

1
2
3 **Verification of TIGGE Multi-model and ECMWF**
4 **Reforecast-Calibrated Probabilistic Precipitation**
5 **Forecasts over the Conterminous US**
6
7

8 Thomas M. Hamill

9
10 *NOAA Earth System Research Laboratory, Physical Sciences Division*

11 *Boulder, Colorado USA*
12

13
14
15 Submitted as an article

16
17 to

18
19 *Monthly Weather Review*
20

21
22
23 17 August 2011
24
25
26
27
28
29
30
31
32
33
34

35 Corresponding author address:

36
37 Dr. Thomas M. Hamill
38 NOAA ESRL, Physical Sciences Division
39 R/PSD 1
40 325 Broadway
41 Boulder, CO 80305-3328
42 tom.hamill@noaa.gov
43 phone: (303) 497-3060 fax: (303) 497-6449

44
45 ABSTRACT
46

47 Probabilistic quantitative precipitation forecasts (PQPFs) were generated
48 from the THORPEX Interactive Grand Global Ensemble (TIGGE) database during July
49 to October 2010. 24-hour accumulated precipitation forecasts were evaluated at 1-
50 degree grid spacing within the conterminous US. Probabilistic forecasts were also
51 generated from ECMWF ensemble system with statistical calibration using
52 reforecasts from the period 2002-2009 and from the multi-model forecasts using
53 the last 30 days of forecasts.

54 ECMWF's EPS forecasts generally had the highest skill, followed by CMC;
55 UKMO and NCEP were the least skillful. CMC forecasts were the most reliable but
56 the least sharp, and NCEP and UKMO forecasts were the least reliable but sharper.

57 Multi-model forecasts were more reliable and skillful than individual EPS
58 forecasts. The improvement was larger for events heavier precipitation events such
59 as $> 10 \text{ mm } 24 \text{ h}^{-1}$ than for smaller events like $> 1 \text{ mm } 24 \text{ h}^{-1}$.

60 ECMWF ensembles were statistically post-processed using the five-member
61 weekly reforecasts for the June - November period of 2002-2009, the period where
62 precipitation analyses were also available. The post-processed ECMWF forecasts
63 were much more skillful and reliable for the heavier precipitation events than raw
64 forecasts, but much less sharp. Multi-model forecasts were generally more skillful
65 than reforecast-calibrated ECMWF forecasts for the light precipitation events but
66 about the same skill for the higher precipitation events; also, they were sharper but
67 somewhat less reliable than ECMWF post-processed ensembles.

68 The evidence presented here suggests that all operational centers, even
69 ECMWF, would benefit from the open sharing of precipitation forecast data.

70

71 1. Introduction

72

73 An ongoing challenge with short- and medium range ensemble prediction
74 systems (EPSes) is how to generate probabilistic forecasts that account for the
75 system errors in the ensemble. System errors include sampling error due to the
76 finite ensemble size and the error introduced by model imperfections such as the
77 grid truncation and the use of deterministic parameterizations (Houtekamer and
78 Mitchell 2005). There are many methods for treating system error, from
79 introducing stochastic aspects into the ensemble prediction system (Buizza et al.
80 1999, Shutts 2005, Berner et al. 2009, Palmer et al. 2009, Charron et al. 2010), using
81 multiple parameterizations (Charron et al. 2010, Berner et al.), using multiple
82 models (Bougeault et al. 2010), and statistical post-processing.

83 Two methods that will be explored and contrasted here are the multi-model
84 methods and statistical post-processing. The underlying hypothesis of multi-model
85 ensembles (Krishnamurti et al. 2000, Wandishin et al. 2001, Mylne et al. 2002,
86 Doblas-Reyes et al. 2005, Hagedorn et al. 2005, Weigel et al. 2008, Johnson and
87 Swinbank 2009, Bougeault et al. 2010, Iversen et al. 2011) is that the many
88 differences between constituent EPSs will result in them generating ensemble
89 forecasts with quasi-independent systematic errors, so the combination may result
90 in a more accurate estimate of the uncertainty. Practically, also, these are
91 ensembles of opportunity. If all centers are willing to share rather than sell their
92 forecast data, the additional members can be used for only the cost of data
93 transmittal and storage, so they may provide an inexpensive way to improve

94 forecast skill. However, there are some potential disadvantages of multi-model
95 ensembles. Developing an accurate, stable weather prediction system is costly, so
96 multi-model ensembles are likely to be less useful when formed from immature
97 systems. System outages may prevent routine access to other centers' ensembles.
98 One or other of the models is likely to have been changed recently, rendering it
99 difficult to understand the multi-model system error characteristics. Also, the
100 hypothesis of quasi-independent errors may not always hold. Practically, each
101 operational center is interested in providing a high-quality product without
102 depending on another center's data. When another center develops a method that
103 improves the forecast significantly, it may be adopted at other operational centers.
104 The similarity could result in some co-linearity of errors and decreased collective
105 usefulness (Lorenz et al. 2011).

106 Another method for addressing system error is through statistical post-
107 processing. Discrepancies between time series of past forecasts from a fixed model
108 and the verifying observations/analyses can be used to modify the real-time
109 forecasts. For some variables such as short-range forecasts of surface temperature,
110 a short time series may be sufficient (Stensrud and Yussouf 2003, Yussouf and
111 Stensrud 2007, Hagedorn et al. 2008). For others such as heavy precipitation and
112 longer-lead forecasts, using a long time series of reforecasts has been shown to
113 dramatically improve the reliability and skill of the probabilistic forecasts (Hamill et
114 al. 2004, Hamill et al. 2006, Hamill and Whitaker 2007, Wilks and Hamill 2007,
115 Hamill et al. 2008). A drawback of using reforecasts is that a forecast time series
116 spanning many years or even decades may be necessary to produce a sufficiently

117 large sample to adjust for systematic errors in rare-event forecasts. Since forecast
118 models are regularly updated, which may change the systematic error
119 characteristics, either a forecast model must be frozen once a reforecast data set has
120 been generated, or a new reforecast data set must be generated every time the
121 modeling system changes significantly. Hence, reforecasting can be computationally
122 expensive and can restrict the ability of a forecast center to upgrade its system
123 rapidly. Recently, statistical post-processing methods have been the subject of
124 much investigation (Gneiting et al. 2005, Raftery et al. 2005, Sloughter et al. 2007,
125 Wilson et al. 2007, Vannitsem and Nicolis 2008, Glahn et al. 2009, Bao et al. 2010).

126 To date, however, there have been no systematic comparisons of multi-
127 model and reforecast-calibrated PQPFs verified over a large enough area and a long
128 enough period of time to confidently assess the relative strengths and weaknesses
129 of these two approaches. This study attempts to provide such a comparison for this
130 high-impact forecast parameter. Using TIGGE forecast data from NCEP, CMC, UKMO,
131 and ECMWF, multi-model ensemble 24-h accumulated probabilistic forecasts of
132 precipitation were generated and then compared against ECMWF forecasts that
133 were statistically adjusted using their reforecast data set. The comparison was
134 performed over the conterminous US (CONUS) during the period July-October 2010.
135 Statistical adjustments were also attempted with multi-model forecasts, trained on
136 the previous 30 days of forecasts and analyses.

137 Below, section 2 describes the data sets used in this experiment, the
138 verification methodology, and the statistical post-processing method. Section 3
139 provides results, and section 4 some conclusions.

140

141 2. Data sets and methods.

