley and east of the Sierra Madre Mountains including the playa lakes, the Pecos River Valley on the New Mexico–Texas border and in other small areas in lower elevation, mountainous areas such as the southern-facing slopes of the White Mountains and the Mogollon Mountains. In contrast, some areas of negative bias are also found in the higher elevation mountains such as the Sacramento Mountains, the White Mountains, Mount Taylor, and the Manzano Mountains.

For RMSE, the lowest error magnitudes, as indicated by the green and blue colors, are distributed over a considerable amount of the domain including the low elevation areas such as the Rio Grande Valley between Texas and Mexico and the area in Mexico southwest of the Sierra Madre Mountains, and some areas in the plains of eastern New Mexico near the Pecos River Valley. However, there are also low values of RMSE in higher elevation mountains such as the western foothills of the White Mountains, the San Andres Mountains, the Mogollon Mountains, the Sacramento Mountains, the Black Range, the Sandia Mountains, and the Manzano Mountains. There are also large areas of lower elevation mesa and mountainous terrain in other parts of Mexico, Texas, and New Mexico with low values of RMSE. The largest magnitude errors, indicated by the orange shades, are only found in a few isolated areas. The most notable of these is an area in northern Mexico west of the Rio Grande River in the playa lakes area and in the eastern foothills of the Sierra Madre Mountains. Some of the smaller spots in Mexico are in the lowest and the highest elevation areas, which indicate that other factors besides elevation may be contributing to the larger magnitude RMSE values. The high elevation mesa area north of the Capitan Mountains again shows a conspicuous area of higher errors compared to the surrounding lower RMSE values. There is no evidence of a strong, consistent dependency of forecast errors on elevation.

4.2.4 WIND

The graphics showing the CSI, FBIAS, ME, and RMSE for WIND for the 3-km WRF are presented in Figs. 36–38. Figure 36 shows the scores for WIND GE 3 m/s and Fig. 37 shows the scores for WIND GE 8 m/s for the 99-day period. Figure 38 shows the ME and RMSE statistics for the same period.

