

1 **Comparing and Combining Deterministic Surface Temperature**
2 **Postprocessing Methods over the US**

3
4
5 Thomas M. Hamill

6 *NOAA Physical Sciences Laboratory*

7 *Boulder, Colorado, USA*

8
9
10
11 to be submitted to *Monthly Weather Review*

12
13
14
15 13 February 2021

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31 Corresponding author information:

32
33 Dr. Thomas M. Hamill
34 NOAA Physical Sciences Laboratory
35 R/PSL 1
36 325 Broadway
37 Boulder, CO 80305 USA
38 e-mail: tom.hamill@noaa.gov
39 phone: (303) 497-3060
40 telefax: (303) 497-6449

41
42 ABSTRACT

43
44 Common methods for the statistical postprocessing of deterministic 2-meter temperature
45 (T_{2m}) forecasts over US and nearby land regions were evaluated at leads from +12 h to +120 h.
46 Forecast data were extracted from the Global Ensemble Forecast System (GEFS) v12
47 reforecast data set and thinned to a 1/2-degree grid encompassing the contiguous US.
48 Analyzed data from the European Centre/Copernicus reanalysis (ERA5) were used for training
49 and validation. Data from the 2000-2018 period were used for training, and 2019 forecasts
50 were validated. The statistical postprocessing methods compared were the raw forecast
51 guidance, a decaying-average bias correction (DAV), quantile mapping (QM), a univariate
52 model output statistics (uMOS) algorithm, and a multi-variate (mvMOS) algorithm. mvMOS
53 used the raw forecast temperature, the DAV bias correction, and the QM adjustment as
54 predictors.

55 Forecasts from all the post-processing methods reduced the root-mean-square error
56 (RMSE) and bias relative to the raw guidance. QM produced forecasts with slightly higher error
57 than DAV, though error differences were not always statistically significant. uMOS and mvMOS
58 produced statistically significant lower RMSEs than DAV at forecast leads longer than 1 day,
59 with mvMOS exhibiting the lowest error. Taylor diagrams showed that the MOS methods
60 reduced the variability of the forecasts while improving forecast-analyzed correlations. QM and
61 DAV modified the distribution of forecasts to more closely exhibit those of the analyzed data.

62 A main conclusion, reinforcing that found by others, is that the judicious statistical
63 combination of guidance from multiple post-processing methods is capable of producing
64 forecasts with improved error statistics relative to any one individual post-processing technique
65 on its own. As each post-processing method applied here is algorithmically relatively simple,
66 this suggests that operational deterministic postprocessing could produce improved T_{2m}

67 guidance with little effort, assuming that reduction of error is the primary criterion for evaluating
68 the post-processing procedure.

69 **1. Introduction.**

70 Much of the attention in the recent literature on the statistical postprocessing of forecasts
71 has shifted to the postprocessing of ensemble prediction system guidance and the production of
72 skillful and reliable probabilistic forecasts. This is reflected in a recent textbook (Vannitsem et
73 al. 2018) highlighting developments in this discipline. Despite the evolution in this direction,
74 many weather prediction centers still produce deterministic forecast guidance from a variety of
75 methods, especially forecasts of more statistically straightforward quantities such as surface
76 temperature and particularly at shorter forecast lead times (days, not weeks). Hence, it is still of
77 practical interest to operational weather prediction centers to understand the potential strengths
78 and weaknesses of several plausible candidate statistical post-processing methods.

79 In this article we compare the characteristics of several algorithmically simple methods
80 when applied to the statistical correction of two-meter above ground surface temperatures (T_{2m}).
81 The algorithms are the decaying-average (DAV) bias correction (Cui et al. 2012), quantile
82 mapping (QM; Hopson and Webster 2010, Voisin et al. 2010, Maraun 2013), and Model Output
83 Statistics (MOS) regression techniques (Glahn and Lowry 1972, Carter et al. 1989). While this
84 is not an exhaustive list, these represent different techniques with different underlying correction
85 principles, and each is used operationally in different contexts. In fact, in the US National
86 Weather Service, each of these is used. The DAV method is used in the National Blend of
87 Models (NBM; Craven et al. 2020). QM is also used in the NBM for precipitation forecasts
88 (Hamill et al. 2017, Hamill and Scheuerer 2018), and MOS is still used for station data
89 postprocessing (e.g., Glahn et al. 2009).

90 Algorithms often employ some approximations when training sample sizes are smaller,
91 such as bolstering the training set with data from “supplemental locations” (Hamill et al. 2017)
92 or pooling of training data over locations spanning large regions (Lowry and Glahn 1976). With
93 newly available global reforecasts from version 12 of the NWS Global Ensemble Forecast
94 System (Hamill et al. 2020, Zhu et al. 2020ab), there is a long-enough training data set that

95 such approximations are not necessary for surface temperature, and each grid point can be
96 processed using only that point's data for training. In particular, this study independently
97 evaluated the raw and post-processed forecasts over a set of $\frac{1}{2}$ -degree grid points in a domain
98 encompassing the contiguous US (CONUS). 2000-2018 T_{2m} forecast and reanalysis data were
99 used as training data, and the forecasts were validated during 2019. These multiple post-
100 processing methods were evaluated with common (root-mean-square error, bias) metrics as
101 well as those less commonly applied to weather predictions such as "Taylor diagrams" (Taylor
102 2001). The hope is that the results will guide the choice of algorithms in future operational
103 postprocessing decisions.

104 Below, section 2 discusses the data used in this study as well as the postprocessing
105 methods and the methods of evaluation. Section 3 provides results, and section 4 concludes.