142

143 *a. Analysis data used.*

144 A recently created precipitation data set, NCEP’s Climatology-Calibrated
145 Precipitation Analysis (CCPA), was used for verification. The CCPA attempts to
146 combine the relative advantages of the 4-km, hourly NCEP Stage-IV precipitation
147 analysis (Lin and Mitchell 2005), and the daily, 0.25-degree NCEP Climate Prediction
148 Center (CPC) Unified Precipitation Analysis (Higgins et al. 1996). The former is
149 based on gage and radar data, the latter solely on gage data. A disadvantage of the
150 Stage-IV product is that it may inherit some of the biases due to the estimation of
151 rainfall from radars. A disadvantage of the CPC product is that there are areas of the
152 US that are only sparsely covered by gage data. The CCPA analysis regressed the
153 Stage-IV analysis (the predictor) to the Higgins et al. CPC analysis (the predictand),
154 thereby reducing bias with respect to the in-situ observations. In several of the
155 driest locations in the western US, the CCPA analysis was set to missing, for the
156 regression analysis was untrustworthy and singular due to no precipitation in either
157 analysis product. In such cases, the CCPA analysis for this study was simply
158 replaced with the Stage-IV analysis. For our purposes, we used CCPA analyses that
159 also were upscaled to 1 degree and accumulated over a 24-h period in a manner that
160 preserved total precipitation, similar to the “remapping” procedure described in
161 Accadia et al. (2003). The CCPA analyses were available from 2002 – current, a
162 shorter period than the ECMWF reforecasts, thus limiting the amount of training
163 data that could be used in the statistical post-processing.

164

165 *b. Forecast model data.*

166 For this experiment, 20 perturbed member forecasts of 24-h accumulated
167 precipitation were extracted from the UKMO, CMC, NCEP, and ECMWF ensemble
168 systems archived in the TIGGE database at ECMWF. Probabilities were calculated
169 directly from the ensemble relative frequency. The forecast period was July to
170 October 2010; only 00 UTC initial time forecasts were extracted in order to allow
171 comparison with a post-processed forecasts using ECMWF's reforecasts, which were
172 generated only from 00 UTC initial conditions. Daily forecasts were examined from
173 +1 to +5 day lead. Regardless of the original model resolution, all centers' forecasts
174 were bi-linearly interpolated to a 1-degree latitude-longitude grid covering the
175 CONUS using ECMWF's TIGGE portal software. ECMWF's interpolation procedure
176 set the amount to zero if there was no precipitation at the nearest neighboring point
177 and the interpolated value was less than 0.05 mm. No control forecasts were
178 included, just the forecasts from the perturbed initial conditions. Other forecast
179 centers' contributions to the TIGGE archive were not used here for various reasons,
180 such as the unavailability of 00 UTC forecasts from the Japan Meteorological Agency.
181 For size consistency and to facilitate skill comparisons, only the first 20 of the full 50
182 ECMWF member forecasts were used in the generation of the multi-model ensemble,
183 though the 50-member ECMWF forecasts were evaluated for skill and reliability.
184 More detailed descriptions of the configuration of these four ensemble systems are
185 described in the Appendix 1.

186 When calibrating ECMWF data with reforecasts, the 5-member weekly
187 reforecasts were extracted from ECMWF's weekly reforecast archive (Hagedorn

188 2008) and similarly interpolated to the 1-degree grid. Since precipitation analysis
189 data was only available for the period from 2002 forward, the forecast training data
190 was the reforecasts for period of June to November, 2002-2009.

191
192 *c. Verification methods.*

193 The primary verification methods used here were Brier Skill Scores (*BSS*),
194 continuous ranked probability skill scores (*CRPSS*), and reliability diagrams (Wilks
195 2006). The *BSS* and *CRPSS* as conventionally calculated (see section 7.4.2 of Wilks
196 (2006)) can exaggerate forecast skill, attributing skill to variations in climatological
197 event probabilities. Thus, the procedures suggested in Hamill and Juras (2006) were
198 used here to avoid this.

199 To calculate the *BSS*, the score was calculated separately for subsets of points
200 that had more uniform climatological probabilities. The overall *BSS* was the average
201 of the skill scores over these subsets. The specific procedure was as follows. Using
202 the 1-degree precipitation analysis data from 2002-2009, for each month the
203 climatological probability of a given precipitation event was estimated from the
204 observed frequency. For a given event such as $> 1 \text{ mm } 24 \text{ h}^{-1}$, the n_s grid points
205 within the CONUS were sorted from lowest to highest event probability. The sorted
206 points were then divided into $k=6$ classes, with the lowest bin containing the $\sim n_s/6$
207 grid points with the lowest event probabilities, the highest bin containing the $n_s/6$
208 points with the highest probabilities, and so on (Fig. 1). Let $\mathbf{BS}^{f1} = [\mathbf{bs}_1^{f1}, \dots, \mathbf{bs}_6^{f1}]$
209 denote a matrix of Brier scores for forecast model $f1$, where \mathbf{bs}_i^{f1} was a $n_d -$
210 dimensional (= 123, the number of case days here) column vector of average Brier

211 scores for the points in the i^{th} class and for forecast model $f1$. An element of this
 212 vector thus provided the average Brier Score for all of the grid points in the i^{th} class
 213 on a particular day; the samples were weighted by the cosine of their latitude to
 214 account for differences in grid box size. The average over the 123 case days
 215 produced a vector $\overline{\mathbf{bs}}^{f1} = [\overline{bs_1^{f1}}, \dots, \overline{bs_6^{f1}}]$. Similarly, for climatology there was an
 216 array of Brier scores, $\mathbf{BS}^c = [\mathbf{bs}_1^c, \dots, \mathbf{bs}_6^c]$ and a vector of their averages over the
 217 123 days, $\overline{\mathbf{bs}}^c = [\overline{bs_1^c}, \dots, \overline{bs_6^c}]$. Following Hamill and Juras (2006) eq. (9), the
 218 overall BSS for model $f1$ was then calculated as

$$219 \quad BSS = \sum_{k=1}^6 \frac{1}{6} \left(1 - \frac{\overline{bs_k^{f1}}}{\overline{bs_k^c}} \right). \quad (1)$$

220 The boundaries between the classes were calculated independently for each event,
 221 so it was possible that a given grid point may have be assigned to different classes
 222 when evaluating, say, the 1- and 10-mm BSS es.

223 BSS confidence intervals were estimated using the paired block bootstrap
 224 approach of Hamill (1999). The input data to the bootstrap approach consisted of
 225 arrays of \mathbf{BS}^{f1} and \mathbf{BS}^{f2} for two competing models, $f1$ and $f2$, as well as \mathbf{BS}^c . Let
 226 $\mathbf{bs}^{f1}(d) = [bs_1^{f1}(d), \dots, bs_6^{f1}(d)]$ be the vector of forecast scores on the d^{th} case day,
 227 and similarly $\mathbf{bs}^{f2}(d)$ the vector for forecast model $f2$. The Brier scores, $\mathbf{bs}^{f1}(d)$
 228 were determined to be approximately statistically independent of $\mathbf{bs}^{f1}(d+1)$ and
 229 thus amenable to a paired resampling strategy, with these distinct vector blocks of
 230 data for each day. The following process was then repeated 10,000 times. For each

231 of the 123 days, a random uniform number between 0 and 1 was generated. If the
232 number was greater than 0.5, $\mathbf{bs}^{f1}(d)$ was randomly selected for inclusion in sample
233 1, $\mathbf{bs}^{f2}(d)$ was selected for inclusion in sample 2, and vice versa if the number was
234 less than or equal to 0.5. The vector of average Brier scores for samples s_1 and s_2
235 were then calculated, $\overline{\mathbf{bs}}^{s_1}$ and $\overline{\mathbf{bs}}^{s_2}$. The *BSS* for samples 1 and 2 were generated
236 via eq. (1), and the difference between the *BSS*s for the two samples was noted.
237 The confidence intervals are the 5th and 95th percentiles of the difference between
238 the *BSS*s of the two samples.

239 These block bootstrap confidence intervals should be regarded as
240 approximations. An assumption underlying this process is that there were 123
241 independent data samples. However, $\mathbf{bs}^{f1}(d)$ and $\mathbf{bs}^{f1}(d+1)$ may have been slightly
242 correlated, especially for the longer-lead forecasts, which will contribute to
243 overestimating the effective sample size and thus underestimating the confidence
244 interval. On the other hand, data from grid points across the CONUS were
245 aggregated in this procedure and thereafter considered as a single block. In reality
246 there may be far more than one independent sample spanning the CONUS, thus
247 leading to an under-estimate of sample size and consequent overestimate of the
248 confidence interval in this approach. Note also that for simplicity of presentation,
249 the skill diagrams will show only one set of confidence intervals, e.g., between NCEP
250 and ECMWF forecasts. Slightly smaller confidence intervals could be expected were
251 they computed using ECMWF and CMC forecasts, given their more similar skills.