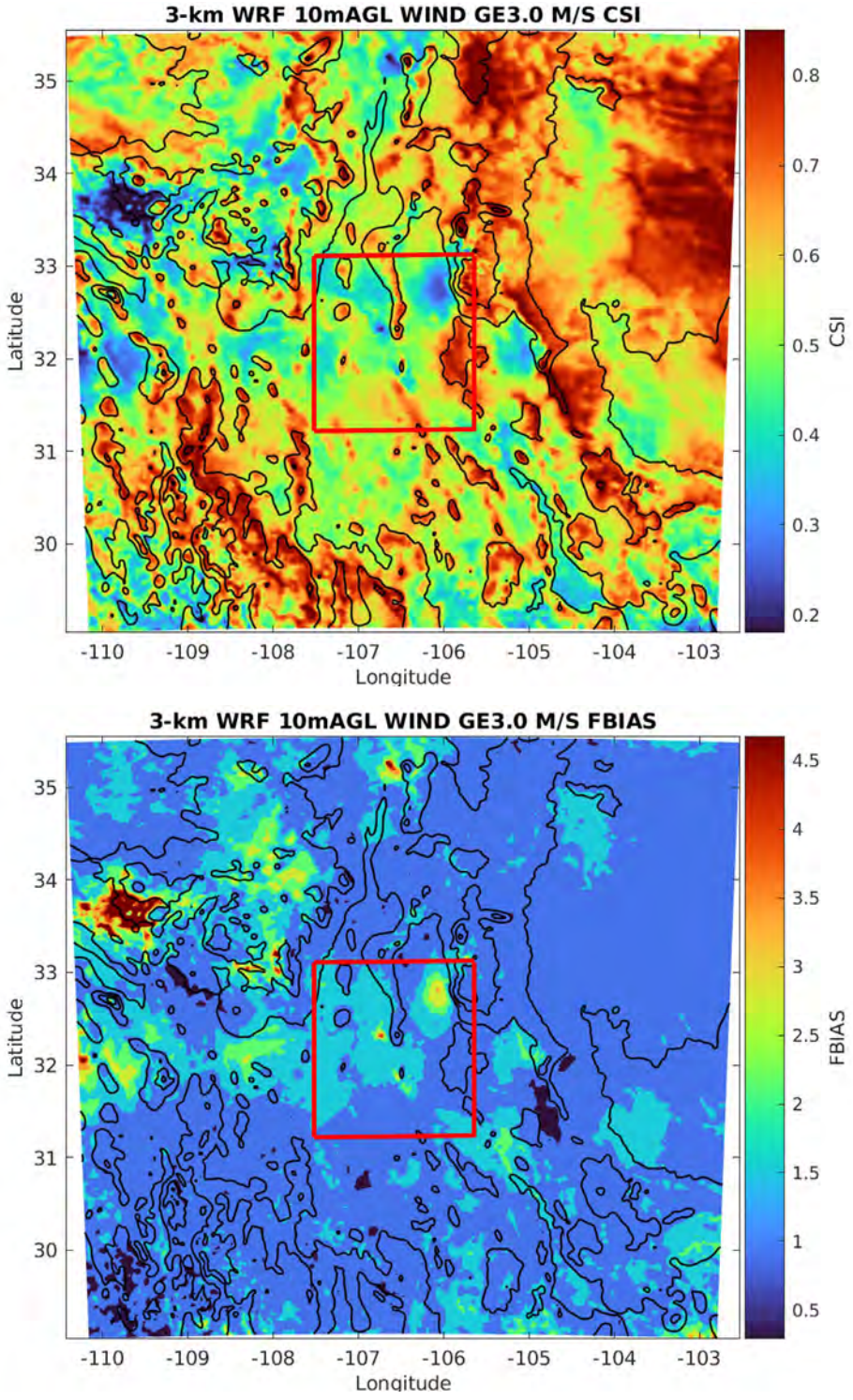

Fig. 36 CSI and FBIAS for 3-km WRF for WIND GE 3 m/s

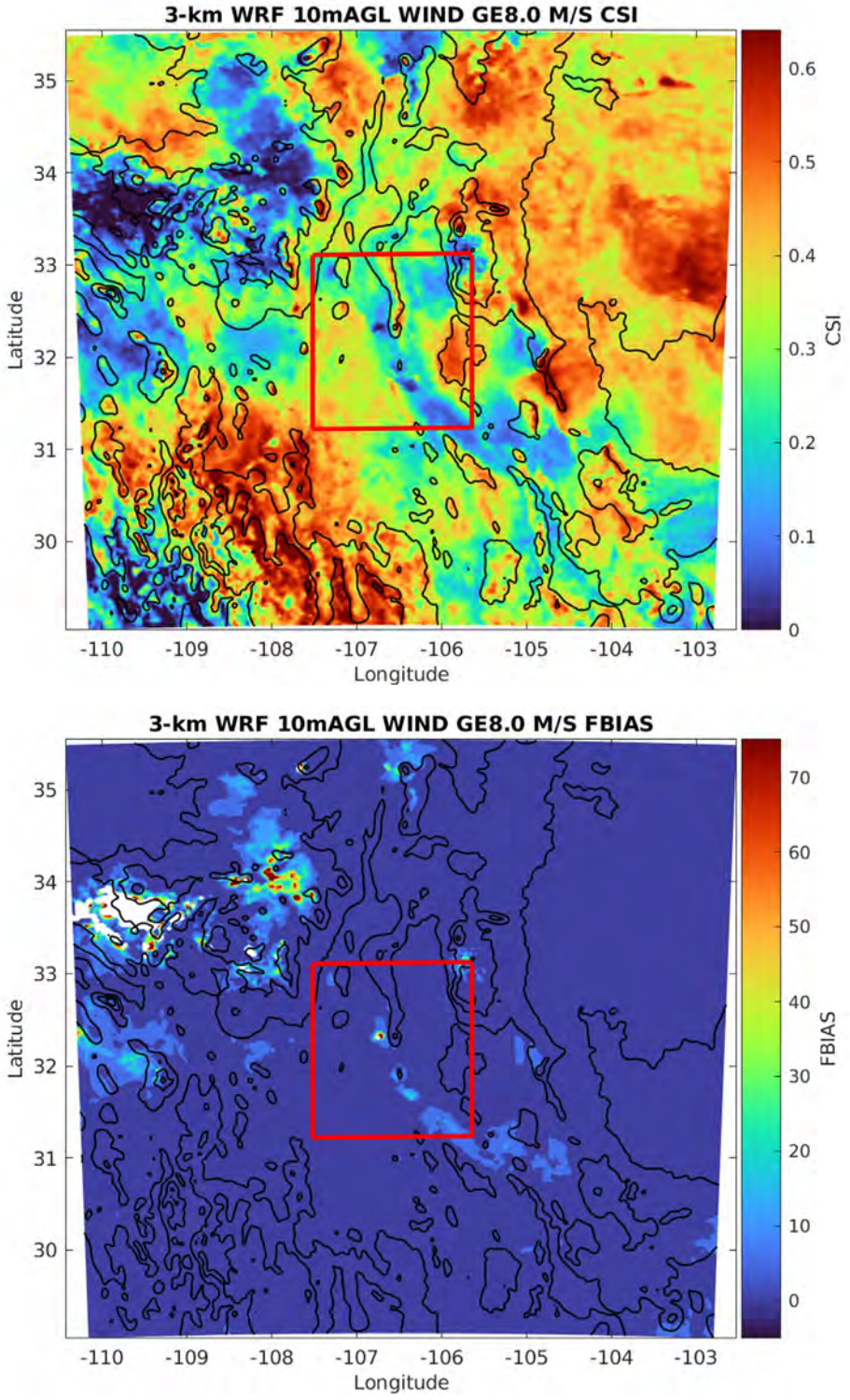

Fig. 37 CSI and FBIAS for 3-km WRF for WIND GE 8 m/s

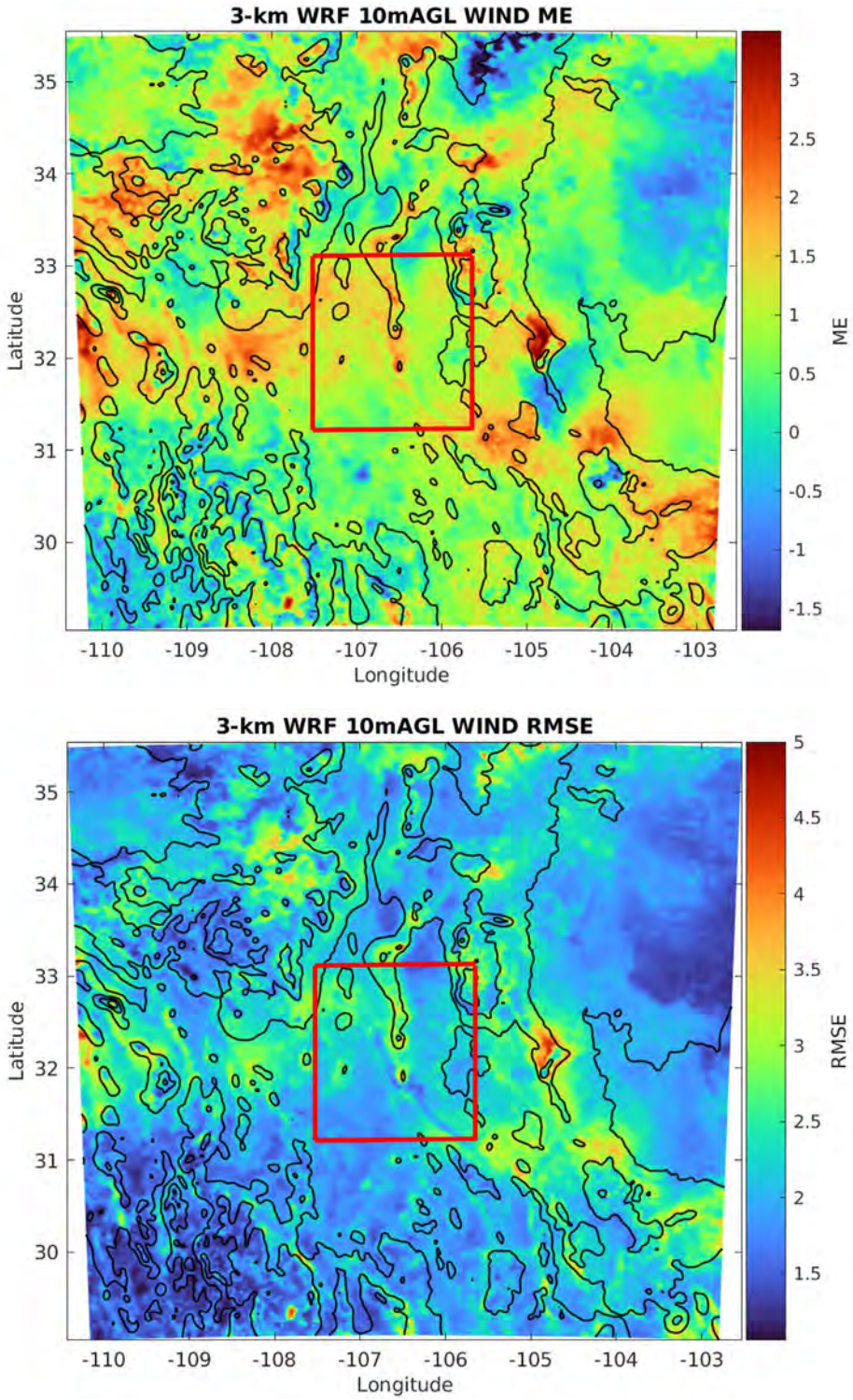

Fig. 38 ME and RMSE for 3-km WRF for WIND

From Fig. 36, the distribution of CSI for WIND GE 3 m/s shows higher skill, considered to be good skill, in the areas with orange colors. The terrain for these areas ranges from high elevation mountains such as the Sacramento Mountains, the San Andres Mountains, the Sangre de Cristo Mountains, and Sierra Madre Mountains to lower elevation areas in the plains of eastern New Mexico and the Texas Panhandle. That said, the orientation, shape, and position of some of the objects with higher scores in the CSI field suggest that there is some dependence on terrain features. For example, the position of the orange areas along the crest of the San Andres Mountains, the Organ Mountains, the Black Range, the Guadalupe Mountains, the Delaware Mountains, and the Sierra Madre Mountains suggests that the forecast skill of the WRF is higher particularly over mountains. The lowest skill, indicated by the green and blue colors, is distributed over the domain as smaller objects situated in lower elevation terrain areas such as in Mexico southwest of the Sierra Madre Mountains, along the Rio Grande River Valley, and the Gila River Valley. Some higher elevation areas also have these objects of lower scores such as the southern foothills of the White Mountains, the Gallo Mountains, and the Zuni Mountains. There appear to be pairings of lower and higher skill associated with the areas where there are large elevation gradients that are oriented in a north-south direction such as the mesa area just west of the San Andres Mountains where a small blue object is juxtaposed with the red object on the crest of the Organ Mountains. Similar pairings are found in the Tularosa Basin just west of the Sacramento Mountains and just west of the Guadalupe Mountains.

The FBIAS values show a range of tendencies from over-forecast to under-forecast. Good values of approximately 1.0, indicated by the lighter blue color, can be found in many areas across the domain in both higher and lower elevations. The areas with significant under-forecast tendency, as indicated by the dark blue colors, are distributed similarly over the domain over high and low elevation areas. Areas with significant over-forecast tendency, indicated by the yellow and orange colors, are also distributed over high and low elevation areas. The areas where there are pairings of low and high CSI scores appear to have similar pairings of areas with over-forecast tendency located just to the west of mountains, which have low bias values.

Figure 37 shows the CSI and FBIAS for WIND GE 8.0 m/s. The CSI scores overall are lower than those for the previous threshold. The areas with higher skill correspond roughly with the same lower and higher elevation areas as those for the lower threshold with a similar pattern, which suggests that the WRF has better skill over some of the higher elevation mountains. The distribution of FBIAS values shows that the range of the bias is much greater than that of the lower threshold, but most of the domain appears to have low (under-forecast) bias values indicated

by the dark blue color. These areas include both lower elevation and high elevation terrain. The highest over-forecast bias values, indicated by the colors ranging between green and orange, are restricted to very isolated, small areas that are in the White Mountains, the Gallo Mountains, the San Mateo Mountains, and Mount Taylor. However, other high elevation mountains, such as the Sacramento Mountains, do not have similar over-forecast areas.

The distribution of ME and RMSE for WIND is shown in Fig. 38. Areas that have good values (approximately 0.0) are indicated by the green color. The largest of these areas include lower elevation terrain features such as the plains of eastern New Mexico and the west Texas Panhandle, the northern Tularosa Basin just east of the San Andres Mountains, and the Rio Grande Valley east of the San Mateo and Magdalena Mountains. Areas with a negative bias, indicated by the blue colors, are limited in areal extent with the largest of these being confined to the plains of eastern New Mexico and the Texas Panhandle, the Sangre de Cristo Mountains, and a small area west of and including the Delaware Mountains in Texas. Areas that have significant positive bias, indicated by the yellow and orange colors, include the White Mountains, portions of the Sacramento Mountains, the Guadalupe Mountains, the Gallo Mountains, the San Andres Mountains, the Organ Mountains, and the Franklin Mountains. However, some lower elevation areas also have the same positive bias such as southwestern New Mexico, a portion of the Rio Grande Valley in west Texas, as well as other small areas in west Texas. The distribution of RMSE values over the domain shows that the largest magnitude WIND errors are confined to the Guadalupe Mountains and small, isolated areas in the Gila River Valley near the western domain boundary in Arizona, and farther south along the boundary in the lower elevation mesa west of the Chiricahua Mountains. The smaller magnitude errors, indicated by the blue color, are found in the plains of eastern New Mexico and the Texas Panhandle, the northern Tularosa Basin, the Rio Grande River Valley east of the San Mateo and Magdalena Mountains, the Mogollon Mountains, and the Sierra Madre Mountains. There are no features that show the strong contrast in skill and bias indicative of sharp contrasts in land surface characteristics.

The scores and statistics for the remaining variables UGRD and VGRD all present the same patterns in terms of the impact of terrain elevation and land surface characteristics on the scores and statistics. Since there are no operational thresholds for these variables, their scores are not presented here but are presented in the Appendix.

4.2.5 Analysis of Impact of Land Surface Features on Scores and Error Statistics

There is some evidence showing the impact of sharp contrasts in land surface characteristics for the 3-km WRF, but overall, the number of instances where this was evident was significantly reduced compared to the 1-km WRF. The strength of the contrast in the scores for the 3-km WRF was weak and difficult to discern owing to the small scale of the land surface features themselves. The relative differences in the impacts between the 1-km and 3-km WRF is discussed later in Section 5.

5. Summary and Conclusion

This assessment was conducted to provide a statistically strong evaluation of the accuracy of the WRF model that was run as part of ARL's WREN_RT system, which provides the forecasts of tactically significant variables and input to decision aids used for battlefield-knowledge products. Previous assessments of the WRF were based on relatively short periods of time that do not have the statistical strength attainable from a large data set of forecast and ground truth data. This assessment was the second of our ongoing program of model verification to apply a set of input data from a continuous 99-day period for the computation of spatial-verification skill scores and error statistics. The assessment used traditional categorical-verification techniques to produce scores and error statistics independently for each grid square in the two domains and aggregate these metrics over the 99-day period. The verification ground truth data were the URMA gridded observations. The data used for the evaluation consisted of WRF forecasts for the middle, 3-km, and inner, 1-km, domains and URMA gridded observations, which were collected for a 99-day winter period from 11 November 2016 to 17 February 2017. The thresholds applied to the forecasts and observations were chosen to provide maximum areal coverage of valid scores and statistics within each model domain by minimizing the occurrence of missing data generated by non-events thus optimizing the analysis with respect to all included terrain features. The domains for the data were the middle and inner nested grids of the WRF model domains located in the southwestern United States and northern Mexico characterized by complex mountain-desert-basin topography.

The MET Series-Analysis tool was used to perform the verification, which involved the ingestion of hourly WRF forecasts and URMA gridded observations of several near-surface meteorological variables for each grid point. The tool computed the differences between the WRF and observed variables at each grid point for the entire domain from which continuous error statistics were calculated for each grid square and aggregated over the 99-day period. Next, the tool applied the thresholds

and computed the categorical skill scores for each grid point and aggregated them oved the 99-day period. MATLAB was used to ingest the output of MET Series-Analysis and generate graphics of the continuous error statistics and the skill scores for selected threshold values for both domains. The plots depicting the CSI, FBIAS, ME and RMSE for all variables were analyzed to gain insight into the impact of terrain features on the accuracy of the WRF model. The accuracy of WRF predictions varied spatially across the model domains (a larger 3-km grid spacing domain and a smaller 1-km grid spacing domain) with some dependence on terrain elevation and also varied depending on the threshold value. During the analysis, it was discovered that there were significant impacts on the scores at specific locations for the 1-km WRF and to a much lesser extent for the 3-km WRF that appear attributable to sharp spatial contrasts in land surface characteristics.

5.1 1-km WRF Domain

The skill of the WRF is judged to be very good where the CSI values are GE 0.9, good where the values are GE 0.8, but LE 0.9, and fair where the values are GE 0.7, but LE 0.8. Generally, in these cases, the FBIAS values were very close to 1.0. The accuracy of the WRF is quantified by the RMSE values that express the typical error magnitude in terms of the units of the variable. The smaller the value the better. The ME is not an accuracy measure but indicates the magnitude of the over-forecast (greater than 0) and under-forecast (less than 0) tendency. (Wilks 2011) The values can be positive or negative and the smaller the absolute value the better. A value of 0 indicates a perfect forecast. The skill of the WRF is judged to be not as good where the CSI values are LT 0.7. Generally, in these cases, the FBIAS values were not as close to 1.0 showing varying degrees of over-forecast (FBIAS values greater than 1.0) and under-forecast (FBIAS values less than 1.0) tendency. These areas would be associated with larger values of RMSE and ME.

The assessment involved analysis of the plots showing the four different scores and statistics looking for patterns in the distribution of the values that might indicate a dependence on terrain features or elevation. Evidence of a dependence on elevation was found that varied depending on the variable and the threshold value. Table 2 summarizes the results of this analysis.

Table 2 Relative comparison of scores and error statistics for low elevation areas (low) vs. mountain areas (mtns) for 1-km WRF

Variable	CSI (> better than)	FBIAS (> better than)	ME (> better than)	RMSE (> better than)
TMP GE 280	Low > mtns	Low > mtns	. . .	. . .
TMP GE 290	Low > mtns	Low > mtns	. . .	. . .
TMP	. . .	. . .	Low > mtns	Mtns > low
DPT GE 265	Low > mtns	Low > mtns	. . .	. . .
DPT GE 280	Low > mtns	Low > mtns	. . .	. . .
DPT	. . .	. . .	Low > mtns	Low = mtns
SPFH GE .004	Low = mtns	Low = mtns	. . .	. . .
SPFH GE .006	Low = mtns	Low = mtns	. . .	. . .
SPFH	. . .	. . .	Low = mtns	Low = mtns
WIND GE 3	Mtns > low	Mtns > low	. . .	. . .
WIND GE 8	Mtns > low	Mtns > low	. . .	. . .
WIND	. . .	. . .	Low > mtns	Low > mtns

For the analysis the terrain was characterized as being either relatively lower elevation areas (low) that typically were the locations of the mesas, river valleys, mountain basins, and desert flatlands whose features generally have little significant relief or mountainous areas (mtns) dominated by relatively high elevation, complex terrain. The notable terrain features and their associated elevations for both domains are depicted in Figs. 2 and 3.