106

107 **2. Data, post-processing, and evaluation methods.**

108

109 *a. Forecast data.*

110 Gridded T_{2m} reforecasts from the US National Weather Service Global Ensemble
111 Forecast System, version 12 (GEFSv12) were used in this study. The ensemble forecast
112 system was described in Zhou et al. (2021), the reforecast data were described in Guan et al.
113 (2021), and the reanalyses used to initialize the reforecasts were described in Hamill et al.
114 (2021). Briefly, v12 of the GEFS provides a major system upgrade; the ensemble prediction
115 system uses a new finite-volume dynamical core, there are major improvements to the
116 deterministic and stochastic physics, and the grid spacing has been refined to ~ 25 km.
117 Ensemble prediction skill is improved in many ways, as described in Zhou et al. (2021). The
118 real-time ensemble is accompanied by a reforecast data set spanning 2000-2019, which is
119 available for free download from Amazon Web Services, [https://noaa-gefs-
retrospective.s3.amazonaws.com/index.html](https://noaa-gefs-
120 retrospective.s3.amazonaws.com/index.html) . During this period, for each day at 00 UTC, a 5-

121 member reforecast ensemble was generated to +16 days lead. Once per week an 11-member
122 ensemble was generated to +35 days. For this simple study, we examined only the
123 deterministic control member from this reforecast ensemble. While data were available on a
124 1/4-degree grid, the data were subsampled to 1/2 degree, confined to a domain from 125 to 60
125 degrees west longitude and 20 to 50 degrees north latitude. We do not expect sub-sampling to
126 affect the results. The domain encompassed the contiguous US and included some of Mexico,
127 southern Canada, and the Carribean (Fig. 1). Only forecasts at grid points considered > 50%
128 land in both the forecast and analyzed data were considered, as there were some oddities
129 where the forecast and analyzed data differed in their land-water classifications, and this
130 profoundly affected the statistics. Forecasts were evaluated from +12 h to +120 h lead time in
131 time steps of 12 h.

132

133 b. *Analysis data.*

134 Coincident “ERA5” reanalyses (Hersbach et al. 2020) from the European Centre for
135 Medium Range Weather Forecasts (ECMWF) / Copernicus Climate Service reanalysis were
136 downloaded and used for statistical model training and validation. The data were extracted on
137 a 1/4-degree grid and sub-sampled to the 1/2-degree grid, coincident in space with the forecast
138 grid. Data were extracted at 12 h intervals from the beginning of 2000 to the end of January
139 2020. ERA5 employs a T_{2m} analysis procedure using station observations, and it was thus
140 deemed to be a reasonably trustworthy gridded reference product.

141

142 c. *The decaying-average bias correction.*

143 This method will be abbreviated as “DAV” hereafter. The method has previously been
144 described in Cui et al. (2012). The approach is quite simple, both algorithmically and in terms
145 of implementation. The application developer chooses a value α that determines the weighting

146 to apply to the most recent discrepancy between forecast and observation (or analysis). For a
147 forecast date t for a particular forecast lead time and grid point, the DAV bias estimate is

148

$$149 \quad \hat{b}_t^{DAV} = (1 - \alpha) \hat{b}_{t-1}^{DAV} + \alpha(f_t - a_t), \quad (1)$$

150

151 where \hat{b}_{t-1}^{DAV} is the bias estimate at the same lead time and grid point but one day previous, and
152 f_t is the sample forecast value of a random variable X_f and a_t is the sample analyzed value at
153 date t . Some particularly appealing characteristics of DAV are: (a) training may be conducted
154 on-the-fly; one need not conduct a separate training, followed by validation. (b) Because of this,
155 storage of training data in an operational environment is not necessary. When considering
156 high-resolution grids over large areas and spanning multiple forecast variables, multiple lead
157 times, and lengthy training periods, this storage can become quite large. Some disadvantages
158 of the DAV method were discussed in Hamill (2018), in particular the difficulty in choosing an
159 optimal value of α in the presence of time-varying unconditional bias.

160 The error of the DAV method was only slightly sensitive to the chosen value of α . Fig. 2
161 shows the RMSE of the DAV method during the 2000-2018 training period as a function of α .
162 For the official validation in 2019 against other techniques, the α that produced the lowest
163 RMSE at each forecast lead time during the training period was chosen.

164

165 *d. Quantile mapping.*

166 Let the cumulative distribution function (CDF) for the forecast at a particular grid point
167 location and time be denoted by

168

$$169 \quad F_f(V) = P(X_f \leq V), \quad (2)$$

170

171 where X_f is again the temperature forecast random variable at time t , and V is a specific
172 temperature value. The $0.0 \leq F_f(X_f) \leq 1.0$. We define a quantile function that maps a
173 cumulative probability p back to the forecast temperature variable:

174

$$175 \quad X_f = F_f^{-1}(p) . \quad (3)$$

176

177 The quantile-mapping (QM) procedure thus maps the forecast temperature sample to its
178 cumulative probability in the climatological distribution of forecasts and then applies the quantile
179 function (also known as the percent-point function) to this cumulative probability associated with
180 the climatological distribution of analyzed data:

181

$$182 \quad \hat{a}_t^{QM} = F_a^{-1} \left[F_f(f_t) \right] . \quad (4)$$

183

184 In this way it estimates an analyzed value sharing the same cumulative probability relative to its
185 analyzed climatological distribution as the sample forecast value to the climatological forecast
186 distribution. The bias estimate is then $\hat{b}_t^{QM} = f_t - \hat{a}_t^{QM}$.

187 The CDFs for forecast and analyses at many grid points were characteristic of non-
188 Gaussian distributions. After some experimentation, a 3-component Gaussian mixture model
189 was chosen to represent the CDFs instead of a one-component Gaussian or other parametric
190 distribution. It used the python module `scikit-learn.mixture`. This module determined weights,
191 means, and standard deviations associated with three Gaussian kernels whose weighted sum
192 provided the closest fit to the empirical distributions of forecasts (or analyzed) data. An
193 example of the fitted distributions and P-P plots (Wilks 2011, sec. 4.5.2) are provided in Figs. 3
194 and 4, respectively. At this grid point and at many others examined, the fitted CDFs appear to
195 produce highly accurate parametric representations of the empirical CDFs. Different

196 distributions were estimated for each grid point, forecast month, and lead time using 2000-2018
197 data and the month of interest including the data from +/- 1 month. For example, +120 h CDFs
198 for the month of February are fit with January - February - March 2000-2018 training data. See
199 Wilks (ibid) for interpretation of the Q-Q plots.