252 The calculation of the *CRPSS* followed a slightly different procedure, similar
 253 to how forecast skill would be computed when the forecasts were of probabilities of
 254 exceedance of various quantiles. An example of this are the terciles commonly used
 255 in CPC's 6-10 day and 8-14 day forecasts. Thus, the *CRPS* at a given grid point was
 256 *not* computed by integrating differences between observed and forecast cumulative
 257 distribution functions (CDFs) over a range of precipitation values. Instead, the
 258 differences between observed and forecast CDFs were integrated over the
 259 percentiles of the CDF, which were determined separately for each model grid point
 260 and each month. Specifically, given n_d case days, for the $s = 1, \dots, n_d \times n_s$ samples,
 261 let $\mathbf{q}^s = [q_1^s, \dots, q_{20}^s]$ be the 20-dimensional vector of the precipitation quantiles
 262 associated with the 2.5th, 7.5th, ..., 97.5th percentiles of the climatological *CDF* for that
 263 point and that month for the s^{th} sample. The average forecast *CRPS_f* was
 264 determined by integrating over every 5th percentile, as

$$265 \quad CRPS_f = \frac{\sum_{s=1}^{n_d \times n_s} \cos(\phi_s) \sum_{iq=1}^{20} 0.05 \times [F^s(q_{iq}^s) - O^s(q_{iq}^s)]^2}{\sum_{s=1}^{n_d \times n_s} \cos(\phi_s)} \quad (2)$$

266 where $F^s(q_{iq}^s)$ represents the forecast's *CDF* for the s^{th} sample evaluated at the q_{iq}^s
 267 quantile, and $O^s(q_{iq}^s)$ represents the same, but for the observed (analyzed). ϕ_s is the
 268 latitude of the grid box, the cosine factor accounting for variations in grid box size.
 269 The analyzed state was assumed perfect, i.e., no analysis errors were incorporated,
 270 so the analyzed *CDF* was a Heaviside function, 0 at the quantiles less than the
 271 analyzed value, 1 at quantiles greater than or equal to the analyzed value. The *CRPS*
 272 of the climatological forecast, *CRPS_c*, was calculated as in eq. (2), but substituting

273 the climatological CDF for the forecast CDF. Finally, the overall skill score was
274 calculated as $CRPSS = 1 - CRPS_f / CRPS_c$. As with the *BSS*, a paired block bootstrap
275 approach was used to estimate the confidence intervals.

276 Two other common verification statistics were also used, root-mean-square
277 (RMS) errors, and bias, the average forecast divided by the average analyzed
278 amount.

279
280 *d. Statistical post-processing methodology.*

281 The extended logistic regression (ELR) approach of Wilks (2009) was used
282 here, a procedure that permitted the development of a single regression equation
283 that was suitable for predicting probabilities of exceeding any precipitation amount.
284 The probability was estimated with a function of the form

$$285 \quad p = \frac{\exp[f(\mathbf{x})]}{1 + \exp[f(\mathbf{x})]}, \quad (3)$$

286 where $f(\mathbf{x})$ was a linear function of the predictor variables. In this case, the
287 predictors were (a) the ensemble-mean forecast \bar{x} raised to the 0.4 power, (b) the
288 product of (a) and the variance σ^2 to the 0.4 power, and (c) the precipitation event
289 threshold T raised to the 0.4 power. The linear function was thus

$$290 \quad f(\mathbf{x}) = b_0 + b_1 \bar{x}^{0.4} + b_2 \bar{x}^{0.4} \sigma^{2 \times 0.4} + b_3 T^{0.4}. \quad (4)$$

291 The choice of these predictors was arrived at through some trial and error. The
292 power transformation of the predictors helped make the input data somewhat more
293 normally distributed. The probabilistic forecast skill was also only mildly
294 dependent on the inclusion/exclusion of the predictor with the product of the

295 transformed mean and variance. Skill was also only slightly dependent on the
296 power of the transform, with 0.4 providing an approximate minimum. Previous
297 values of power transformations in the literature have ranged from $\frac{1}{2}$ in Hamill and
298 Whitaker (2006) and Schmeits and Kok (2010), $\frac{1}{3}$ in Sloughter et al. (2007), and
299 $\frac{1}{4}$ in Hamill et al. (2008) and Roulin and Vannitsem (2011). The use of the product
300 of the ensemble mean and variance follows Wilks and Hamill (2007). The additional
301 predictor incorporating T permitted the single regression equation to be used to
302 predict probabilities across the range of possible amounts. A disadvantage of this
303 ELR approach (as opposed to approached such as the analog approach discussed in
304 Hamill and Whitaker (2006)) was that this algorithm was not able to correct for
305 possible position biases in forecast features.

306 ELR was applied both to calibrating real-time multi-model forecasts and to
307 calibrating ECMWF forecasts alone using the weekly reforecasts. It was found that
308 forecast skill increased if some method is applied to increase the modest training
309 sample sizes. A discussion of how sample sizes were augmented using data from
310 other forecast grid points is discussed in Appendix 2. .

311 Roulin and Vannitsem (2011) noted that since the ECMWF reforecast size (5
312 members) was smaller than the operational ensemble size (50 members; or in the
313 case here, 20 members selected from the 50), the regression coefficients may be
314 somewhat biased when trained with a smaller ensemble compared to what they
315 would be were they trained with a larger ensemble. Hence, when the coefficients
316 are used to correct the larger real-time ensemble, they may produce somewhat
317 biased probabilistic forecasts. They adjusted the values of the 5-member ensemble

318 training data to better estimate the values that would be obtained with the larger
319 real-time ensemble. An analogous approach was tried here but did not improve the
320 forecast skill. The results discussed below will omit this adjustment.

321 3. Results.

322 a. Properties of forecasts from the individual centers.

323 Before considering the multi-model and ECMWF reforecast-calibrated
324 forecast properties, let us consider the ensemble forecasts from the individual
325 centers. Figure 2 shows 1- and 10-mm *BSS* and *CRPSS*. ECMWF generally produced
326 the most skillful precipitation forecasts. Depending on the metric, either NCEP or
327 UKMO produced the least skillful forecasts. Interestingly, though UKMO forecasts
328 appeared to be generally more skillful than NCEP forecasts in *BSS*, they appeared to
329 be consistently worse in *CRPSS*. This was a consequence of the *CRPSS* verification
330 algorithm as implemented here, which attempted to equally weight the *CRPSS* at all
331 grid points, irrespective of whether the climatological probability was extremely
332 high or extremely low. The conventionally calculated *CRPSS* weighted the
333 climatologically wet areas more than the dry. Figure 3 shows maps of the day +3
334 *CRPSS* scores (see the online appendix for *CRPSS* maps for the other lead times).
335 The UKMO probabilistic forecasts had negative skill in the extremely dry regions of
336 the western US. The RMS errors of the ensemble-mean forecasts in the dry regions
337 of all the models were very small and relatively similar (Fig. 4a; for other lead times,
338 see the online appendix). However, the UKMO forecasts exhibited a large moist bias
339 in the climatologically dry regions (Fig. 4b), which resulted in a very large over-
340 forecast of probabilities and poor skill for those points. This was apparently due to

341 a drizzle over-forecast bias in that version of the UKMO's forecast model (D. Barker,
342 personal communication, 2011). Figure 4b also illustrates some other interesting
343 characteristics of the ensemble systems. NCEP over-forecasted rainfall for the grid
344 points and dates where the climatological probability was already quite high. CMC
345 forecasts were also biased, exhibiting a moist bias at the lowest climatological
346 probabilities but dry biases for most of the rest of the larger climatological
347 probabilities. ECMWF forecasts were the least biased, with a moderate over-
348 forecast bias at the low climatological probabilities.

349 Figure 5 provides reliability diagrams of day +3 forecasts for the > 10-mm 24
350 h⁻¹ event. Other reliability diagrams for other lead times and for the > 1-mm 24 h⁻¹
351 event are available in the online appendix. CMC forecasts were generally the most
352 reliable, though they were not as sharp as the ECMWF forecasts and hence had a
353 lower *BSS*. UKMO and NCEP forecasts were much less reliable, though NCEP
354 forecasts were slightly sharper than the others. ECMWF 50-member forecasts were
355 slightly more reliable and skillful than their 20-member subset.