From Table 2, the scores for TMP (CSI, FBIAS and ME) were better over lower elevation terrain than for mountainous terrain. The opposite was true for TMP RMSE. The same was true for DPT except there was no clear-cut result for RMSE. The results for SPFH showed no clear relationship between the scores and terrain elevation. For WIND, the relationship between CSI and FBIAS and terrain elevation was the opposite of that for TMP and DPT with the better scores associated with higher elevation. For the WIND ME and RMSE statistics, the relationship was reversed to favor lower elevations having the better values. There were notable exceptions to these results for all variables, but overall, the previous relationships stood out as more dominant.

The assessment also revealed the presence of objects in the fields of the various scores and statistics that indicated the impact of land surface features on the scores and statistics. The objects were characterized by sharp contrasts in the scores over short distances. In some cases, the objects had geometric and linear features and appeared in locations corresponding to the locations of cities and other urbanized areas. In other cases, the objects coincided with the locations of area lakes. In one specific case, the object coincided with the location of the white sand dune field of the WSNP. The objects were compared with objects defined in the Noah LSM that WRF uses to simulate the land surface. It appears that there is a close

correspondence of these objects with those revealed in the scores and statistics. An analysis of the relative impacts of these objects on the scores and statistics was performed to better understand the relationships between them. The analysis results are presented in Table 3.

Table 3 Relative impacts of sharp changes in land surface characteristics on scores and error statistics for 1-km WRF

Variable	Urban areas	Lakes	White sands	Resemblance to Noah LSM
FBIAS TMP GE 280	Adverse	Adverse	None	High
FBIAS TMP GE 290	Favorable	None	Adverse	High
TMP ME	Favorable	Adverse	None	High
TMP RMSE	None	None	Adverse	Medium
FBIAS DPT GE 265	Favorable	Favorable	None	Medium
FBIAS DPT GE 280	None	None	None	None
DPT ME	Favorable	Adverse	None	Medium
DPT RMSE	Favorable	Adverse	None	Low (urban), High (lakes)
FBIAS SPFH GE .004	Favorable	Adverse	None	Medium (urban), Low (lakes)
FBIAS SPFH GE .006	Adverse	Adverse	None	Low
SPFH ME	Neither	Neither	None	Low
SPFH RMSE	None	Adverse	None	High
FBIAS WIND GE 3	Adverse	None	None	Medium
FBIAS WIND GE 8	Adverse	None	None	Low
WIND ME	None	None	None	None
WIND RMSE	Favorable	None	None	Low

For the analysis, the character of the relative impact on the scores and statistics for the three different types of objects was determined as being a net adverse impact or a net favorable impact relative to the value of the score in the areas surrounding the object. In addition, a subjective assessment of the degree of resemblance between the objects in the scores and the corresponding objects in the Noah LSM was made to provide an indication of the strength of the linkage between the two objects.

From Table 3, for TMP, the objects in every case closely resembled the corresponding object from the Noah LSM. The impact, however, varied depending on the specific score or statistic and the type of object (urban, lake, or white sand dunes). In the case of urban areas, the contributing components from the Noah LSM come from vegetation type and skin temperature that contain matching objects. In the case of the lakes, the contributing components from the Noah LSM come from vegetation type, skin temperature, land versus water type, and soil type that contain matching objects. In the case of the white sands, the contributing components from the Noah LSM come from vegetation type and soil type that contain matching objects. The net impact on the scores may be attributable to one or the other of these components or possibly some combination of them. More research is needed to isolate the relative contributions from the components to understand whether the

impact on the bias is internal to the WRF or whether the bias may be coming from the URMA gridded observations or a combination of both sources. Similar results were found for DPT, SPFH, and WIND, but the signatures of the impacts on the scores were not as well defined as with TMP and the linkages with the corresponding objects in the Noah LSM were not as strong. The white sands object was not discernable in the scores for these other three variables. Impacts from land surface characteristics were largely absent from the scores and statistics of the 3-km WRF. There was one instance of an impact on the CSI and FBIAS scores for TMP GE 290 K associated with the white sand dunes that was adverse. There are corresponding objects for the white sands in the Noah LSM vegetation type and soil type for the 3-km WRF.

These results raise some interesting questions about the impact of the Noah LSM in WRF. Regarding the impacts in urban areas, further investigation is needed in cases when the WRF urban parameterization is not used to determine what default soil category is being used. Knowing this information might prove useful to identifying aspects that may be impacted such as the heat and moisture fluxes and the Bowen ratio. Regarding the impacts over the lakes, more investigation is needed to understand the how poorly initialized water temperature values might result in a localized increase in bias. Can such a bias that impacts only a few grid points adversely affect the skill scores? Additional research is needed to quantify the contribution of bias coming from the URMA gridded observations and to understand if these biases serve to amplify or de-amplify the scores and error statistics. It is likely that the lack of sufficient point observations in key areas such as the higher elevation mountains, the white sands dune fields, or the small inland water bodies has introduced a natural bias in the gridded observations, which in turn may have enhanced or offset biases in the WRF forecasts for these areas. Certainly, there is the possibility that assessing of the skill of the 1-km WRF using the URMA analysis on a 2.5-km grid may have affected the scores of the 1-km WRF in a negative way due to the potential lack of skill of the URMA in capturing smaller scale, terrain-induced features.

5.2 3-km WRF Domain

The assessment showed that the WRF on a 3-km grid performed similar to the 1-km WRF in terms of the relationship between the scores and terrain elevation. Table 4 summarizes the results of the analysis of dependence of the scores on elevation for the 3-km WRF.

Table 4 Relative comparison of scores and error statistics for low elevation areas (low) vs. mountain areas (mtns) for 3-km WRF

Variable	CSI (> better than)	FBIAS (> better than)	ME (> better than)	RMSE (> better than)
TMP GE 280	Low > mtns	Low > mtns	. . .	. . .
TMP GE 290	Low > mtns	Low > mtns	. . .	. . .
TMP	. . .	. . .	Low > mtns	**Low = mtns**
DPT GE 265	Low > mtns	**Low = mtns**	. . .	. . .
DPT GE 280	Low > mtns	Low > mtns	. . .	. . .
DPT	. . .	. . .	**Low = mtns**	Low = mtns
SPFH GE .004	Low = mtns	Low = mtns	. . .	. . .
SPFH GE .006	Low = mtns	Low = mtns	. . .	. . .
SPFH	. . .	. . .	Low = mtns	Low = mtns
WIND GE 3	Mtns > low	**Mtns = low**	. . .	. . .
WIND GE 8	Mtns > low	**Mtns = low**	. . .	. . .
WIND	. . .	. . .	Low > mtns	Low > mtns

The instances where there were differences between the results for the 1-km WRF and the 3-km WRF are in bold font. No differences were noted in the results for the CSI scores between the two model nests for all four variables. For FBIAS, there were some small differences noted in the relationship between elevation and the scores. For DPT GE 265, there was no significant difference in the scores at low elevations compared to the higher elevations, which was not the case for the 1-km WRF where the scores for the lower elevations were better than those of the higher elevations. The same difference was also noted in the FBIAS for WIND GE 3.0 m/s and GE 8 m/s where the 3-km showed no difference attributable to elevation. For ME, the results for the 1-km WRF for DPT showed a clear dependence on elevation with the lower elevation areas having the lowest bias error compared to the 3-km WRF where there was no apparent dependence. The results for both models were identical for the other variables. For RMSE, the results for the 1-km WRF for TMP showed a clear dependence on elevation with the lower elevation areas having the lowest error magnitude compared to the 3-km WRF where there was no apparent dependence. The results for both models were identical for the other variables.

A comparison was made to assess the impact on the scores attributable to sharp changes in land surface characteristics between the 1-km WRF and the 3-km WRF. The 1-km WRF results clearly showed significantly more impacts than the 3-km WRF. Table 5 summarizes the results of the analysis of impacts of land surface characteristics.

Table 5 Relative impacts of sharp changes in land surface characteristics on scores and error statistics for 3-km WRF

Variable	Urban areas	Lakes	White sands	Resemblance to Noah LSM
FBIAS TMP GE 280	Adverse	None	None	**Low**
FBIAS TMP GE 290	None	None	Adverse	**Medium**
TMP ME	None	None	None	None
TMP RMSE	None	None	None	None
FBIAS DPT GE 265	Favorable	None	None	**Low**
FBIAS DPT GE 280	None	None	None	None
DPT ME	None	None	None	None
DPT RMSE	None	None	None	None
FBIAS SPFH GE .004	None	None	None	None
FBIAS SPFH GE .006	None	None	None	None
SPFH ME	None	None	None	None
SPFH RMSE	None	None	None	None
FBIAS WIND GE 3	Adverse	None	None	**Low**
FBIAS WIND GE 8	Adverse	None	None	Low
WIND ME	None	None	None	None
WIND RMSE	None	None	None	None

The instances where there were differences between the results for the 1-km WRF and the 3-km WRF are in bold font. The most noteworthy difference was the significant reduction in the presence of sharp changes in the scores in the 3-km WRF compared to the 1-km WRF. For the few instances where a sharp contrast was noted in the 3-km WRF TMP, DPT, and WIND, the character of the impact (adverse/favorable) did not change from that for the same variables for the 1-km WRF. Regarding the resemblance to the Noah LSM, there were a few differences noted for TMP, DPT, and WIND. In all cases, the resemblance decreased for the 3-km WRF. The lack of sharp changes in the scores attributable to changes in the land surface characteristics appears to be caused by the difficulty in discerning these features on the coarser grid. The features are small in areal extent and become "washed out" or eliminated at larger scale sized. This makes an analysis of the contribution to the differences attributable to differences in the model configurations difficult to accomplish.

6. Future Work

Additional assessments using different techniques are needed to characterize the skill of the WRF more completely in view of the uncertainty in the scores from this study arising from the exceptions and anomalies discussed in Section 4. To improve the assessment of the impacts attributable to changes in land surface characteristics for the 3-km WRF, better visualization of the scores within the 1-km WRF domain may be achievable by running the MET Series-Analysis tool so that the scores are

plotted over the 1-km WRF domain only rather than over the entire 3-km domain. This would result in plots of the same size as those for the 1-km WRF, which would facilitate comparisons. Conducting further assessments would lead to a better understanding of the impact of land surface characteristics on model errors. Furthermore, understanding which surface effects, or combinations thereof, from the Noah LSM are contributing to the errors would be of value to modelers striving to improve model performance. To achieve this, studies are needed that provide more in-depth assessments of the 1-km WRF nest using new methodologies to target the identification of the relative impacts from the various land surface types. To better understand these impacts, one possibility may be to apply techniques that can isolate the land surface type in the Noah LSM, which is contributing significantly to the bias. This information could assist in efforts to improve model performance. Smith and Penc (2017) describe a promising approach that uses the statistical Design of Experiments (DoE) technique and present a method for developing the design matrix for applying the technique to NWP forecasts. Smith et al. (2019) demonstrates an application of this method to NWP. Although other methods are available to study these factor level effects, for example Stein and Alpert (1993), Cleveland et al. (2020) demonstrates that factor methods are a special case of DoE. The DoE technique involves a controlled statistical analysis of numerous model runs that were configured and run as prescribed by the design matrix.

Another approach may provide better insight into the impacts of land surface characteristics. A promising new source of ground truth, available from an array of meteorological towers in the Multipurpose Sensing Area (MSA), has been developed by ARL at WSMR. The MSA consists of numerous instrumented towers that are placed in regularly spaced arrays at the JER. Having point observations on such a dense grid and at multiple levels provides a wealth of data that can be used for verification.