200

201 *e. Univariate MOS.*

202 Univariate MOS, or uMOS hereafter, is an application of simple linear regression to bias
203 correction. This assumes that an estimate of the analyzed temperature may be determined
204 through a regression equation of the form

205

$$206 \quad \hat{a}_t^{uMOS} = c_0 + c_1 f_t \quad (5)$$

207

208 where c_0 and c_1 are the fitted intercept and slope. The error (or residual) $e_t = a_t - \hat{a}_t$ is
209 commonly assumed to be normally distributed with zero mean. In practice, non-linear
210 relationships and heteroscedasticity were present at many grid points, as will be discussed in
211 the results, but for generality, no grid-point specific remedial measures such as power
212 transformation of data were employed. Linear regression is reviewed in many texts, including
213 Wilks (2011, section 7.2.1). As with the QM, separate regression equations were fit for each
214 grid point at each forecast lead time, month by month, using 2000-2018 training data and a 3-
215 month period centered on the month of interest. The implicit bias estimate was thus

$$216 \quad \hat{b}_t^{uMOS} = f_t - \hat{a}_t^{uMOS}.$$

217

218 *f. Multi-variate MOS.*

219 Multi-variate MOS using multiple forecast variables as predictors has a long heritage in
220 the US National Weather Service (Glahn and Lowry, Carter *et al.* 1989) and in many other

221 forecast agencies. Commonly, multiple forecast fields including variables above the surface are
222 used as additional predictors, variables such as forecast cloud cover, thicknesses between
223 pressure levels, and so forth. Application of this approach with the data at hand could be more
224 challenging, for a screening regression approach to the selection of predictors might result in
225 different predictor choices for different parts of the domain. Rather than use this approach with
226 its training and data management complexity, multi-variate here implies something slightly
227 different; instead of multiple forecast variables, the bias corrections from other approaches are
228 used as predictors. Specifically, we estimate the analyzed state with a regression equation of
229 the form

230

$$231 \quad \hat{a}_t^{mvMOS} = c_0 + c_1 f_t + c_2 \hat{b}_t^{DAV} + c_3 \hat{b}_t^{QM}. \quad (6)$$

232

233 This allows us to determine whether a method that uses information from alternative bias
234 correction approaches may improve the forecasts. No interaction terms were included.

235 Previously, multi-method synthesis showed promise for probabilistic forecasts (Möller and Groß
236 2016, Bassetti et al. 2017, Yang et al. 2017, Baran and Lerch 2017). The implicit bias correction
237 is then $\hat{b}_t^{mvMOS} = f_t - \hat{a}_t^{mvMOS}$. The training data periods were the same as with uMOS, but a
238 first sweep through the data was necessary to generate the DAV and QM bias estimates.

239

240 g. *Verification methods.*

241 Commonly used verification methods will be applied, focusing on error and bias. Root-
242 mean square error (RMSE) statistics will be provided. For a bias-correction method with n
243 samples, the estimated RMSE is

244

245
$$RM\hat{S}E = \left\{ \frac{1}{n-1} \sum_{i=1}^n [\hat{a}_t(i) - a_t(i)]^2 \right\}^{1/2}$$
 (7)

246

247 and the estimated unconditional mean bias (BIA) is

248

249
$$B\hat{I}A = \frac{1}{n} \sum_{i=1}^n [\hat{a}_t(i) - a_t(i)]$$
 (8)

250

251 Differences of RMSE for the various bias-correction methods were evaluated relative to the
 252 DAV method. 5th and 95th percentile confidence intervals for these RMSE differences were
 253 generated with the paired block-bootstrap procedure described in Hamill (1999); 100
 254 resamplings were performed.

255 Taylor diagrams were also generated as a way of understanding the forecast
 256 characteristics (Taylor 2001, Wilks 2011, sec. 8.6.3). These diagrams were plotted in polar
 257 coordinates. The radial distance from the origin represented the ratio of the climatological
 258 standard deviations of forecast vs. analyses. This was the mean forecast variability divided by
 259 the mean analyzed variability, where variability measured the standard deviation of the sample.
 260 The angle, computed clockwise from the 12 o'clock position, represented the forecast vs.
 261 observed correlation. For this application of Taylor diagrams, a sample will be plotted for each
 262 forecast grid point, so that the potential variability of the error decomposition across the domain
 263 can be examined.

264

265 **3. Results.**

266 Figure 5 provides RMSE and BIA statistics averaged over all land points within the
 267 domain. The DAV method provided a statistically significant decrease in RMSE relative to the
 268 raw guidance, and this improvement in skill amounted to 1-3 days gain lead time; for example,

269 the +3 day DAV forecasts were nearly as skillful as the +1 day raw forecasts. QM generally
270 produced forecasts with slightly higher RMSE than DAV, but the uMOS and mvMOS forecasts,
271 especially after +1.5 days, provided a significant reduction in RMSE relative to DAV. Overall,
272 the mvMOS forecasts produced the lowest domain-average errors, and this was generally true
273 across seasons (not shown). Domain-average biases were reduced in all methods, but the
274 DAV method produced forecasts with the lowest bias. Apparently the bias characteristics of
275 2000-2018 were somewhat dissimilar to those of 2019, for QM, uMOS, and mvMOS all
276 demonstrated slight warm biases, which should not be evident were the bias characteristics the
277 same in the training and validation periods. The uMOS and mvMOS results were slightly lower
278 in error when the forecast data predictor was changed to be a deviation from climatology (not
279 shown). However, to facilitate more direct comparison against the other methods, only the
280 results using the unmodified forecasts are presented.

281 The reduced error of mvMOS is an interesting result, supported by other literature
282 (Möller and Groß 2016, Bassetti et al. 2017, Yang et al. 2017, Baran and Lerch 2018). Different
283 bias-correction methods have different strengths. To the extent that bias is less dependent on
284 whether the forecast temperature is comparatively more warm or cold but instead more
285 dependent on local discrepancies, then DAV performs well. The QM method does not attempt
286 to minimize error but seeks the analyzed value associated with today's quantile in the forecast
287 distribution. This may result in large mappings if the CDFs are different in character. The
288 MOS methods by design minimize RMSE, but as will be discussed, at the expense of other
289 forecast characteristics.

290 How responsive were the various statistical adjustment techniques to changes of
291 weather? As an example, time series of +24h forecast data for a grid point near Boulder, CO,
292 USA are presented in Fig. 6. The top panel displays a time series of the +24 h lead GEFSv12
293 forecast and ERA5 analyzed data, as well as the ERA5 climatology, fitted with cubic splines as
294 in Hamill and Scheuerer (2020). The bottom panel shows a corresponding time series of the

295 various bias-correction methods. The 1-day lag autocorrelation coefficients are provided in the
296 inset legend. Per its design, the DAV method changed the least from day to day, with the
297 largest autocorrelations. The other methods' bias corrections were more responsive to the
298 weather.