356 Did the individual EPSes produce forecasts with any systematic biases in the
357 position of precipitation features? In subjective analyses of individual forecasts, it
358 appeared that several of the forecast models had subtle systematic northward
359 biases in the northern central US. Figure 6 shows the 10-mm observed contour and
360 the 0.5 probability contour for the > 10-mm 24 h⁻¹ event from the day +3 ECMWF
361 forecasts. Here, the 25 cases with the largest areal coverage of observed
362 precipitation between 105° and 80° west longitude and 35° and 50° north latitude

363 were chosen. Similar plots for the other forecast models are included in the online
364 appendix.

365

366 *b. Properties of multi-model and statistically post-processed forecasts.*

367 Consider first two actual forecast cases, presented in Figs. 6 and 7, showing
368 probabilities from the 20-member ensembles and from the 80-member multi-model
369 ensemble. The first case, covering the 24-h period ending 00 UTC 21 July 2010,
370 illustrates that sometimes the forecast models could be overly similar to each other.
371 Here all the forecast precipitation shields were significantly north of the observed
372 shield. A multi-model forecast would not be expected to provide much benefit in
373 such a situation. Figure 7 shows the same, but for 24-h period ending 00 UTC 7
374 August 2010. Here the multi-model forecast provided some improvement. On this
375 day the CMC and UKMO areas of high probabilities were too far north, the NCEP area
376 too far south, but the higher probabilities in the multi-model forecasts were more
377 coincident with the analyzed regions exceeding 10 mm. Most of the analyzed areas
378 with greater than 10 mm were covered by nonzero multi-model probabilities.

379 Figure 9 provides *BSSes* and *CRPSS* for the multi-model and the post-
380 processed forecasts. For the light precipitation forecasts ($> 1.0 \text{ mm } 24 \text{ h}^{-1}$), the
381 multi-model forecasts improved the skill by approximately +1 day relative to
382 ECMWF at the earliest lead times; a 2-day multi-model forecast could now be made
383 as skillfully as a +1 day ECMWF forecast. The improvement in skill was a more
384 modest $\sim +0.3$ days at the longer forecast lead times. The improvement from
385 reforecast-based post-processing over the raw ECMWF system was much smaller
386 and was even slightly negative at the day +5 lead for the 1-mm event. The calibrated

387 multi-model forecast product improved skill over the basic multi-model forecast by
388 a tiny amount at day +1 but degraded the skill after day +3. This is consistent with
389 previous results; at the longer lead times, the growth of errors makes it more
390 difficult to differentiate the model bias from the chaotically induced errors with
391 short training data sets (Hamill et al. 2004).

392 Even more impressive increases in skill were evident for the $> 10\text{-mm } 24 \text{ h}^{-1}$
393 event. Both the reforecast-based calibration and the multi-model approach
394 increased forecast skill by an equivalent of up to +2 days of additional lead time.
395 Again, the calibration of the multi-model forecasts provided modest improvement at
396 the early leads and degradation at the longer leads relative to the unprocessed
397 multi-model.

398 Measured in *CRPSS*, the multi-model forecasts produced the most skillful
399 forecasts, exceeding the skill of reforecast-calibrated ECMWF forecasts by a small
400 amount. Consider now *where* the forecasts were improved or degraded by the
401 various approaches. Figure 10 provides maps of the day +3 *CRPSS*; maps for other
402 lead times are in the online appendix. The patterns of multi-model skill are rather
403 similar to those of the most skillful ensemble system, ECMWF (Fig. 3a). The
404 reforecast-calibrated ECMWF forecasts appear to have increased the skill most
405 notably in the driest regions of the western US.

406 Figure 11 shows day +3 $> 10\text{-mm } 24 \text{ h}^{-1}$ event reliability diagrams for the
407 multi-model, the calibrated multi-model, and reforecast-calibrated ECMWF
408 forecasts. The raw multi-model forecasts were slightly more reliable than any of the
409 forecasts from the individual centers (Fig. 5) and retained a slight over-forecast bias

410 at the higher probabilities. The improvement in reliability was more substantial for
411 the $> 1\text{-mm } 24 \text{ h}^{-1}$ event; see diagrams in the online appendix. The reforecast-
412 calibrated probabilities exhibited a slight under-forecast bias and were not as sharp
413 as those from the multi-model forecasts. Was this due to some inhomogeneity
414 between the 2002-2009 training data and the 2010 real-time forecasts? Figure 12
415 shows that the absolute errors of the precipitation forecasts in 2010 had a shorter
416 tail than those in either 2002 or 2006; fewer large forecast busts occurred in 2010.
417 When the regression analysis from 2002-2009 data was applied to correct the 2010
418 forecasts, the assumption was that the 2010 forecasts would be equally unskillful.
419 In fact they were better, and as a consequence the post-processed forecasts were
420 less sharp than they could have been. Though it was not attempted here, it might be
421 possible to apply ad-hoc corrections to the training data to improve the regression
422 analysis. Perhaps a slight adjustment of the training data ensemble mean toward the
423 analyzed data would make its accuracy more closely resembles that of the 2010 data,
424 sharpening and making the ELR forecasts more reliable and skillful.

425 Figure 13 shows the areal coverage of the 0.5 probability contours for
426 selected cases, but here for the multi-model forecasts; these should be compared
427 with Fig. 8 for ECMWF-only forecasts. Figure 14 also shows the areal coverage, but
428 for reforecast-calibrated ECMWF forecasts. The areal coverage was only slightly
429 smaller for the multi-model forecasts than it was for the ECMWF forecasts,
430 illustrating that the multi-model forecasts did not lose a tremendous amount of
431 sharpness. In comparison, the reforecast-calibrated product in Fig. 14 showed a
432 marked decrease in the areal coverage; many grid points with probability $p > 0.5$ in

433 the ECMWF ensemble had $p < 0.5$ after calibration. Figures 15 and 16 show for the
434 cases plotted in Figs. 7 and 8 a bit more detail on what happened with typical multi-
435 model and reforecast-calibrated probability forecasts. The multi-model forecasts
436 retained their sharpness, but not always desirably so. For example, in Fig. 15, the
437 multi-model forecasts retain relatively high probabilities in eastern Iowa and
438 northern Illinois, whereas the analyzed area was displaced further south. The
439 reforecast-calibrated product decreased the areal coverage of high probabilities,
440 appropriately so in this case, reducing the false alarms. However, as seen in
441 inspection of Figs. 13-14, there were many cases when the sharpness retained in the
442 multi-model forecasts was desirable.

443 Overall, the impressive skill improvements provide evidence for the merit of
444 both multi-model and reforecast approaches. Should other forecast centers share
445 precipitation ensemble data, large gains in probabilistic precipitation forecast skill
446 are possible for little more than the cost of data transmission and storage.
447 Alternatively, should any one center produce and utilize reforecasts, they can
448 improve their own forecasts significantly, assuming a comparably long time series
449 of observations or analyses are available. The improvement here noted with
450 reforecasts may have also been modest because the training data was limited on
451 account of a short time series of analyses, dating back to only 2002; only around
452 40% of the available reforecast data was used.

453

454 **4. Conclusions.**

455 This article examined probabilistic multi-model weather forecasts of
456 precipitation over the CONUS and the relative advantages and disadvantages of

457 these forecasts when compared to statistically post-processed ECMWF forecasts.
458 20-member forecasts were extracted from the ECMWF, NCEP, UKMO, and CMC
459 global ensemble systems at 1-degree resolution between June and October 2010.
460 Daily 24-h accumulated probabilistic precipitation forecasts were generated from
461 the subsequent 80-member ensemble for lead times of +1 to +5 days and compared
462 to gridded precipitation analyses. Two statistically post-processed products were
463 also evaluated, the first being multi-model forecasts that were adjusted using
464 extended logistic regression and that were trained on the previous 30 days of
465 forecasts and analyses. The second was ECMWF forecasts, which were statistically
466 adjusted using forecast/analysis data for the period 2002-2009, the time period
467 when both reforecasts and analyses were available.