7. References

Benjamin S, Weygandt S, Brown J, Hu M, Alexander C, Smirnova T, Olson J, James E, Dowell D, Grell G, et al. A North American hourly assimilation and model forecast cycle: the rapid refresh. Mon Weather Rev. 2016;144(4):1669–1694.

Case JL, White KD, Guyer B, Meyer J, Srikishen J, Blankenship C, Zavodsky BT. Real-time land information system over the continental US for situational awareness and local numerical weather prediction applications. American Meteorological Society (AMS) Annual Meeting; 2016 Jan. No. MSFC-E-DAA-TN28946.

Cleveland JL, Smith JA, Collins JP. Factor effects in numerical simulations. J Atmos Sci. 2020;77(7):2439–2451.

Dawson L, Raby J, Smith J. The automation of nowcast model assessment processes. Army Research Laboratory (US); 2016 Sep. Report No.: ARL-MR-0940.

De Pondeca Manuel SFV, Manikin G, DiMego G, Benjamin S, Parrish D, Purser RJ, Wu WS, Horel J, Myrick D, Lin Y, et al. The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: current status and development. Weather Forecast. 2011;26(5):593–612.

Ebert E. Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. Meteorol Appl. 2008;15:51–64.

Grell GA, Freitas S. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. Atmos Chem Phys. 2014;14:5233–5250. doi:10.5194/acp-14-5233-2014.

Gunn A, Wanker M, Lancaster N, Edmonds DA, Ewing RC, Jerolmack DJ. Circadian rhythm of dune-field activity. Geophysical Research Letters. 2021;48(5):e2020GL090924. https://doi.org/10.1029/2020GL.

Halley Gotway J, Newman K, Soh H, Opatz J, Jensen T, Prestopnik J, Goodrich L, Fillmore D, Brown D, Bullock R, Fowler T. The MET version 10.0.0 user's guide. Developmental Testbed Center; 2021 [accessed 2022 Sep 6]. https://github.com/dtcenter/MET/releases.

Iacono MJ, Delamere JS, Mlawer EJ, Shephard MW, Clough SA, Collins WD. Radiative forcing by long–lived greenhouse gases: calculations with the AER radiative transfer models. J Geophys Res. 2008;113:D13103. doi:10.1029/2008JD009944.

Jensen T, Brown B, Bullock R, Fowler T, Gotway JH, Newman K. Model evaluation tools version 9.0.1 user's guide; 2020 Apr [accessed 2021 June 29]. Developmental Testbed Center. https://dtcenter.org/sites/default/files/community-code/met/docs/user-guide/MET_Users_Guide_v9.0.pdf 479 pp.

Jolliffe IT, Stephenson DB. Forecast verification: a practitioner's guide in atmospheric science. 2nd ed. John Wiley and Sons; 2012.

Jones SL, Adams-Selin R, Hunt ED, Creighton GA, Cetola JD. Adapting WRF-CHEM GOCART for fine-scale dust forecasting. AGU Fall Meeting Abstracts; 2010. Vol. 1.

Jones SL, Adams-Selin R, Hunt ED, Creighton GA, Cetola JD. Update on modifications to WRF-CHEM GOCART for fine-scale dust forecasting at AFWA. AGU Fall Meeting Abstracts; 2012.

Massey JD, Steenburgh WJ, Hoch SW, Jensen DD. Simulated and observed surface energy fluxes and resulting playa breezes during the MATERHORN field campaigns. J Appl Meteorol Climatol. 2017;56(4):915–935.

Mittermaier M, Roberts N, Thompson SA. A long-term assessment of precipitation forecast skill using the fractions skill score. Meteorol Appl. 2013;20:176–186.

Morris M, Carley J, Colon E, Gibbs A, De Pondeca M, Levine S. A quality assessment of the real-time mesoscale analysis (RTMA) for aviation. Weather Forecast. 2020;35:977–996.

Nakanishi M, Niino H. An improved Mellor-Yamada level 3 model: its numerical stability and application to a regional prediction of advection fog. Boundary Layer Meteorology. 2006;119:397–407. doi:10.1007/s10546-005-9030-8.

[NCAR] User's guide for the advanced research WRF (ARW) modeling system version 3.8. National Center for Atmospheric Research; 2016 [accessed 2020 Dec 4]. http://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3.8/contents.html.

[NCEP] Unified post processor (UPP). Ver. 3.0. National Centers for Environmental Prediction; 2020 [accessed 2021 Apr 21]. https://dtcenter.org/sites/default/files/community-code/upp-users-guide-v3.pdf.

[NOAA] NCEP central operations, real-time mesoscale analysis products. National Oceanic and Atmospheric Administration; 2017 [accessed 2020 Nov 05]. https://www.nco.ncep.noaa.gov/pmb/products/rtma/#URMA.

Pondeca M, Levine S, Carley J, Lin Y, Zhu Y, Purser J, McQueen J, Yang R, Gibbs A, Parrish D, DiMego G. Ongoing improvements to the NCEP real time mesoscale analysis (RTMA) and unrestricted mesoscale analysis (URMA) and NCEP/EMC; 2015 [accessed 2020 Nov 6]. https://www.wcrp-climate.org/WGNE/BlueBook/2015/individual-articles/01_Pondeca_Manuel_etal_RTMA.pdf.

Raby J. Application of a fuzzy verification technique for assessment of the Weather Running Estimate – Nowcast (WRE-N) model. Army Research Laboratory (US); 2016 Oct. Report No.: ARL-TR-7849.

Raby JW, Cai H. Verification of spatial forecasts of continuous meteorological variables using categorical and object-based methods. Army Research Laboratory (US); 2016 Aug. Report No.: ARL-TR-7751.

Raby J, Cai H, Dawson L, Dumais R. An evaluation of the unrestricted mesoscale analysis as gridded observations for spatial model verification. DEVCOM Army Research Laboratory (US); 2020 Nov. Report No.: ARL-TR-9115.

Raby J, Cai H, Dawson L, Reen B. A 99-day assessment of the Weather Research and Forecasting Model over the southwest United States. DEVCOM Army Research Laboratory (US); 2021 July. Report No.: ARL-TR-9237.

Reen BP. A brief guide to observation nudging in WRF. University Corporation for Atmospheric Research; 2016 [accessed 2020 Dec 4]. http://www2.mmm.ucar.edu/wrf/users/docs/ObsNudgingGuide.pdf.

Reen BP, Dawson LP. The Weather Running Estimate–Nowcast Realtime (WREN_RT) system, version 1.03. Army Research Laboratory (US); 2018 Sep. Report No.: ARL-TR-8533. https://apps.dtic.mil/sti/pdfs/AD1060869.pdf.

Ruth D, Huntemann T, Plumb D. Verification of the national blend of models. 28th Conference on Weather Analysis and Forecasting/24th Conference on Numerical Weather Prediction; 2017. Amer Meteor Soc, 7B.3. https://ams.confex.com/ams/97Annual/webprogram/Paper305573.html.

Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda M, Huang XY, Wang W, Powers JG. A description of the advanced research WRF version 3. University Corporation for Atmospheric Research; 2008. Report No.: NCAR/TN-475+STR. doi:10.5065/D68S4MVH.

Smith JA, Penc RS. A design of experiments approach to evaluating parameterization schemes for numerical weather prediction: problem definition and proposed solution approach. Joint Statistical Meetings

Proceedings, Section on Statistics in Defense and National Security, Conference on Applied Statistics in Defense; 2015 Aug 8–13. 2017 Jan. p. 4183–4192.

Smith JA, Cleveland JL, Raby JW, Penc R. Applying design of experiments to numerical weather prediction. Annual Joint Statistical Meeting, American Statistical Association; 2019.

Stauffer D. Uncertainty in environmental NWP modeling. In: Fernando HJ, editor. Handbook of environmental fluid dynamics. Vol. 2, systems, pollution, modeling, and measurements. CRC Press; 2012. Chapter 29.

Stein U, Alpert P. Factor separation in numerical simulations. J Atmos Sci. 1993;50(14):2107–2115.

Tewari M, Chen F, Wang W, Dudhia J, LeMone MA, Mitchell K, Ek M, Gayno G, Wegiel J, Cuenca RH. 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. American Meteorological Society; 2004. p. 11–15.

Thompson G, Eidhammer T. A study of aerosol impacts on clouds and precipitation development in a large winter cyclone. J Atmos Sci. 2014;71(10):3636–3658. doi:10.1175/JAS-D-13-0305.1.

[UCAR] Operational models encyclopedia. University Corporation for Atmospheric Research; 2015 [accessed 2021 July 02]. https://sites.google.com/ucar.edu/operational-models-encyclo/deterministic-models/analyses/rtma-urma.

Wang Y, Benson MJ. Large-eddy simulation of turbulent flows over an urban building array with the ABLE-LBM and comparison with 3D MRI observed data sets. Environ Fluid Mech. 2021;21:287–304. https://doi.org/10.1007/s10652-020-09770-6.

Weygandt S, Alexander C, Ge G, Hu M, Ladwig T, Hartsough C, Carley J, Zhao G, Pondeca M, Yang R. Evaluation of a prototype version of the 3D-real-time mesoscale analysis (3D-RTMA) for situational awareness and nowcast applications. 35th Conference on Environmental Information Processing Technologies; 2019 Jan 8 [accessed 2020 Oct 12]. https://ams.confex.com/ams/2019Annual/webprogram/Paper353171.html.

Wilks DS. Statistical methods in the atmospheric sciences. 3rd ed. Academic Press; 2011.

Appendix. Critical Success Index (CSI), Frequency Bias (FBIAS), Mean Error (ME), and Root Mean Squared Error (RMSE) for U Wind Component (UGRD), V Wind Component (VGRD)