299 Let's consider more closely the rapid oscillations of the regression methods during July
300 in Fig. 6(b). Figure 7(a) shows the CDFs used in the quantile-mapping function at this location.
301 The CDFs aligned closely with each other, so the mappings were quite modest, and only
302 modest changes in bias estimates occurred from day to day during this month; however, in
303 other months the QM corrections were more weather sensitive, such as in February. In
304 contrast, the regression methods produced differing corrections for different forecast
305 temperatures at this grid point during July. Figure 7(b) shows the 2000-2018 training data for
306 this location as well as the fitted uMOS regression curve. In this case, the one-size-fits-all
307 regression approach, with no remedial measures to address issues such as heteroscedasticity,
308 appeared to be a model shortcoming. The training data were in fact heteroscedastic, with larger
309 differences between forecasts and observations at lower temperatures. Further, the marginal
310 distributions showed that the underlying data were multi-modal in nature, with peak probability
311 density at the higher temperatures; because of the larger number of samples with higher
312 temperatures, the regression fit was more closely optimized to these samples. As a
313 consequence, the regression model did not appear to provide a high-quality fit at the lower
314 temperatures; in this instance when the forecast temperatures were comparatively low, the
315 regression model predicted a cold forecast bias. The actual (forecast, analyzed) samples for
316 July 2019 were presented in Fig. 7(b) as the bolder red points, several of which have colder
317 forecast temperatures and predicted cold biases based on the regression line. With a daily
318 change in forecast temperature from warm to cold, there was a corresponding change in the
319 estimated forecast bias from too warm to too cold, and hence large oscillations occurred with
320 the change in forecast temperatures from one day to the next.

321 Despite its challenges exhibited in Fig. 7, the MOS methods did produce comparatively
322 lower RMSE on average, but what other forecast characteristics did they have relative to the
323 other methods? This can be examined in part with Taylor diagrams (Taylor 2001, Wilks 2011,
324 sec 8.6.3). Figure 8 provides such diagrams for the +24 h forecasts for raw, DAV, QM, and
325 mvMOS methods during the Jul-Aug-Sep 2019 period. See the references above for more
326 interpretation of these diagrams. Differently colored dots denote the magnitude of the analysis
327 standard deviation, i.e., the red dots denoted locations with little weather variability during the
328 sample period while the brown dots were locations with the most weather variability.

329 The raw forecasts exhibited much scatter in the Taylor diagram standard deviation ratio,
330 sometimes with the forecast sample during this season having more variability than the
331 analyzed data, and sometimes less. These variations in the standard deviation ratio were
332 muted only somewhat with the DAV method. The QM method, consistent with its goal of
333 producing mappings that represented draws from the analyzed climatology, had a narrower
334 range of standard deviation ratios that were more concentrated around the 1.0 ratio. The
335 practical effect of this as a forecast procedure is that *this method retains more of the variability*
336 *in the observations*. In contrast, the mvMOS procedure, especially for the forecasts at locations
337 with smaller analyzed weather variability, produced less variability in the corrected forecasts
338 than in the analyzed, as denoted by the ratio that on average was lower than 1.0, especially
339 when climatological variability was small. It is possible that a human forecaster, say, seeking to
340 predict the magnitude of a warm or cold event, might prefer the QM guidance relative to one of
341 the MOS procedures' guidance, given that the former retained more of the synoptic-scale
342 variability.

343

344 4. Conclusions.

345 This brief study provided an intercomparison of statistical postprocessing methods
346 applied to deterministic surface-temperature forecasts on a ½-degree grid over the CONUS and

347 surrounding land regions out to +5 days lead time. The control member from the Global
348 Ensemble Forecast System version 12 reforecast data were used to provide the forecasts,
349 2000-2018 for training and 2019 for validation. ECMWF reanalyses, specifically ERA5, were
350 used for training and validation. The four methods that were considered were the decaying-
351 average bias correction (DAV), quantile mapping (QM), a univariate Model Output Statistics
352 (uMOS), otherwise known as linear regression, and a multi-variate MOS (mvMOS) that used the
353 surface-temperature forecasts as well as bias estimates from DAV and QM as predictors.
354 Except at the earliest leads, the MOS techniques produced forecasts with the lowest error with
355 mvMOS providing errors lower than uMOS. Through an examination of Taylor diagrams, it was
356 revealed that while the mvMOS reduced the error, especially at locations with low climatological
357 variability across a season, it also reduced the variability in the post-processed forecasts
358 relative to the raw guidance. On the other hand, QM and DAV methods retained much of the
359 seasonal variability in the raw forecasts. Which method a forecaster may prefer could depend
360 on whether they are optimizing for RMSE (choose a MOS method) or for realistic prediction of
361 the magnitude of unusual events (choose DAV or QM). The DAV method produced bias
362 corrections that were more consistent in time, while the QM and MOS techniques were more
363 sensitive to the weather of the day.

364 A main conclusion is that because different post-processing methods may have differing
365 strengths and weaknesses, the judicious combination of them may be able to, in some metrics,
366 provide guidance that is improved relative to any one on its own. In particular, the mvMOS
367 method here, which combined DAV, QM, and MOS approaches, produced guidance with the
368 lowest RMSE. Since each postprocessing method is relatively straightforward to implement, an
369 operational combination of these could be a practical solution that would provide improved
370 guidance for many customers.

371 This study was not comprehensive; it considered only an area around the US, and it
372 used a long training data set and considered only surface temperature, not other variables of

373 interest such as winds or cloud cover or precipitation. Nonetheless, the optimistic results,
374 confirmed by other supporting literature, suggest that the judicious combination of multiple post-
375 processing methods may provide a practical way to reduce errors with modest effort.

376

377 **Acknowledgments:** ERA5 reanalysis data from the Copernicus Climate Service were used in
378 this study. Publication costs were provided by the NOAA Weather Program Office grant
379 U8R2WRE-P00.

380

381 **References**

382

383 Baran, S. & S. Lerch, 2018: Combining predictive distributions for the statistical post-processing
384 of ensemble forecasts. *Int. J. Forecasting*, **34 (3)**, 477-496.

385 Bassetti, F., R. Casarin, and F. Ravazzolo, 2017: Bayesian nonparametric calibration and
386 combination of predictive distributions. *J. Amer. Stat. Assoc.*, doi:
387 [10.1080/01621459.1273117](https://doi.org/10.1080/01621459.1273117).

388 Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the
389 National Meteorological Center's numerical weather prediction system. *Wea.*
390 *Forecasting*, **4**, 401–412.