468 Considering first the skill of forecasts from the individual EPSes, ECMWF
469 forecasts generally were the most skillful in terms of Brier skill scores and the
470 continuous ranked probability skill score. CMC forecasts were the most reliable but
471 the least sharp, while NCEP and UKMO forecasts were more sharp but less reliable.

472 Multi-model probabilistic forecast products were substantially more skillful
473 than the best of the individual centers' probabilistic forecasts. The improvement
474 was on the order of an extra 0.5 to 1 day of forecast lead time for light precipitation
475 events and as much as 2 days for heavier precipitation events. The reforecast-
476 calibrated ECMWF forecasts did not exhibit nearly as much skill improvement at the
477 1-mm event as they did at the 10-mm event. Relative to the multi-model forecasts,
478 the skills were similar at 10 mm, but the reforecast-calibrated was more reliable
479 while the multi-model was sharper.

480 The results exhibited here with reforecast calibration were not as impressive
481 as they have been in previous studies, e.g., Hamill and Whitaker (2006) and Hamill
482 et al. (2008). There are at least four reasons for this. First, the training data was not
483 as accurate as the real-time data in this application (Fig. 12), and this inhomogeneity
484 degraded the regression analysis. The second is that gratifying improvements have
485 been made to models and EPSes so that they produce more skillful and reliable
486 forecasts than they did in the past; it's tougher to improve upon ECMWF's 2010's
487 model output than its 2005 model output. The third reason is that even with
488 reforecasts, there really was a limited training data set in this study, here due to the
489 unavailability of precipitation analyses prior to 2002. The fourth reason is that in
490 prior studies, the ensemble forecasts (at coarse resolution) were evaluated against
491 analysis data at finer resolution, so that the reforecast calibration process was also
492 producing a statistical downscaling. This point is worth keeping in mind when
493 considering the relative merits of reforecast calibration vs. multi-model approaches.
494 If the desired output is forecast data at the grid scale, multi-models may have
495 substantial appeal. If the desired output is point data or high-resolution gridded
496 data, the statistical downscaling is more straightforward when reforecasts are used.

497 I was pleasantly surprised by the magnitude of skill improvements
498 demonstrated here from multi-model ensembles, improvements which were larger
499 than those seen with 2-meter temperatures (Hagedorn et al. 2011). From our own
500 experience, however, I recommend some caution against over-interpreting these
501 results. This study examined a combination of data from four mature EPSes based
502 on mature models and assimilation systems. Each center's system has been refined

503 through the collective efforts of hundreds if not thousands of person-years of
504 research and development. A combination of less developed EPSes may not
505 provide nearly the same gratifying result.

506 Nonetheless, these results demonstrate the potential value of multi-model
507 ensembles. The THORPEX program, organized by the World Meteorological
508 Organization, has promoted the concept of a multi-model based “Global Interactive
509 Forecast System” (Bougeault et al. 2010), whereby the operational centers share
510 data that will facilitate the production of multi-model products for high-impact
511 weather events. This study provides additional evidence for the validity and the
512 potential benefits of such a system. Currently several centers have restrictive data
513 policies; full access to their data is reserved for paying customers, and those
514 customers cannot thereafter share the data they purchased. I hope that the
515 approach embraced in the US and Canada will be embraced by other centers
516 worldwide, for the mutual benefit of all. In the US and Canada, the data is effectively
517 free since the research, development, and production were funded by public
518 taxpayer funds.

519 Finally, can we all have “the best of both worlds?” That is, will NWP centers
520 both agree to share their data freely and internationally, and will they produce
521 reforecast data sets so that each model can be calibrated to remove systematic
522 errors prior to their combination? There is evidence that such approaches will
523 provide substantial benefit. The climate community is working on sharing multi-
524 model information and hindcasts to facilitate the error correction for intra-seasonal
525 and seasonal forecasts, and for weather and weather-to-climate applications, there

526 have also been successful demonstrations of multi-model calibrated forecasts
527 (Vislocky and Fritsch 1995, Whitaker et al. 2006). NOAA is currently developing a
528 new reforecast data set for its global ensemble prediction system, and I hope that
529 other centers will be inspired to do so as well.

530
531
532

Acknowledgments

533 TIGGE data was supplied from ECMWF's TIGGE data portal. I thank ECMWF for the
534 development of this portal software and for the archival of this (immense) data set.
535 Florian Pappenberger of ECMWF was very helpful in extracting and pre-processing
536 the reforecasts. Yan Luo of NCEP/EMC was helpful in obtaining the CCPA analysis
537 data. Tom Galarneau of NCAR/MMM is thanked for providing an informal review.
538 This study was funded in part by the NOAA THORPEX program.

539

Appendix 1.

541 Here are additional details on the forecast models and ensemble systems used in
542 this experiment.

543
544

a. NCEP

545 NCEP used the GFS model in their ensemble system at T190L28 resolution.
546 The GFS treatment of vertical mixing, including the planetary boundary layer, was
547 based on Hong and Pan (1996) and Troen and Mahrt (1986). Short-wave radiation
548 followed Chou (1992), Chou and Lee (1996), Chou et al. (1998), while long-wave
549 radiation implements the Rapid Radiative Transfer Model (RRTM) of Mlawer et al.
550 (1997). Penetrative convection was developed by Pan and Wu (1995), based on a

551 Grell (1993) implementation of Arakawa and Schubert (1974). The effect of non-
552 precipitating shallow clouds is incorporated following Tiedtke (1983) . Cloud
553 formation on resolved scales was treated according to a Zhao and Carr (1997)
554 modification of Sundqvist et al. (1989). Attenuation of gravity waves propagating
555 into the mesosphere was accomplished with Rayleigh drag. A description of the GFS
556 model is available from the NCEP Environmental Modeling Center (EMC), with
557 changes as of 2003 described at www.emc.ncep.noaa.gov/gmb/moorthi/gam.html.

558 The control initial condition around which the perturbed initial conditions
559 were centered was produced by the T382 Global Statistical Interpolation (GSI)
560 analysis (Kleist et al. 2009) at T384L64 resolution. Perturbed initial conditions
561 were generated with the ensemble transform with rescaling technique of Wei et al.
562 (2008). Stochastic perturbations were included, following Hou et al. (2008). More
563 details on changes to the NCEP ensemble system can be found at
564 http://www.emc.ncep.noaa.gov/gmb/yzhu/html/ENS_IMP.html.

565
566 *b. Canadian Meteorological Centre*

567 The CMC EPS used the Global Environmental Multiscale Model (GEM), a
568 hydrostatic primitive equation model with a terrain-following pressure vertical
569 coordinate. Further documentation on the GEM model can be found at
570 [http://collaboration.cmc.ec.gc.ca/science/rpn/gef_html_public/DOCUMENTATION/](http://collaboration.cmc.ec.gc.ca/science/rpn/gef_html_public/DOCUMENTATION/GENERAL/general.html)
571 [GENERAL/general.html](http://collaboration.cmc.ec.gc.ca/science/rpn/gef_html_public/DOCUMENTATION/GENERAL/general.html) and in Charron et al. (2010). The CMC ensemble system
572 used a horizontal computational grid of 400x200 grid points, or approximately 0.9
573 degrees, and 28 vertical levels. The EnKF initial conditions were used, following
574 Charron et al. (2010) and Houtekamer et al. (2009) and references therein. The 20

575 forecast ensemble members used a variety of perturbed physics; changing gravity
576 wave drag parameters, land-surface process type, condensation scheme type,
577 convection scheme type, shallow convection scheme type, mixing-length
578 formulation, and turbulent vertical diffusion parameter. More details on these are
579 provided at http://www.weatheroffice.gc.ca/ensemble/verifs/model_e.html.