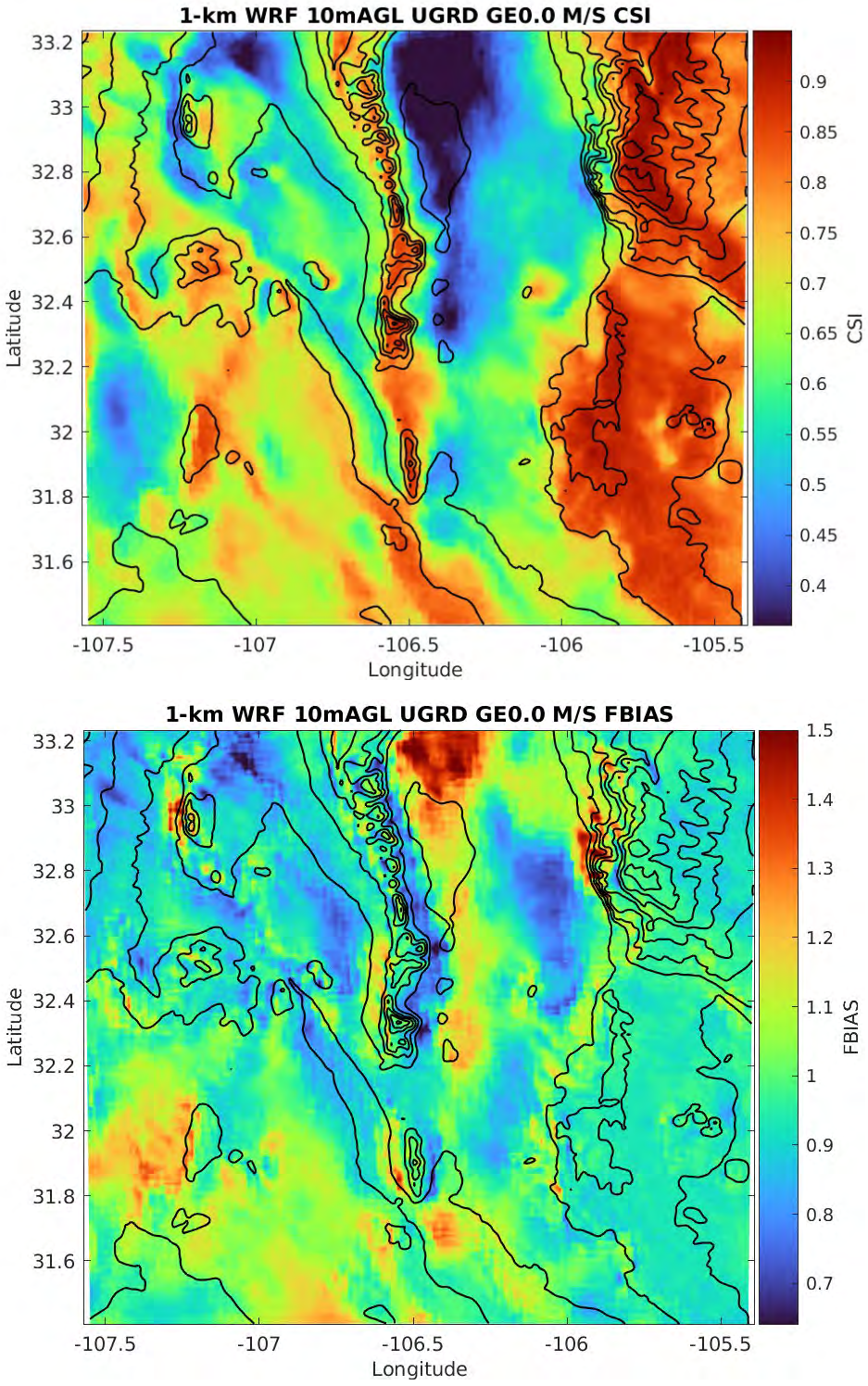

Fig. A-1 Critical success index (CSI) and frequency bias (FBIAS) for 1-km Weather Research and Forecasting (WRF) for U wind component (UGRD) greater than or equal to (GE) 0 m/s

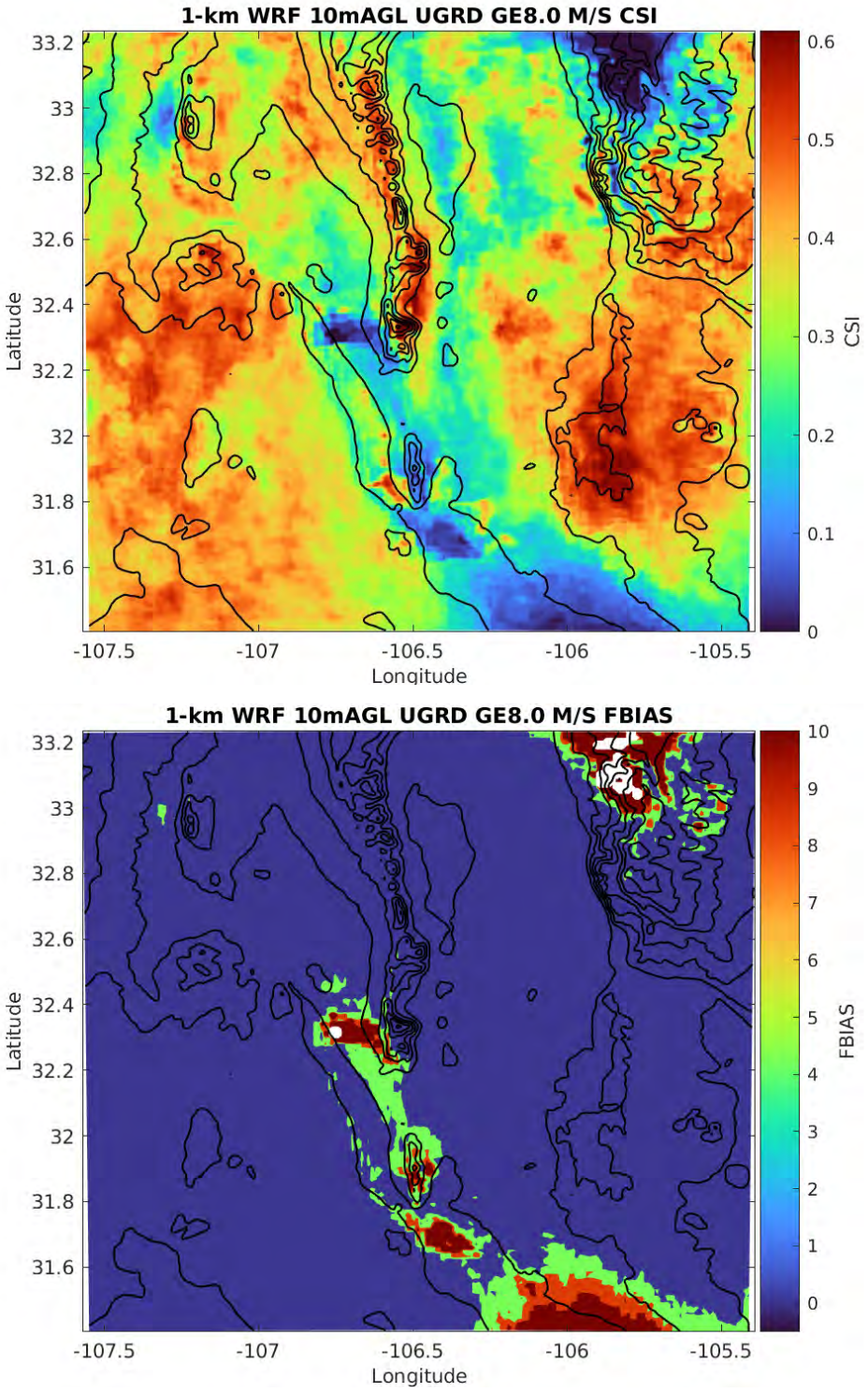

Fig. A-2 CSI and FBIAS for 1-km WRF for UGRD GE 8 m/s

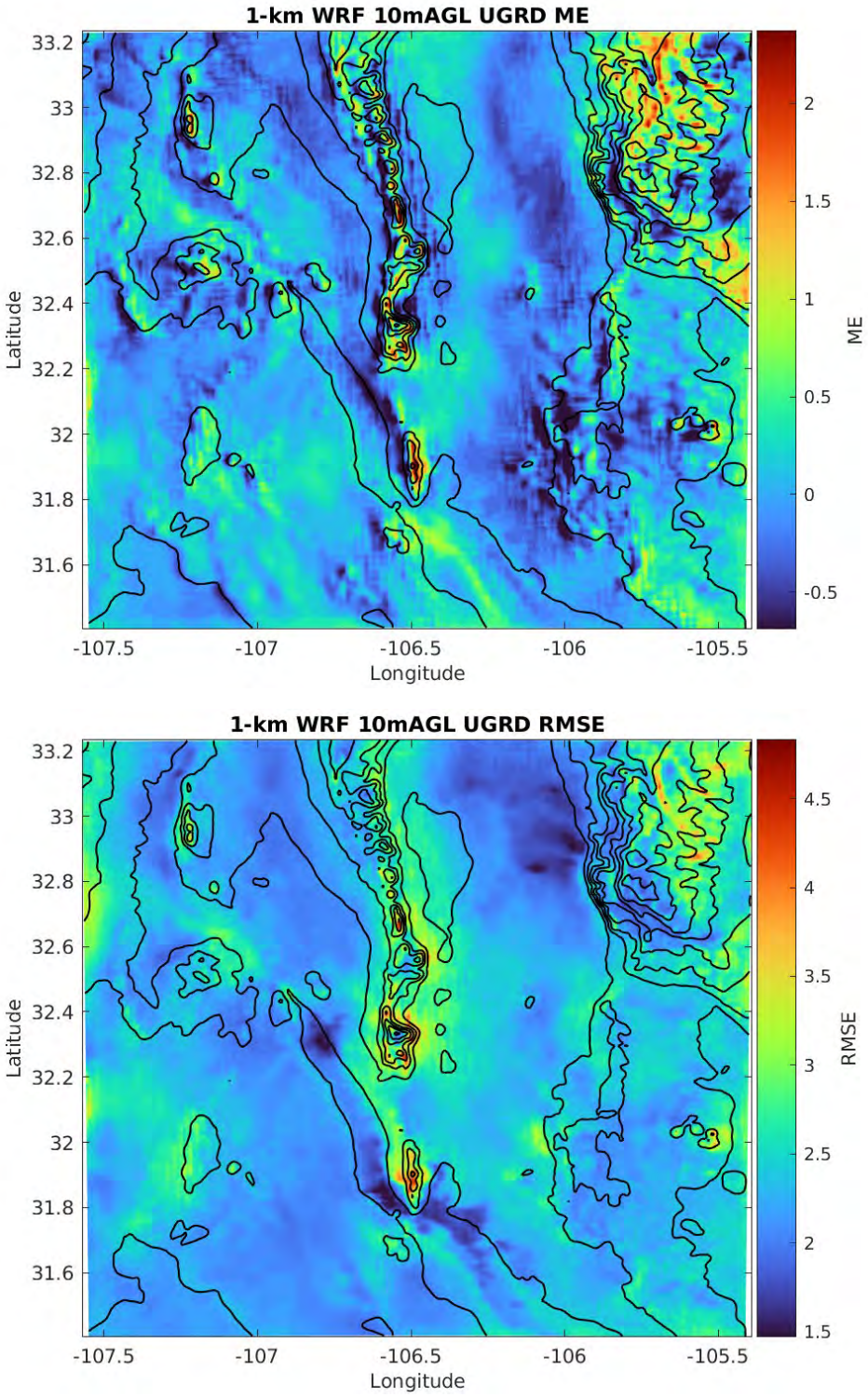

Fig. A-3 Mean error (ME) and root mean squared error (RMSE) for 1-km WRF for UGRD