391 Craven, J.P., D.E. Rudack, and P.E. Shafer, 2020: National Blend of Models: A statistically
392 post-processed multi-model ensemble. *J. Operational Meteor.*, **8 (1)**, 1-14, doi:
393 <https://doi.org/10.15191/nwajom.2020.0801>

394 Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea.*
395 *Forecasting*, **27**, 396–410, <https://doi.org/10.1175/WAF-D-11-00011.1>.

396 Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective
397 weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.

398 Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The Gridding of MOS.
399 *Wea. Forecasting*, **24**, 520–529, <https://doi.org/10.1175/2008WAF2007080.1>.

400 Guan, H., and others, 2021: The GEFsv12 reforecast dataset for supporting sub-seasonal and
401 hydrometeorological applications. In preparation.

402 Hamill, T. M., 1999: [Hypothesis tests for evaluating numerical precipitation forecasts](#). *Wea.*
403 *Forecasting*, **14**, 155-167.

404 Hamill, T.M., E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: [The U.S.](#)
405 [National Blend of Models for Statistical Postprocessing of Probability of Precipitation and](#)
406 [Deterministic Precipitation Amount](#). *Mon. Wea. Rev.*, **145**, 3441-3463,
407 <https://doi.org/10.1175/MWR-D-16-0331.1>

408 Hamill, T. M., and Scheuerer, M., 2018: [Probabilistic precipitation forecast postprocessing using](#)
409 [quantile mapping and rank-weighted best-member dressing](#). *Mon. Wea. Rev.*, **146**,
410 4079-4098. Also: [Online appendix 1](#).

411 Hamill, T. M., 2018: [Practical Aspects of Statistical Postprocessing](#). Chapter 7 in *Statistical*
412 *Postprocessing of Ensemble Forecasts* (Elsevier Press).

413 Hamill, T. M., and others, 2021: [The reanalysis for the Global Ensemble Forecast System,](#)
414 [version 12](#). *Mon. Wea. Rev.*, submitted.

415 Hamill, T. M. , and M. Scheuerer, 2020: [Improving ensemble weather prediction system](#)
416 [initialization: disentangling the contributions from model systematic errors and initial](#)
417 [perturbation size](#). *Mon. Wea. Rev.*, in press. Also: [online Appendix](#)

418 Hersbach, H., B. Bell, P. Berrisford, et al., 2020: The ERA5 global reanalysis. *Quart. J Royal.*
419 *Meteor. Soc.*, 146, 1999 - 2049. <https://doi.org/10.1002/qj.3803>

420 Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major
421 river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**,
422 618-641. doi: <https://doi.org/10.1175/2009JHM1006>.

423 Lowry, D. A., and H. R. Glahn, 1976 : An operational model for forecasting probability of
424 precipitation—PEATMOS PoP. *Mon. Wea. Rev.*, **104**, 221-232.

425 Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation
426 issue. *J. Climate*, 26, 2137- 2143. doi: <https://doi.org/10.1175/JCLI-D-12-00821.1>

427 Möller, A., and J. Groß, 2016: Probabilistic temperature forecasting based on an ensemble AR
428 modification. *Quart. J. Royal Meteor. Soc.*, **142**, 1385-1394.

429 Taylor, K. E., 2001: Summarizing multiple aspects of forecast performance in a single diagram.
430 *J. Geophys. Res.*, **D106**, 7183-7192.

431 Vannitsem, S., D. S. Wilks, and J. W. Messner, 2018: *Statistical Postprocessing of Ensemble*
432 *Forecasts*, Vannitsem, Wilks, and Messner, eds. Elsevier Press, 347 pp.

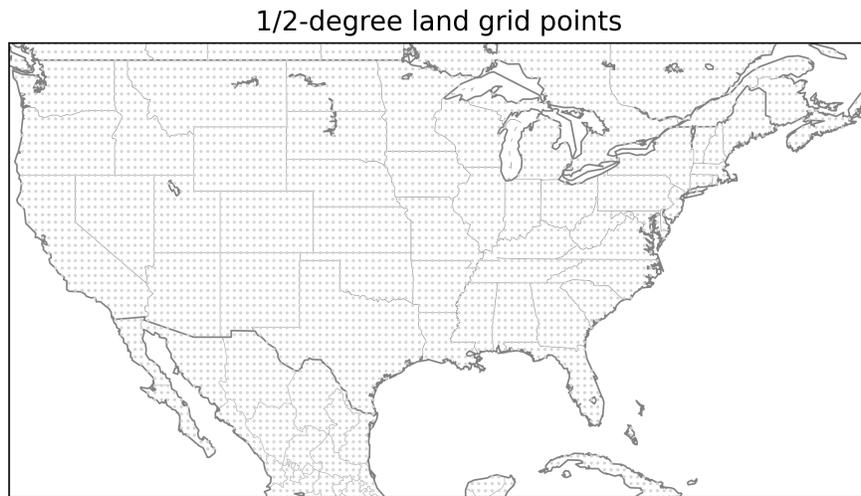
433 Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010 : Calibration and downscaling methods
434 for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, 25, 1603-1627.
435 doi:<https://doi.org/10.1175/2010WAF2222367.1>.

436 Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences (3rd Ed.)*. Academic
437 Press, 676 pp.

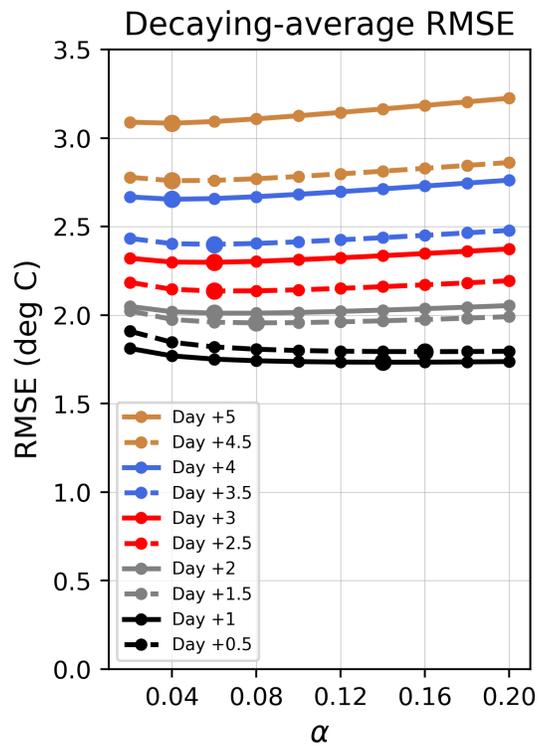
438 Yang., X., S. Sharma, R. Siddique, S. J. Greybush, , and A. Mejia, 2017: Postprocessing of
439 GEFS precipitation ensemble reforecasts over the US Mid-Atlantic region. *Mon. Wea.*
440 *Rev.*, **145**, 1641-1658.