580
581 *c. European Centre for Medium-Range Weather Forecasts.*

582
583 The ECMWF EPS used the ECMWF Integrated Forecast System (IFS) model,
584 versions 36r2. Model resolution was T639L62 for both versions; details on the IFS
585 are provided at www.ecmwf.int/research/ifsdocs/. The changes to the ensemble
586 stochastic treatments in the 8 Sep 2009 implementation are described in Palmer et
587 al. (2009). The ensemble was initialized with a combination of initial-time and
588 evolved total-energy singular vectors (Buizza and Palmer 1995, Molteni et al. 1996,
589 Barkmeijer et al. 1998, Barkmeijer et al. 1999, Leutbecher 2005) and utilized
590 stochastic perturbations to physical tendencies. An overview of the ensemble
591 system was provided in Buizza et al. (2007) and references therein. For
592 consistency with the analysis of other EPSs, only the first 20 perturbed members
593 were used here.

594
595 *e. United Kingdom Met Office.*

596
597 The UK Met Office (UKMO) ensemble system was “MOGREPS,” the Met Office
598 Global and Regional Ensemble Prediction System. TC track forecasts from this
599 system came from its global component, which was described in Bowler et al. (2008,

600 2009). The global system was run at a resolution of 1.25° longitude and 0.83°
601 latitude on a regular latitude-longitude grid. 38 vertical levels were employed.
602 Initial condition perturbations were generated from an implementation of the
603 ensemble transform Kalman filter (Hunt et al. 2006, Bowler et al. 2009). The mean
604 initial state was generated from the UKMO 4D-Var system (Rawlins et al. 2007). The
605 model included a parameterization of one type of model uncertainty via its
606 stochastic kinetic-energy backscatter scheme, following Shutts (2005).

607

608 **Appendix 2:**

609 This appendix discusses the method used to augment the training sample
610 size used in the regression analyses. Were only the data at the grid point of interest
611 used for training, when calibrating using the multi-model ensemble using the past
612 30 days of forecasts and analyses, this would provide, of course, only 30 training
613 samples. Older forecasts could be used, but precipitation biases are often seasonally
614 dependent, so the older data may degrade the results despite augmenting the
615 sample size. Also, with such a multi-model ensemble, the farther back into the past
616 one seeks training data, the more likely it is that at least one of the models will have
617 had a major upgrade and concomitant change in systematic error characteristics.
618 Despite ECMWF providing a multi-decadal reforecast, in practice the sample sizes
619 were too small here, too. When using the 2002-2009 weekly, 5-member ECMWF
620 reforecasts (including reforecast dates +/- 6 weeks around the week of interest),
621 this provided a total of 13 weeks \times 8 years = 104 samples. In both cases, these were
622 relatively small samples to estimate four regression parameters, and especially for
623 rare events such as heavy precipitation, experience has shown that larger training
624 samples improved the regression analysis.

625 Hence, following the general philosophy demonstrated and discussed in
626 Hamill et al. (2008) and inspired by the regionalization used in some Model Output
627 Statistics algorithms (Lowry and Glahn 1976), the training data set for a particular
628 grid point was augmented by finding 25 other grid points that had relatively similar
629 climatological analyzed CDFs. Consider a particular location (λ, ϕ) at which we seek

630 to augment the sample size, and another location (λ^s, ϕ^s) we are considering as a
631 location with suitable supplemental training data. Differences between the analyzed
632 cumulative probabilities at (λ, ϕ) and (λ^s, ϕ^s) were measured at the 1, 2.5, 5, 10, 25,
633 and 50 mm 24 h^{-1} amounts and then weighted by similar respective factors of [1,
634 2.5, 5, 10, 25, 50]. That is, a cumulative probability difference of 0.1 at 1 mm and
635 0.1/50 at 50 mm were judged to have the same weighted difference. The maximum
636 weighted difference at any of the possible precipitation amounts was then noted for
637 this (λ^s, ϕ^s) . Having evaluated the maximum of the weighted differences all the grid
638 points less than 8 grid points distant from the grid point of interest (λ, ϕ) , the 25
639 grid points with the smallest weighted differences were identified, and the training
640 sample for (λ, ϕ) was augmented by the forecasts-analysis pairs at these locations.
641 This approach increased sample size, but it's possible that the forecast bias might
642 have been different at the supplemental locations, and hence not an unalloyed
643 benefit. For more discussion of this, see Hamill et al. (2008, section 3a).

644

645

646

647

648

649

650 **References**

651

652

653 Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of
654 Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple
655 Nearest-Neighbor Average Method on High-Resolution Verification Grids.
656 *Weather and Forecasting*, **18**, 918-932.

657 Arakawa, A., and W. H. Schubert, 1974: Interaction of a Cumulus Cloud Ensemble
658 with the Large-Scale Environment, Part I. *Journal of the Atmospheric Sciences*,
659 **31**, 674-701.

660 Bao, L., T. Gneiting, E. P. Grit, P. Guttorp, and A. E. Raftery, 2010: Bias Correction
661 and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind
662 Direction. *Monthly Weather Review*, **138**, 1811-1821.

663 Barkmeijer, J., F. Bouttier, and M. Van Gijzen, 1998: Singular vectors and estimates of
664 the analysis-error covariance metric. *Quarterly Journal of the Royal
665 Meteorological Society*, **124**, 1695-1713.

666 Barkmeijer, J., R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors
667 and their potential use in the ECMWF ensemble prediction system. *Quarterly
668 Journal of the Royal Meteorological Society*, **125**, 2333-2351.

669 Berner, J., S.-Y. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty
670 in a mesoscale ensemble prediction system: Stochastic versus multi-physics
671 representations. *Monthly Weather Review*, **in press.**, null.

672 Berner, J., G. J. Shutts, M. Leutbecher, and T. N. Palmer, 2009: A Spectral Stochastic
673 Kinetic Energy Backscatter Scheme and Its Impact on Flow-Dependent

674 Predictability in the ECMWF Ensemble Prediction System. *Journal of the*
675 *Atmospheric Sciences*, **66**, 603-626.

676 Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global
677 Ensemble. *Bulletin of the American Meteorological Society*, **91**, 1059-1072.

678 Bowler, N. E., A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local
679 ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction
680 system. *Quarterly Journal of the Royal Meteorological Society*, **135**, 767-776.

681 Bowler, N. E., A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare, 2008: The
682 MOGREPS short-range ensemble prediction system. *Quarterly Journal of the*
683 *Royal Meteorological Society*, **134**, 703-722.

684 Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, and F. Vitart, 2007:
685 The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System).
686 *Quarterly Journal of the Royal Meteorological Society*, **133**, 681-695.

687 Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model
688 uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of*
689 *the Royal Meteorological Society*, **125**, 2887-2908.

690 Buizza, R., and T. N. Palmer, 1995: The Singular-Vector Structure of the Atmospheric
691 Global Circulation. *Journal of the Atmospheric Sciences*, **52**, 1434-1456.

692 Charron, M., G. r. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and
693 L. Michelin, 2010: Toward Random Sampling of Model Error in the Canadian
694 Ensemble Prediction System. *Monthly Weather Review*, **138**, 1877-1901.

695 Chou, M.-D., 1992: A Solar Radiation Model for Use in Climate Studies. *Journal of the*
696 *Atmospheric Sciences*, **49**, 762-772.

697 Chou, M.-D., and K.-T. Lee, 1996: Parameterizations for the Absorption of Solar
698 Radiation by Water Vapor and Ozone. *Journal of the Atmospheric Sciences*, **53**,
699 1203-1208.

700 Chou, M.-D., M. J. Suarez, C.-H. Ho, M. M.-H. Yan, and K.-T. Lee, 1998:
701 Parameterizations for Cloud Overlapping and Shortwave Single-Scattering
702 Properties for Use in General Circulation and Cloud Ensemble Models.
703 *Journal of Climate*, **11**, 202-214.

704 Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the
705 success of multi-model ensembles in seasonal forecasting – II. Calibration
706 and combination. *Tellus A*, **57**, 234-252.

707 Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B.
708 Jackson, 2009: MOS Uncertainty Estimates in an Ensemble Framework.
709 *Monthly Weather Review*, **137**, 246-268.

710 Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated
711 Probabilistic Forecasting Using Ensemble Model Output Statistics and
712 Minimum CRPS Estimation. *Monthly Weather Review*, **133**, 1098-1118.