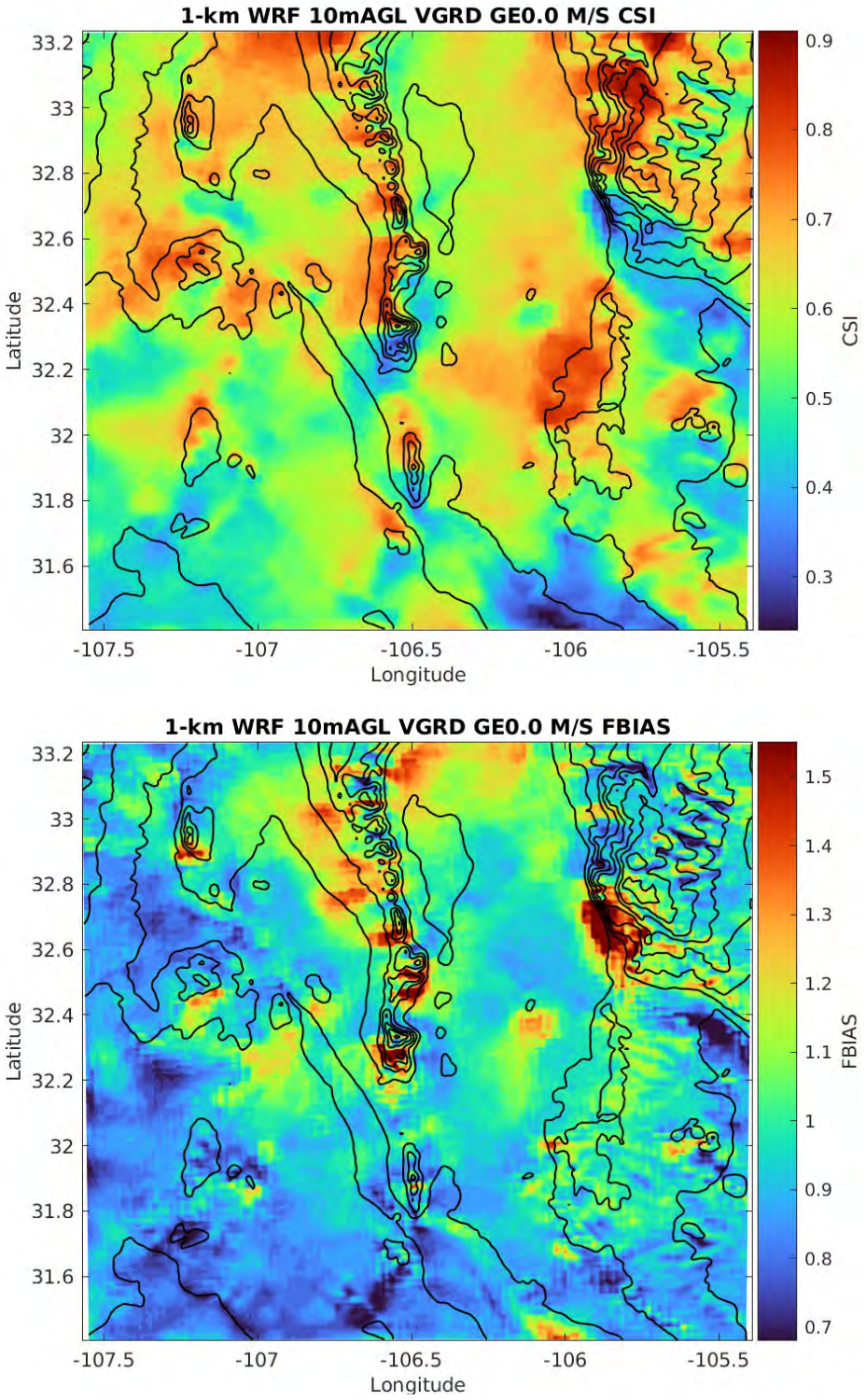

Fig. A-4 CSI and FBIAS for 1-km WRF for V wind component (VGRD) GE 0 m/s

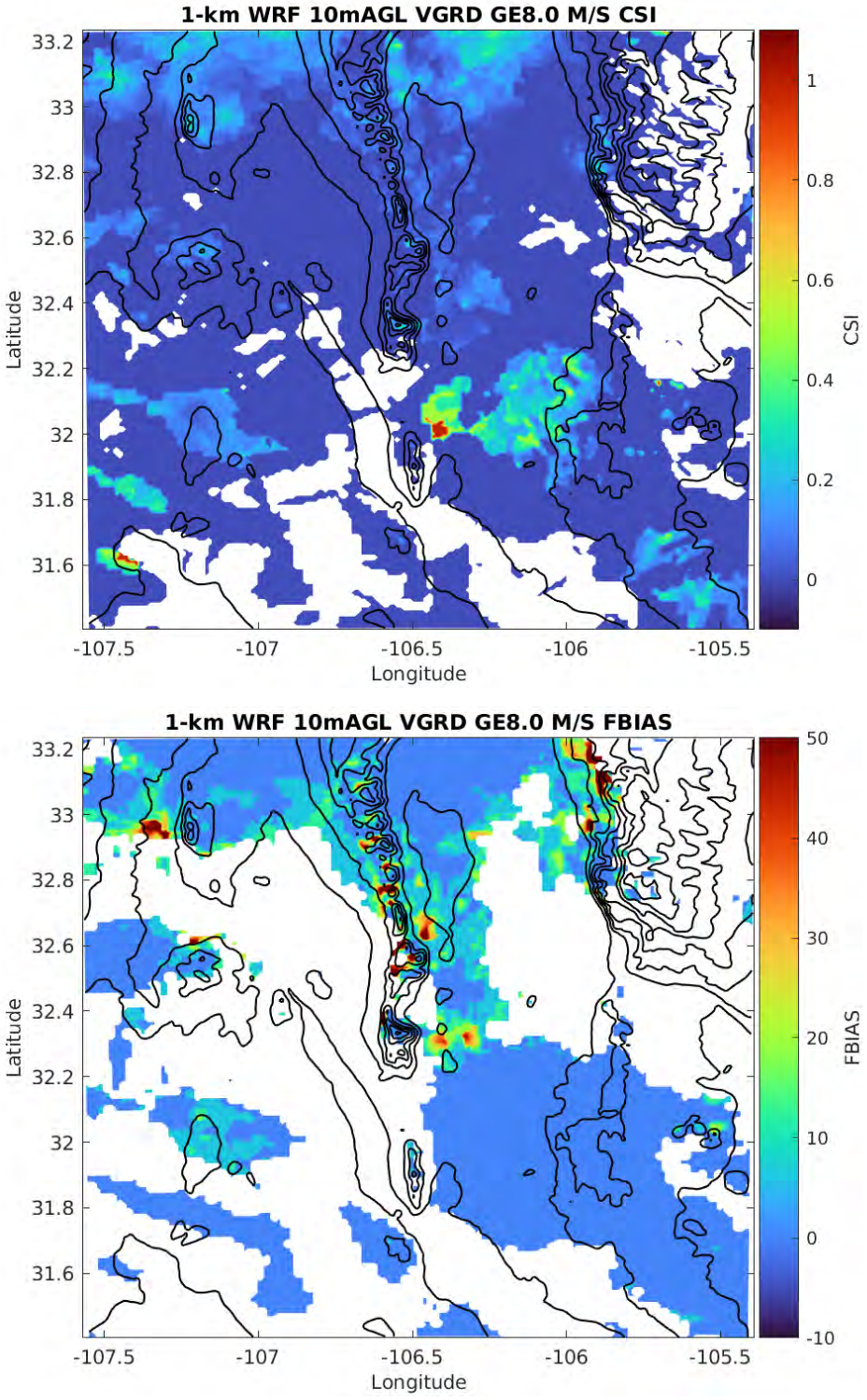

Fig. A-5 CSI and FBIAS for 1-km WRF for VGRD GE 8 m/s

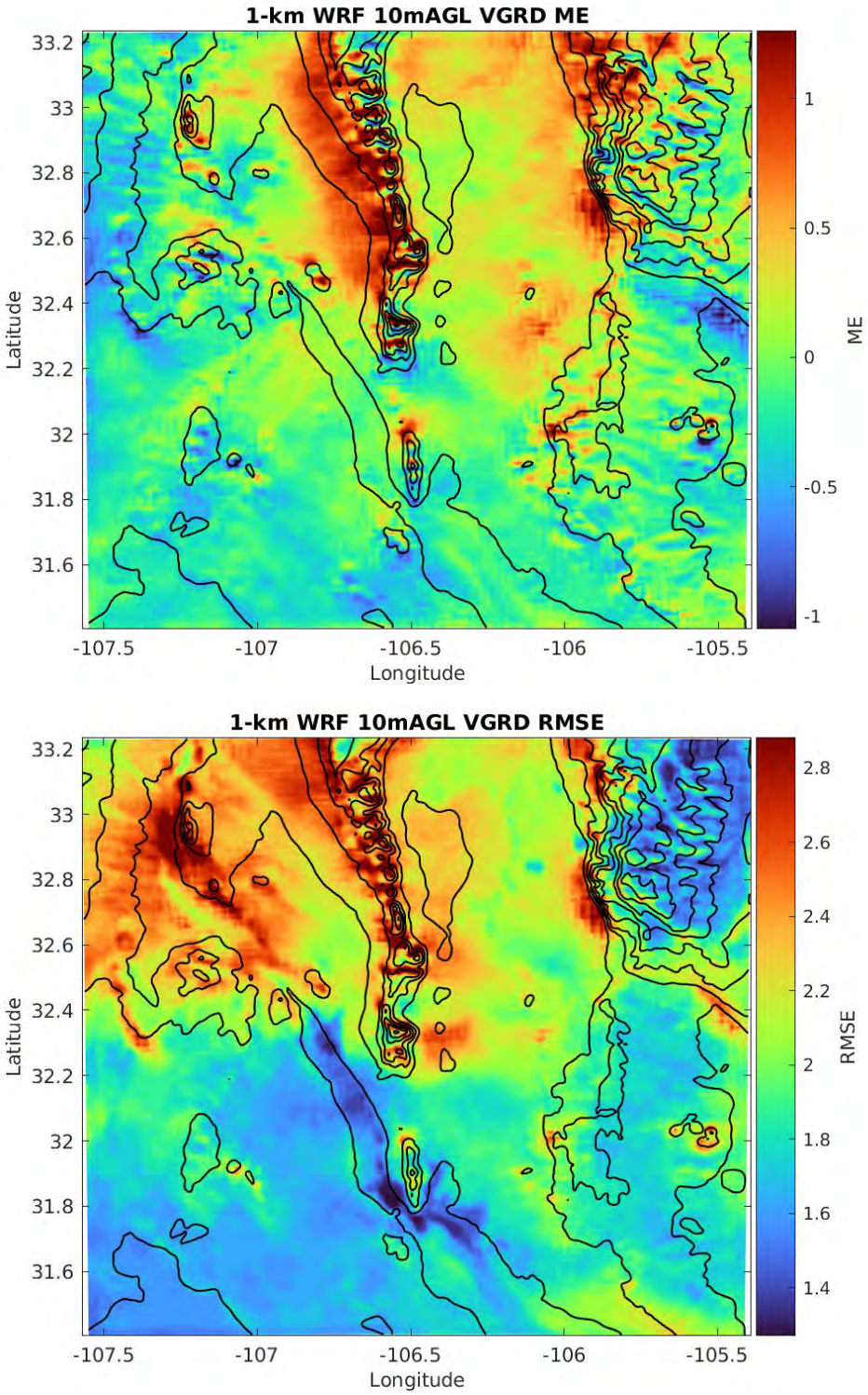

Fig. A-6 ME and RMSE for 1-km WRF for VGRD

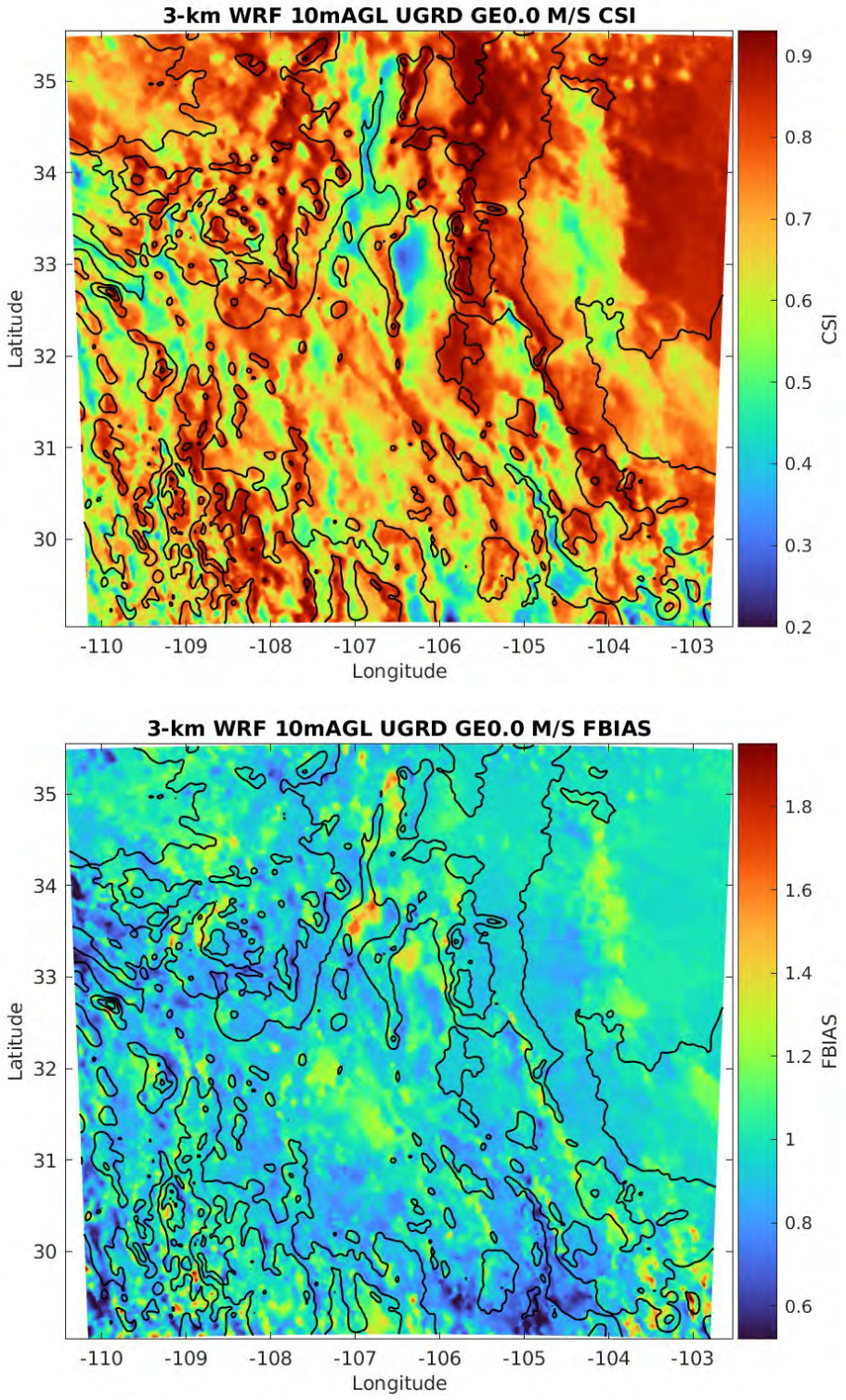

Fig. A-7 CSI and FBIAS for 3-km WRF for UGRD GE 0 m/s

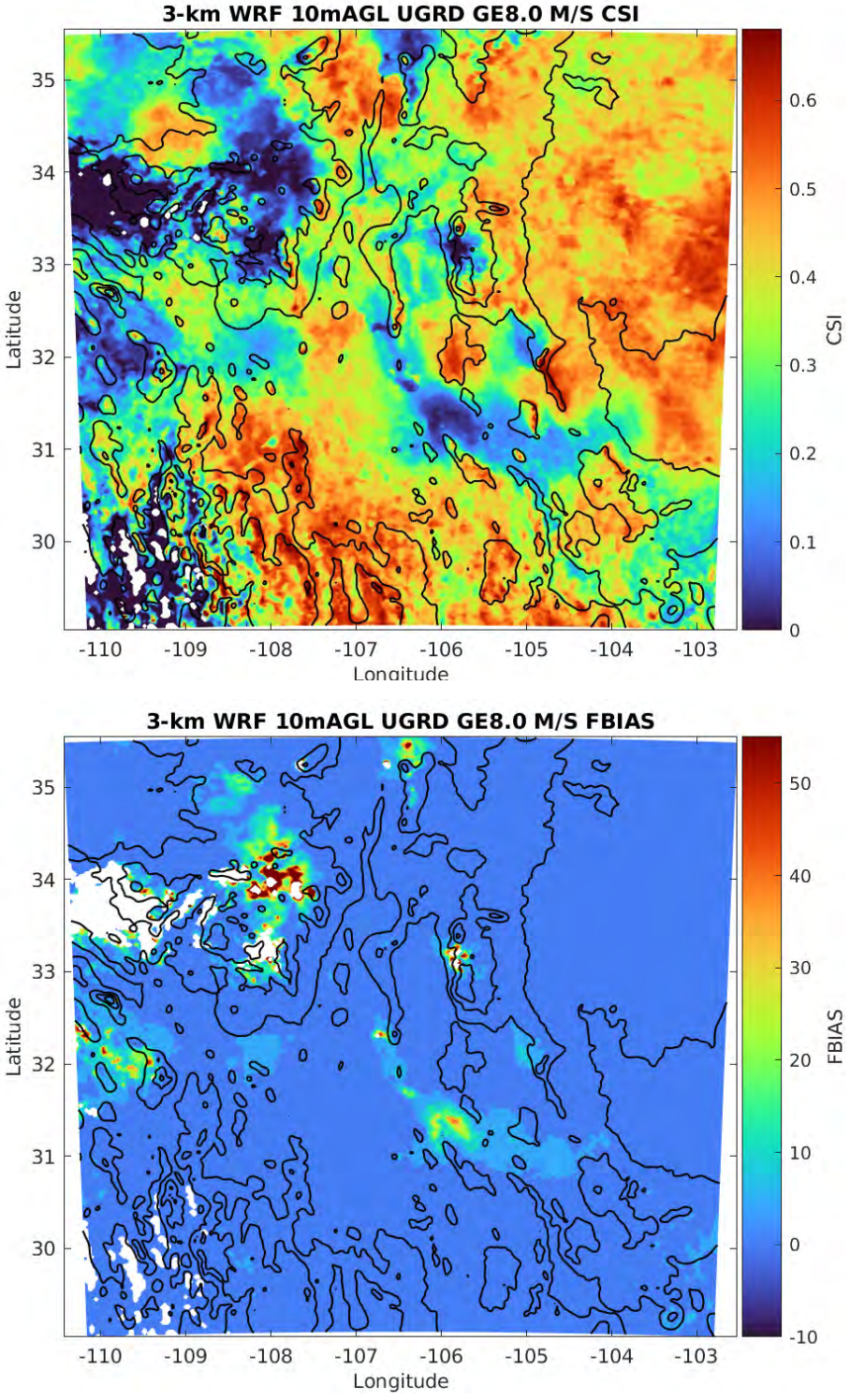

Fig. A-8 CSI and FBIAS for 3-km WRF for UGRD GE 8 m/s

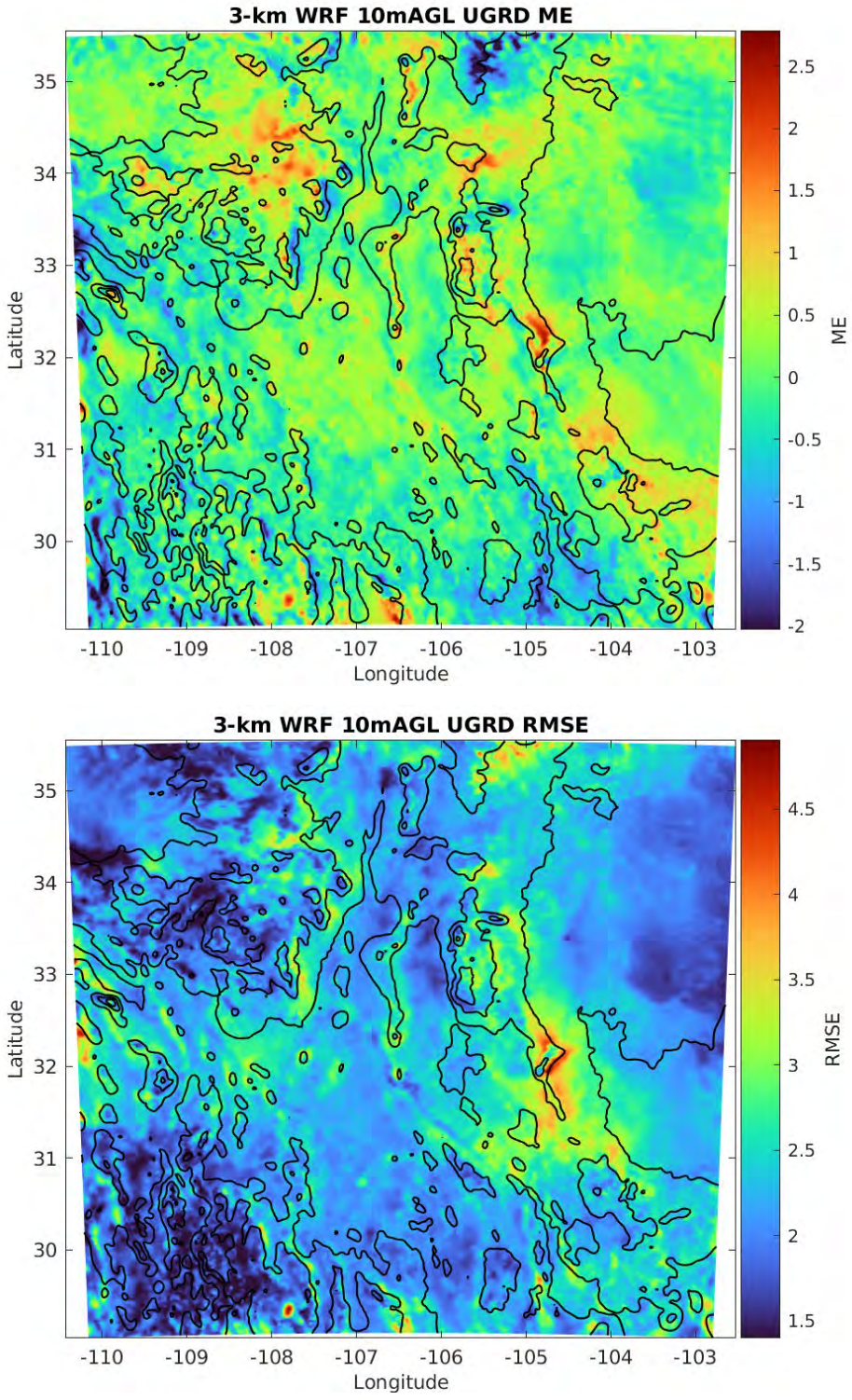

Fig. A-9 ME and RMSE for 3-km WRF for UGRD

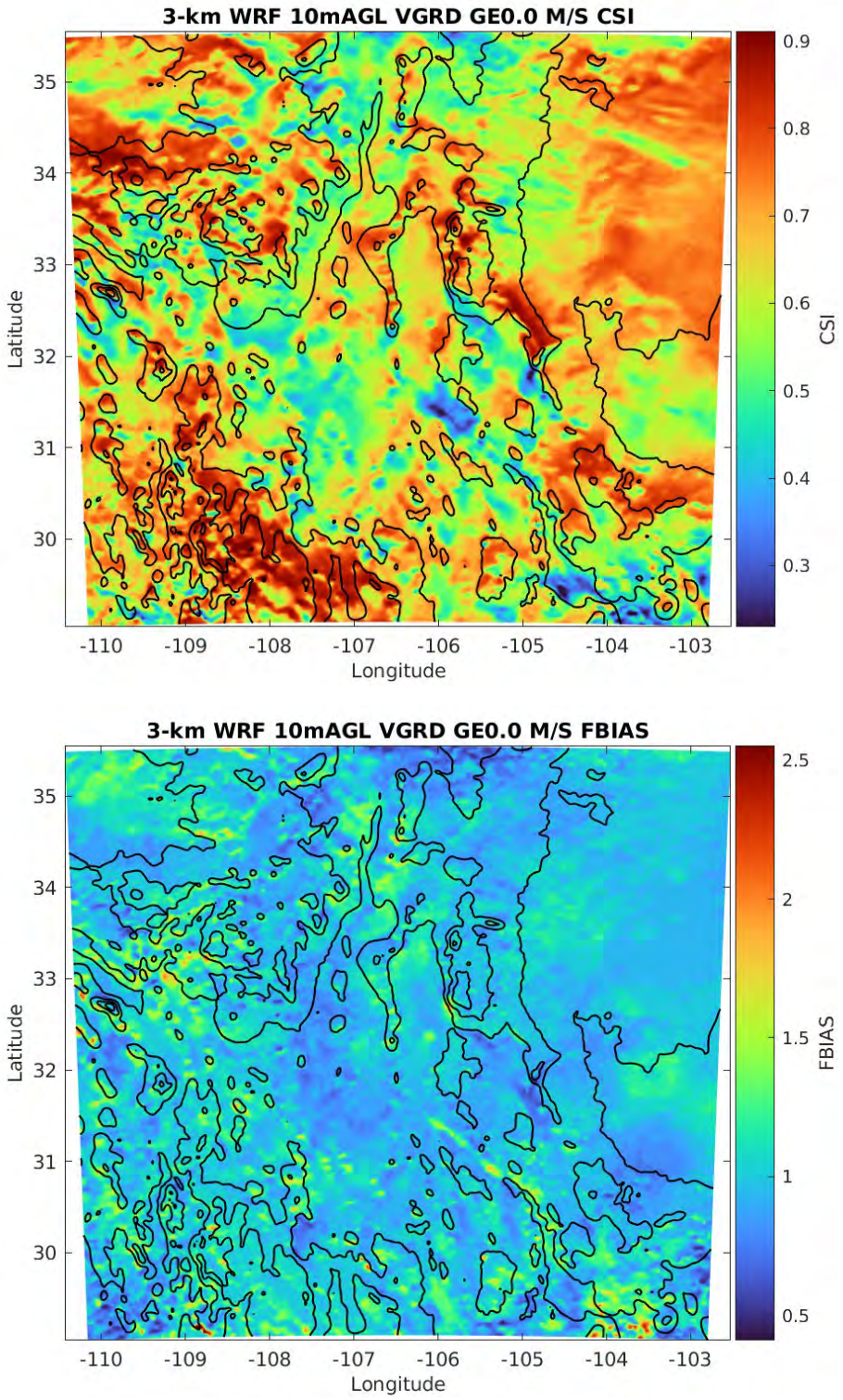

Fig. A-10 CSI and FBIAS for 3-km WRF for VGRD GE 0 m/s

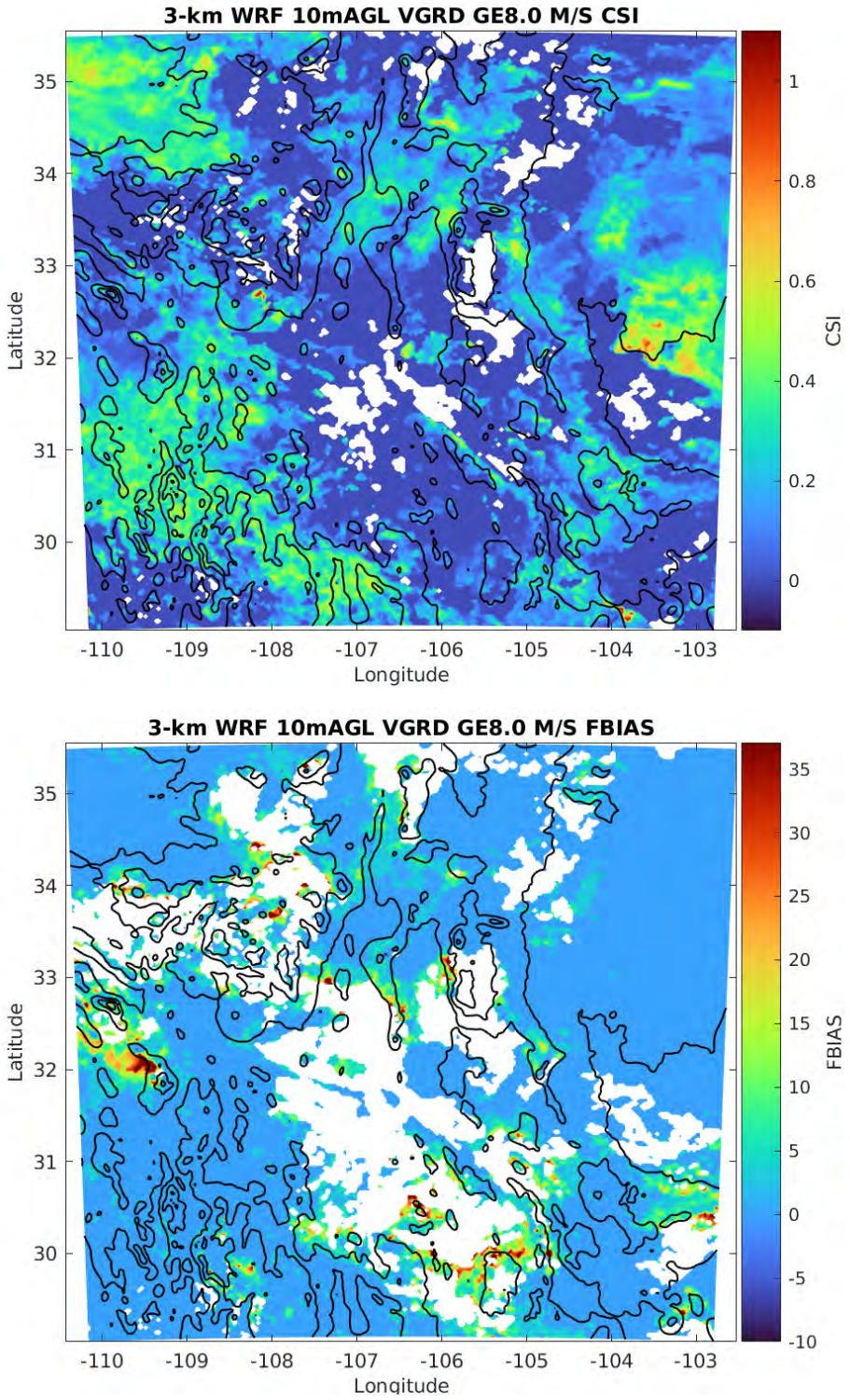

Fig. A-11 CSI and FBIAS for 3-km WRF for VGRD GE 8 m/s

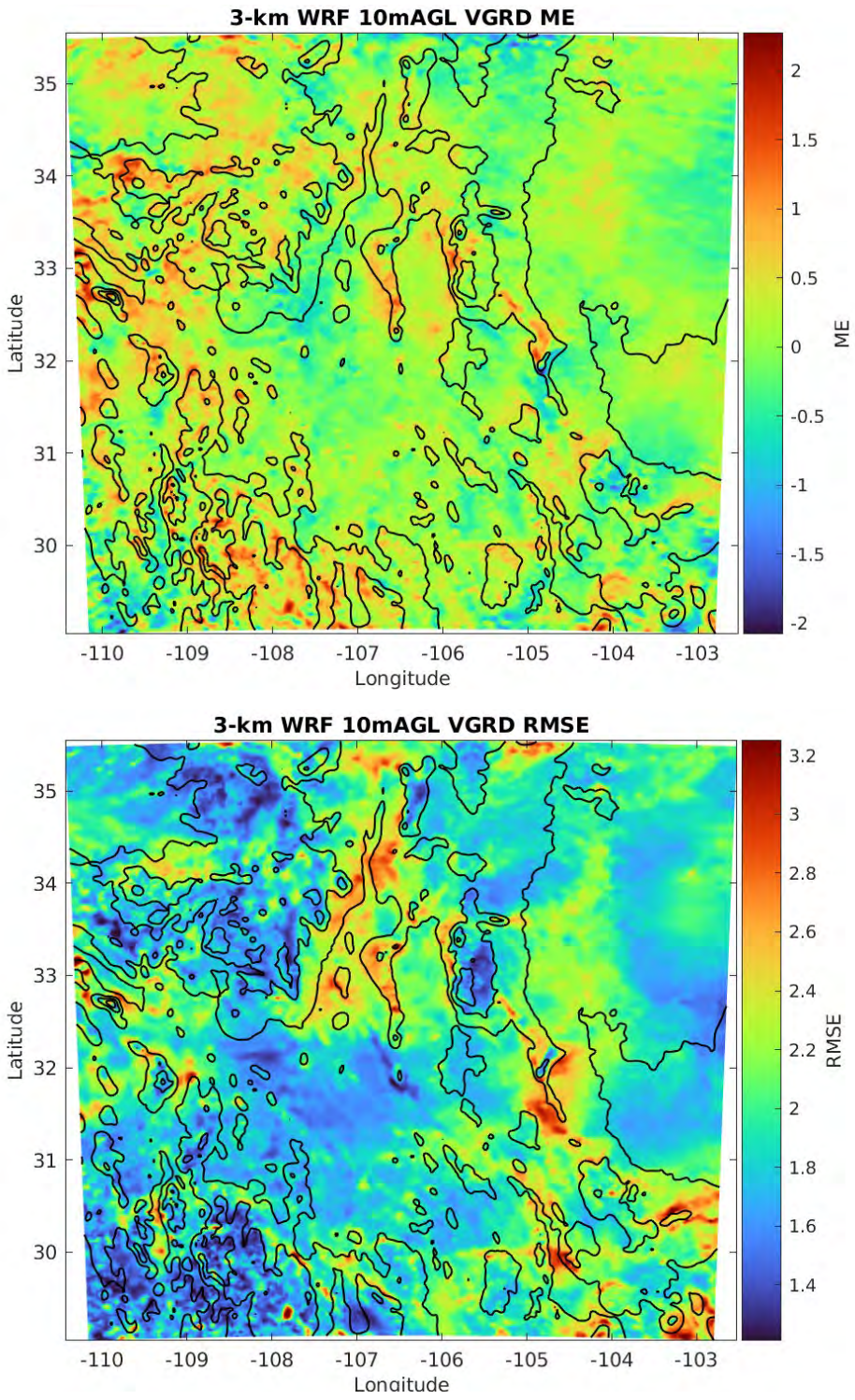

Fig. A-12 ME and RMSE for 3-km WRF for VGRD

List of Symbols, Abbreviations, and Acronyms

2-D	two-dimensional
2DVAR	two-dimensional variational data assimilation