441 Zhou, X., and others 2021: The Introduction of the NCEP Global Ensemble Forecast System
442 Version 12, in preparation.

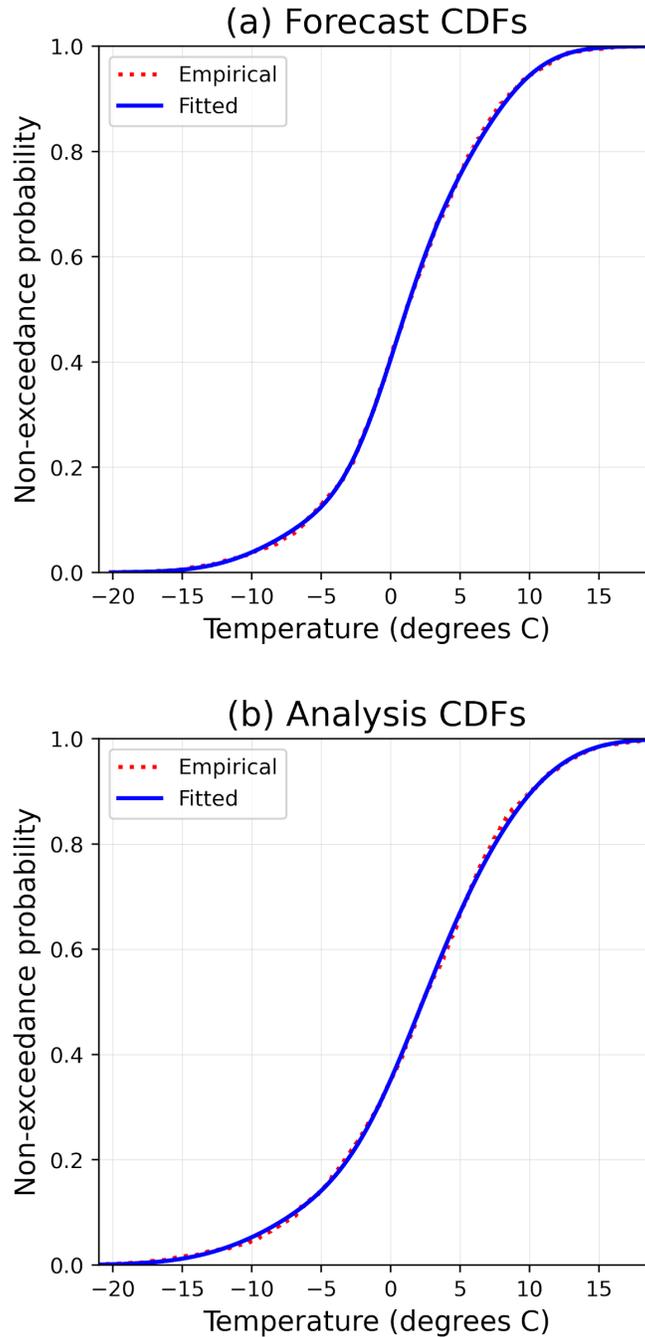
443



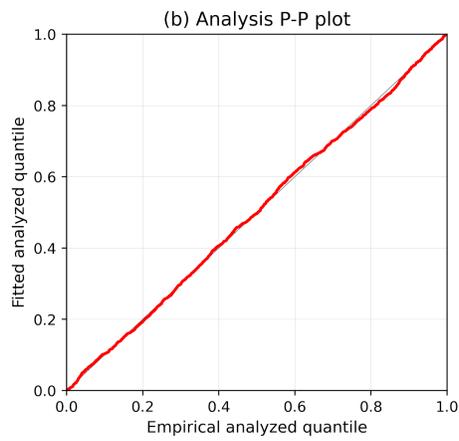
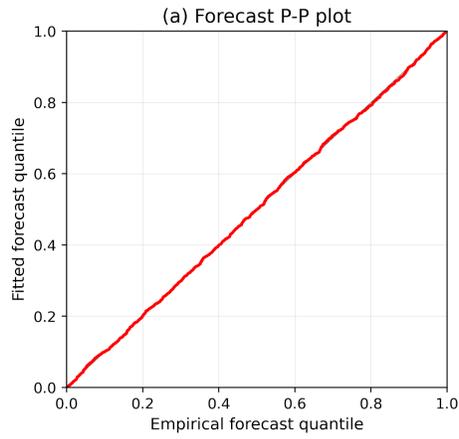
444
 445 **Figure 1.** Grid points (dots) considered for evaluation of postprocessing methods in this study.
 446



447
 448 **Figure 2:** Root-mean-square error of the decaying-average bias correction method as a function
 449 of α for various forecast lead times. The larger dot denotes the value with the lowest error in
 450 the training period.



451
 452 **Figure 3.** Examples of empirical (dashed red, underneath) and fitted (blue, overtop) CDFs
 453 estimated with a 3-component Gaussian mixture, here for January data at +24 h lead time near
 454 Boulder, CO, USA (105° west longitude, 40° north latitude). (a) 2-m temperature forecast data,
 455 and (b) corresponding analysis data.