713 Grell, G. A., 1993: Prognostic Evaluation of Assumptions Used by Cumulus
714 Parameterizations. *Monthly Weather Review*, **121**, 764-787.

715 Hagedorn, R., 2008: Using the ECMWF reforecast data set to calibrate EPS
716 reforecasts. *ECMWF Newsletter*, **117**, 8-13.

717 Hagedorn, R., R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2011:
718 Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF

719 ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*,
720 **submitted; available from martin.leutbecher@ecmwf.int.**

721 Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the
722 success of multi-model ensembles in seasonal forecasting – I. Basic concept.
723 *Tellus A*, **57**, 219-233.

724 Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic Forecast
725 Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter
726 Temperatures. *Monthly Weather Review*, **136**, 2608-2619.

727 Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic Forecast
728 Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II:
729 Precipitation. *Monthly Weather Review*, **136**, 2620-2632.

730 Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the
731 varying climatology? *Quarterly Journal of the Royal Meteorological Society*,
732 **132**, 2905-2923.

733 Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic Quantitative Precipitation
734 Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly*
735 *Weather Review*, **134**, 3209-3229.

736 ———, 2007: Ensemble Calibration of 500-hPa Geopotential Height and 850-hPa and
737 2-m Temperatures Using Reforecasts. *Monthly Weather Review*, **135**, 3273-
738 3280.

739 Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An Important
740 Dataset for Improving Weather Predictions. *Bulletin of the American*
741 *Meteorological Society*, **87**, 33-46.

742 Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble Reforecasting: Improving
743 Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly*
744 *Weather Review*, **132**, 1434-1447.

745 Higgins, R. W., J. E. Janowiak, and Y.-P. Yao, 1996: A Gridded Hourly Precipitation
746 Data Base for the United States (1963-1993). *NCEP/Climate Prediction Center*
747 *ATLAS No. 1, U. S. DEPARTMENT OF COMMERCE, National Oceanic and*
748 *Atmospheric Administration, National Weather Service.*

749 Hong, S.-Y., and H.-L. Pan, 1996: Nonlocal Boundary Layer Vertical Diffusion in a
750 Medium-Range Forecast Model. *Monthly Weather Review*, **124**, 2322-2339.

751 Hou, D., Z. Toth, Y. Zhu, and Y. Yang, 2008: Impact of a Stochastic Perturbation
752 Scheme on NCEP Global Ensemble Forecast System. *Proceedings, 19th AMS*
753 *conference on Probability and Statistics. New Orleans, LA, 20-24 Jan. 2008.*

754 Houtekamer, P. L., and H. L. Mitchell, 2005: Ensemble Kalman filtering. *Quarterly*
755 *Journal of the Royal Meteorological Society*, **131**, 3269-3289.

756 Houtekamer, P. L., H. L. Mitchell, and X. Deng, 2009: Model Error Representation in
757 an Operational Ensemble Kalman Filter. *Monthly Weather Review*, **137**, 2126-
758 2143.

759 Hunt, B., E. Kostelich, and I. Szunyogh, 2006: Efficient Data Assimilation for
760 Spatiotemporal Chaos: a Local Ensemble Transform Kalman Filter.

761 Iversen, T., A. Deckmyn, C. Santos, K. A. I. Sattler, J. B. Bremnes, H. Feddersen, and I.-L.
762 Frogner, 2011: Evaluation of 'GLAMEPS'—a proposed multimodel EPS for
763 short range forecasting. *Tellus A*, **63**, 513-530.

764 Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble
765 combination and calibration. *Quarterly Journal of the Royal Meteorological*
766 *Society*, **135**, 777-794.

767 Kleist, D. T., D. F. Parrish, J. C. Derber, R. Treadon, W.-S. Wu, and S. Lord, 2009:
768 Introduction of the GSI into the NCEP Global Data Assimilation System.
769 *Weather and Forecasting*, **24**, 1691-1705.

770 Krishnamurti, T. N., and Coauthors, 2000: Multimodel Ensemble Forecasts for
771 Weather and Seasonal Climate. *Journal of Climate*, **13**, 4196-4216.

772 Leutbecher, M., 2005: On Ensemble Prediction Using Singular Vectors Started from
773 Forecasts. *Monthly Weather Review*, **133**, 3038-3046.

774 Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses:
775 development and applications. *Preprints, 19th Conf. on Hydrology, American*
776 *Meteorological Society, San Diego, CA 9-13 January 2005, Paper 1.2.* .

777 Lorenz, J., H. Rauhut, F. Schweitzer, and D. Helbing, 2011: How social influence can
778 undermine the wisdom of crowd effect. *Proceedings of the National Academy*
779 *of Sciences*.

780 Lowry, D. A., and H. R. Glahn, 1976: An Operational Model for Forecasting
781 Probability of Precipitation, ÆPEATMOS PoP. *Monthly Weather Review*, **104**,
782 221-232.

783 Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997:
784 Radiative transfer for inhomogeneous atmospheres: RRTM, a validated
785 correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16663-16682.

786 Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble
787 Prediction System: Methodology and validation. *Quarterly Journal of the*
788 *Royal Meteorological Society*, **122**, 73-119.

789 Mylne, K. R., R. E. Evans, and R. T. Clark, 2002: Multi-model multi-analysis ensembles
790 in quasi-operational medium-range forecasting. *Quarterly Journal of the*
791 *Royal Meteorological Society*, **128**, 361-384.

792 Palmer, T. N., and Coauthors, 2009: Stochastic parameterization and model
793 uncertainty. *ECMWF Tech Memo 589*.

794 Pan, H.-L., and W.-S. Wu, 1995: Implementing a mass flux convection
795 parameterization package for the NMC Medium-Range Forecast Model. *NMC*
796 *Office Note*, **409**, 40.

797 Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian
798 Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*,
799 **133**, 1155-1174.

800 Rawlins, F., and Coauthors, 2007: The Met Office global four-dimensional variational
801 data assimilation scheme. *Quarterly Journal of the Royal Meteorological*
802 *Society*, **133**, 347-362.

803 Roulin, E., and S. Vannitsem, 2011: Post-processing of ensemble precipitation
804 predictions with extended logistic regression based on hindcasts. *Monthly*
805 *Weather Review*, **139**, Available from Emmanuel.Roulin@meteo.be.

806 Schmeits, M. J., and K. J. Kok, 2010: A Comparison between Raw Ensemble Output,
807 (Modified) Bayesian Model Averaging, and Extended Logistic Regression

808 Using ECMWF Ensemble Precipitation Reforecasts. *Monthly Weather Review*,
809 **138**, 4199-4211.

810 Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble
811 prediction systems. *Quarterly Journal of the Royal Meteorological Society*, **131**,
812 3079-3102.

813 Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic
814 Quantitative Precipitation Forecasting Using Bayesian Model Averaging.
815 *Monthly Weather Review*, **135**, 3209-3220.

816 Stensrud, D. J., and N. Yussouf, 2003: Short-Range Ensemble Predictions of 2-m
817 Temperature and Dewpoint Temperature over New England. *Monthly*
818 *Weather Review*, **131**, 2510-2524.

819 Sundqvist, H., E. Berge, and J. E. Kristjánsson, 1989: Condensation and Cloud
820 Parameterization Studies with a Mesoscale Numerical Weather Prediction
821 Model. *Monthly Weather Review*, **117**, 1641-1657.

822 Tiedtke, M., 1983: The sensitivity of the time-mean large-scale flow to cumulus
823 convection in the ECMWF model. . *ECMWF Workshop on Convection in Large-*
824 *Scale Models*, 297-316.

825 Troen, I. B., and L. Mahrt, 1986: A simple model of the atmospheric boundary layer;
826 sensitivity to surface evaporation. *Boundary-Layer Meteorology*, **37**, 129-148.

827 Vannitsem, S., and C. Nicolis, 2008: Dynamical Properties of Model Output Statistics
828 Forecasts. *Monthly Weather Review*, **136**, 405-419.

829 Vislocky, R. L., and J. M. Fritsch, 1995: Improved Model Output Statistics Forecasts
830 through Model Consensus. *Bulletin of the American Meteorological Society*, **76**,
831 1157-1164.

832 Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a
833 Short-Range Multimodel Ensemble System. *Monthly Weather Review*, **129**,
834 729-747.