3-D	three-dimensional
ACARS	Aircraft Communications, Addressing, and Reporting System
AFB	Air Force Base
AGL	above ground level
ARL	Army Research Laboratory
CONUS	continental United States
CSI	Critical Success Index
DEVCOM	US Army Combat Capabilities Development Command
DOE	US Department of Energy
DoE	Design of Experiments
DPT	dew-point temperature
FBIAS	frequency bias
GE	“greater than or equal to” logical statement
GT	“greater than” logical statement
HRRR	High-Resolution Rapid Refresh
JER	Jornada Experimental Range
K	Kelvin
LE	“less than or equal to” logical statement
LIDAR	light detection and ranging
LSM	land surface model
LT	“less than” logical statement
m/s	meters per second
MADIS	Meteorological Assimilation Data Ingest System
ME	mean error
MET	Model Evaluation Tools
METAR	Météorologique Aviation Régulière
MSA	Multipurpose Sensing Area
MYNN	Mellor-Yamada Nakanishi Niino

NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction
NOAA	National Oceanic and Atmospheric Agency
NSF	US National Science Foundation
NWP	Numerical Weather Prediction
NWS	National Weather Service
PBL	planetary boundary layer
RAP	Rapid Refresh
RMSE	root-mean-squared error
RTG SST	real-time, global, sea surface temperature
RTMA	Real-Time Mesoscale Analysis
SPFH	specific humidity
TMP	temperature
UGRD	U wind component
UPP	Unified Post Processor
URMA	UnRestricted Mesoscale Analysis
USAF	US Air Force
UTC	Coordinated Universal Time
VGRD	V wind component
WIND	wind speed
WRE-N	Weather Running Estimate – Nowcast
WREN_RT	Weather Running Estimate – Nowcast Real-Time
WRF	Weather Research and Forecasting
WRF-ARW	Weather Research and Forecasting – Advanced Research
WRF-Chem	Weather Research and Forecasting – Chemistry
WSMR	White Sands Missile Range
WSNP	White Sands National Park

1 (PDF) DEFENSE TECHNICAL INFORMATION CTR DTIC OCA

1 (PDF) DEVCOM ARL FCDD RLB CI TECH LIB

6 (PDF) DEVCOM ARL FCDD RLA ID J RABY R DUMAIS B REEN C HOCUT FCDD RLA IF H CAI FCDD RLA NA L DAWSON