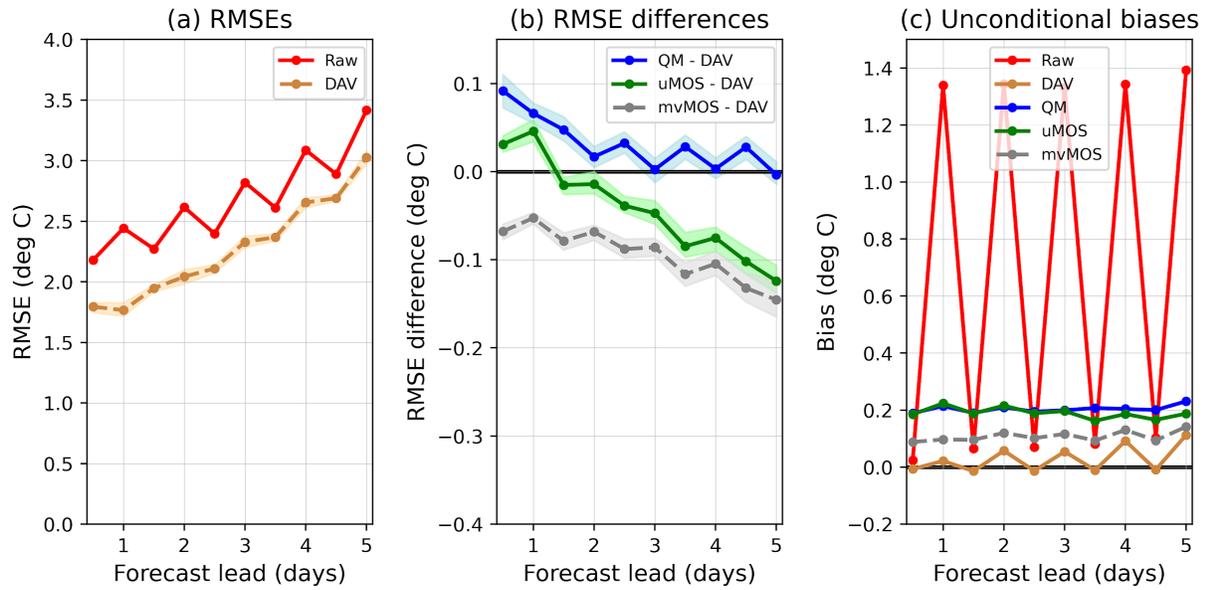
456

457

458

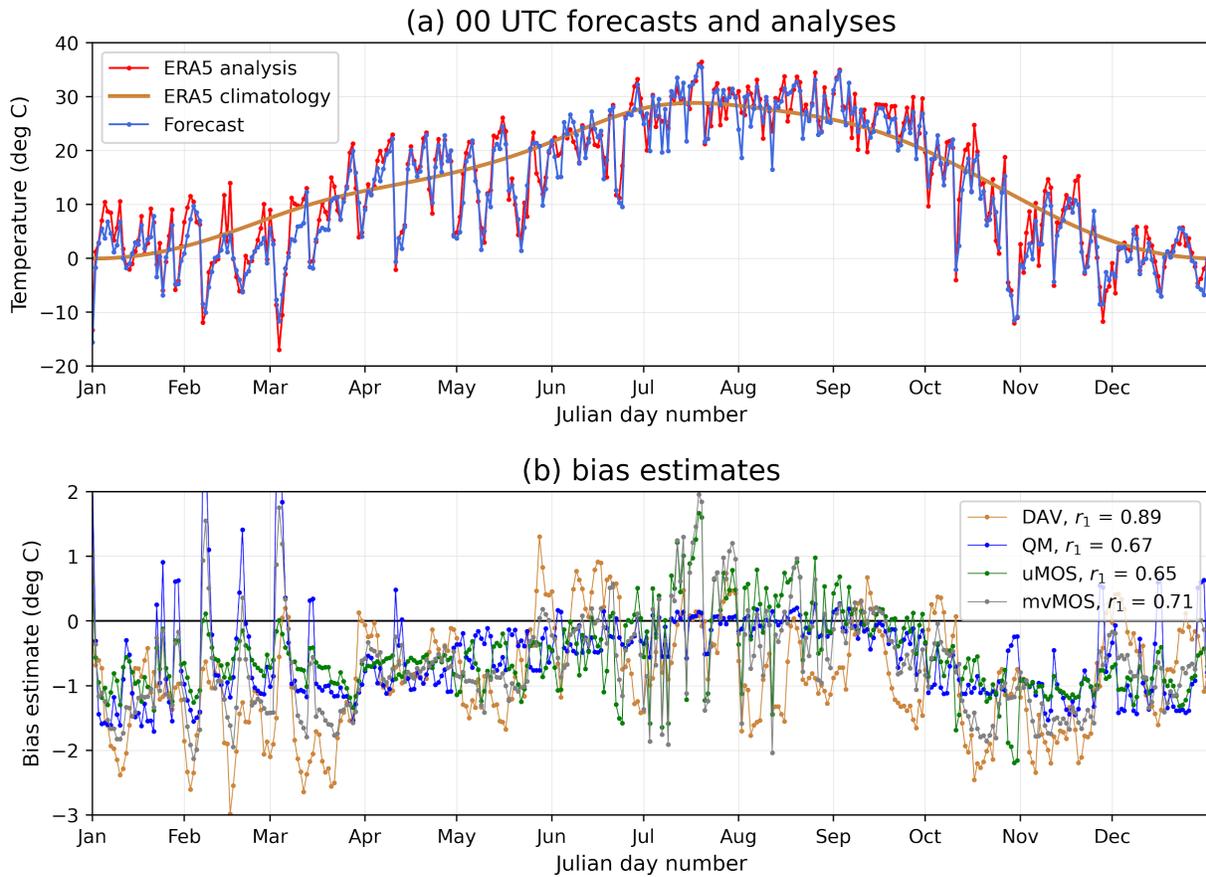
Figure 4: Probability-probability (P-P) plots corresponding to the data in Fig. 2.

459
460
461



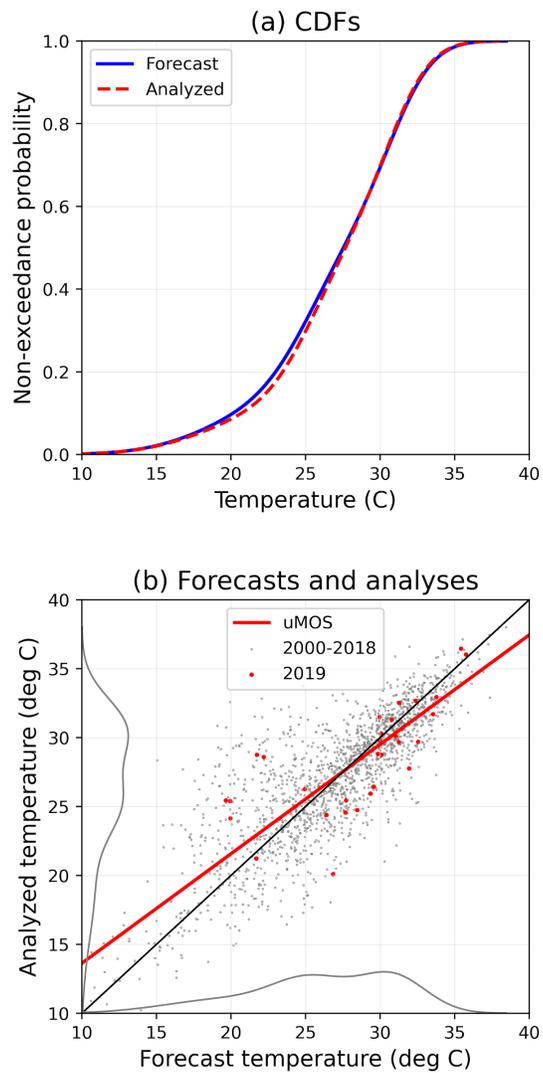
462
463
464
465
466
467
468
469
470
471
472

Figure 5: Domain-averaged errors of raw forecasts and various bias-correction methods. (a) RMSE of raw (solid red) and DAV (dashed brown) bias correction methods as a function of forecast lead time. 5th and 95th percentile confidence interval of differences between the two forecasts are plotted as light brown around the DAV method. (b) RMSE differences of the QM - DAV method (blue; lower is an improvement over DAV), uMOS (green), and mvMOS (gray). Confidence intervals are plotted in lighter-shade colors as in panel (a), but here the confidence intervals represent differences with respect to the DAV method. (c) Unconditional bias for raw forecasts and the various bias correction methods.

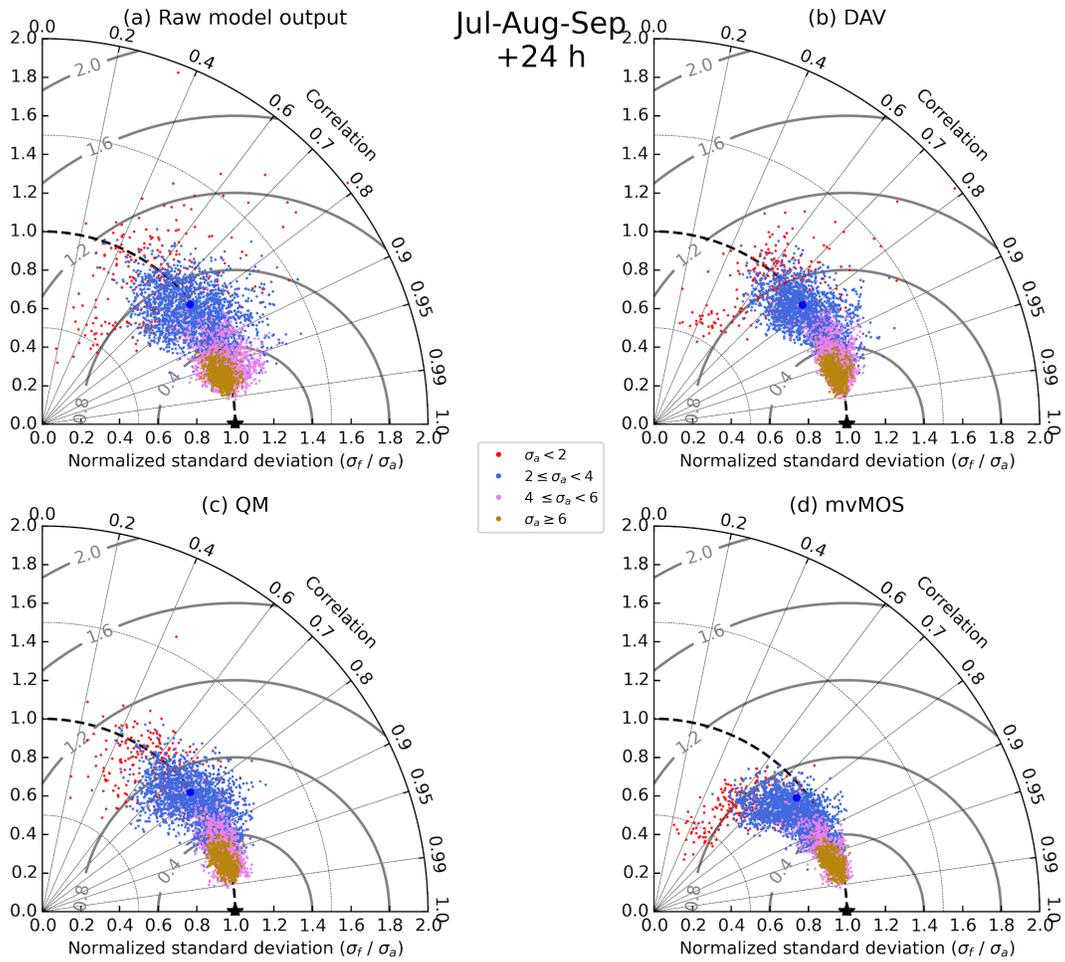


473
474
475
476
477
478
479

Figure 6: +24 h forecast and ERA5 analyzed time series of 00 UTC data for a grid point near Boulder, CO, USA during 2019. (a) ERA5 analyses (red) and GEFSv12 forecasts (blue). (b) Bias estimates from various methods. One-day lag autocorrelations are provided in the inset legend.



480
 481 **Figure 7:** (a) Fitted cumulative distribution functions (CDFs) used in the quantile-mapping
 482 procedure for +24 h lead forecasts at a grid point near Boulder, CO, USA during July. (b)
 483 Scatterplot of +24 analyzed vs. forecast analyzed 2000-2018 training data for July at Boulder
 484 CO (small gray dots) and marginal probability density functions (gray lines along each axis).
 485 uMOS fitted linear regression line is presented in red, and the 2019 (forecast, analyzed) pairs
 486 are shown as the larger red dots.



487
 488
 489
 490
 491
 492
 493
 494
 495
 496
 497

Figure 8: Taylor diagrams for July-August-September and +24 h lead time for (a) raw forecasts, (b) DAV, (c) QM, and (d) mvMOS. A sample from each land grid point is plotted as a separate dot. The radial magnitude indicates the ratio of the sample forecast standard deviation during the season divided by the sample analysis standard deviation. Correlation increases clockwise from the 12 o'clock position (0.0) to the 3 o'clock position (1.0). Gray lines denote lines of equal standardized RMSE. Individual dots are colored by that grid point's sample analysis standard deviation σ_a . The dots' color legend is provided in the plot center.