835 Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the
836 ensemble transform (ET) technique in the NCEP global operational forecast
837 system. *Tellus A*, **60**, 62-79.

838 Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination
839 really enhance the prediction skill of probabilistic ensemble forecasts?
840 *Quarterly Journal of the Royal Meteorological Society*, **134**, 241-260.

841 Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving Week-2 Forecasts with
842 Multimodel Reforecast Ensembles. *Monthly Weather Review*, **134**, 2279-2284.

843 Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences (2nd Ed.)*.
844 Academic Press, 627. pp.

845 —, 2009: Extending logistic regression to provide full-probability-distribution
846 MOS forecasts. *Meteorological Applications*, **16**, 361-368.

847 Wilks, D. S., and T. M. Hamill, 2007: Comparison of Ensemble-MOS Methods Using
848 GFS Reforecasts. *Monthly Weather Review*, **135**, 2379-2390.

849 Wilson, L. J., S. Beauguard, A. E. Raftery, and R. Verret, 2007: Calibrated Surface
850 Temperature Forecasts from the Canadian Ensemble Prediction System
851 Using Bayesian Model Averaging. *Monthly Weather Review*, **135**, 1364-1385.

852 Yussouf, N., and D. J. Stensrud, 2007: Bias-Corrected Short-Range Ensemble
853 Forecasts of Near-Surface Variables during the 2005/06 Cool Season.
854 *Weather and Forecasting*, **22**, 1274-1286.

855 Zhao, Q., and F. H. Carr, 1997: A Prognostic Cloud Scheme for Operational NWP
856 Models. *Monthly Weather Review*, **125**, 1931-1953.

857

858

859

860 **Figure captions**

861

862

863 **Figure 1:** Illustration of the process for determining precipitation classes used in

864 the calculation of *BSS*. (a) Climatological probability of $> 1\text{-mm } 24\text{h}^{-1}$

865 precipitation as determined from Stage-IV data for September 2002-2009.

866 (b) Climatological class assigned to each grid point for September, $1\text{-mm } 24$

867 h^{-1} event.

868 **Figure 2:** Brier skill scores of various forecasts for the (a) $> 1\text{-mm } 24 \text{ h}^{-1}$ event, (b)

869 $> 10\text{-mm } 24 \text{ h}^{-1}$ event, and (c) continuous ranked probability skill scores, all

870 as a function of forecast lead time. Error bars denote confidence intervals,

871 the 5th and 95th percentiles of a paired block bootstrap between ECMWF and

872 NCEP forecasts.

873 **Figure 3:** Maps of average *CRPSS* for day +3 forecasts for (a) ECMWF, (b) NCEP, (c)

874 UKMO, and (d) CMC.

875 **Figure 4:** (a) RMS errors, and (b) bias for day +3 forecasts, each as a function of the

876 climatological probability of greater than $1\text{-mm } 24 \text{ h}^{-1}$. Light grey bars in

877 panel (a) denote the relative frequency of each climatological probability.

878 **Figure 5:** Reliability diagrams for day +3 forecasts for the $> 10\text{-mm } 24 \text{ h}^{-1}$ event. (a)

879 ECMWF, (b) NCEP, (c) CMC, and (d) UKMO. The dark line on each is the 20-

880 member reliability curve. The lighter grey line on panel (a) is the reliability

881 for the full 50-member ensemble. The inset histogram bars show the relative

882 frequency of usage for each probability bin. The black lines on the inset are

883 the relative frequency of usage for the climatological distribution across all

884 the sample points. The grey dots on the inset histogram of panel (a) are the
885 relative frequency of usage for the ECMWF full 50-member ensemble.

886 **Figure 6:** Analyzed $> 10\text{-mm } 24 \text{ h}^{-1}$ precipitation boundary (black line) and area
887 exceeding 10 mm (grey shading) for 25 cases with the largest areal coverage
888 of greater than 10 mm in the upper Midwest US. Red lines indicate the 0.5
889 probability contour from the ECMWF ensemble for the day +3 forecasts of $>$
890 10 mm 24 h^{-1} .

891 **Figure 7:** (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010.
892 10-mm 24 h^{-1} contour is denoted by the thick black line. (b) Probability of
893 greater than 10 mm 24 h^{-1} for day +3 forecast from the ECMWF ensemble for
894 the same period. The analyzed 10-mm contour from panel (a) is repeated.
895 (c) as in (b), but for NCEP. (d) CMC, (e) UK Met Office, and (f) multi-model
896 combination.

897 **Figure 8:** As in Fig. 7, but for 24-h period ending 00 UTC 8 August 2010.

898 **Figure 9:** Brier skill scores of various forecasts for (a) $> 1\text{-mm } 24 \text{ h}^{-1}$ event, and (b)
899 $> 10\text{-mm } 24 \text{ h}^{-1}$ event, and (c) continuous ranked probability skill scores, all
900 as a function of forecast lead time. “Multi-model/cal” refers to forecasts from
901 the multi-model, calibrated using ELR. “ECMWF/reforecast” refers to
902 ECMWF forecasts calibrated using ELR and the reforecast data set. Error bars
903 denote confidence intervals, the 5th and 95th percentiles of a paired block
904 bootstrap between ECMWF and NCEP forecasts.

905 **Figure 10:** Maps of *CRPSS* for day +3 forecasts for (a) multi-model, (b) multi-model
906 with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.

907 **Figure 11:** As in Fig. 3, reliability diagrams for day +3 forecasts at the > 10-mm
908 24 h⁻¹ event. (a) Multi-model forecasts, (b) multi-model with ELR calibration,
909 and (c) ECMWF with ELR calibration using reforecasts.

910 **Figure 12:** A histogram of the absolute errors of day +3 ensemble-mean
911 precipitation forecasts for the 2002 and 2006 reforecasts and for the 2010,
912 20-member real-time ensemble.

913 **Figure 13:** As in Fig. 6, but for multi-model forecasts.

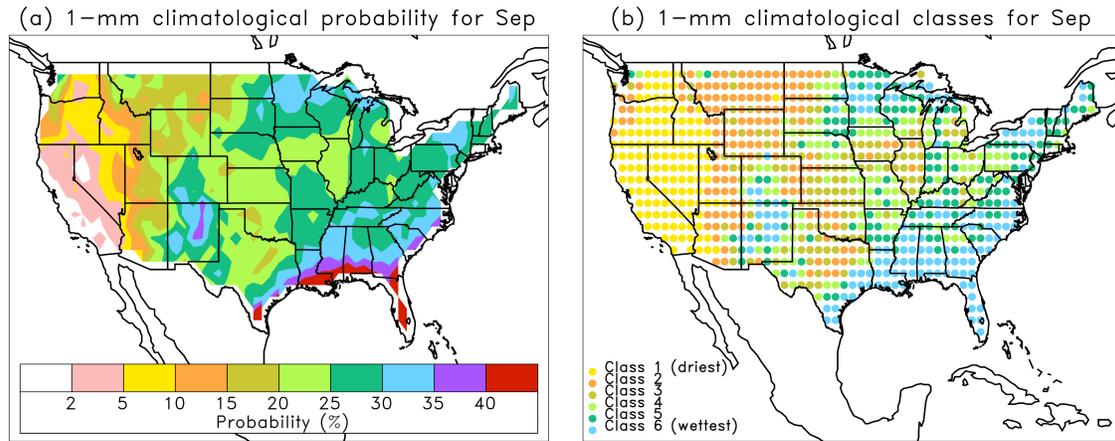
914 **Figure 14:** As in Fig. 6, but for reforecast-calibrated ECMWF forecasts.

915 **Figure 15:** (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July
916 2010. 10-mm contour is denoted by the thick black line. (b) Probability of
917 greater than 10 mm 24 h⁻¹ for day +3 forecast from the ECMWF ensemble for
918 the same period. (c) as in (b), but for multi-model ensemble, and (d) as in (b),
919 but for reforecast-calibrated ECMWF ensemble.

920 **Figure 16:** As in Fig. 15, but for the 24-h period ending 00 UTC 8 August 2010.

921
922
923
924
925
926
927

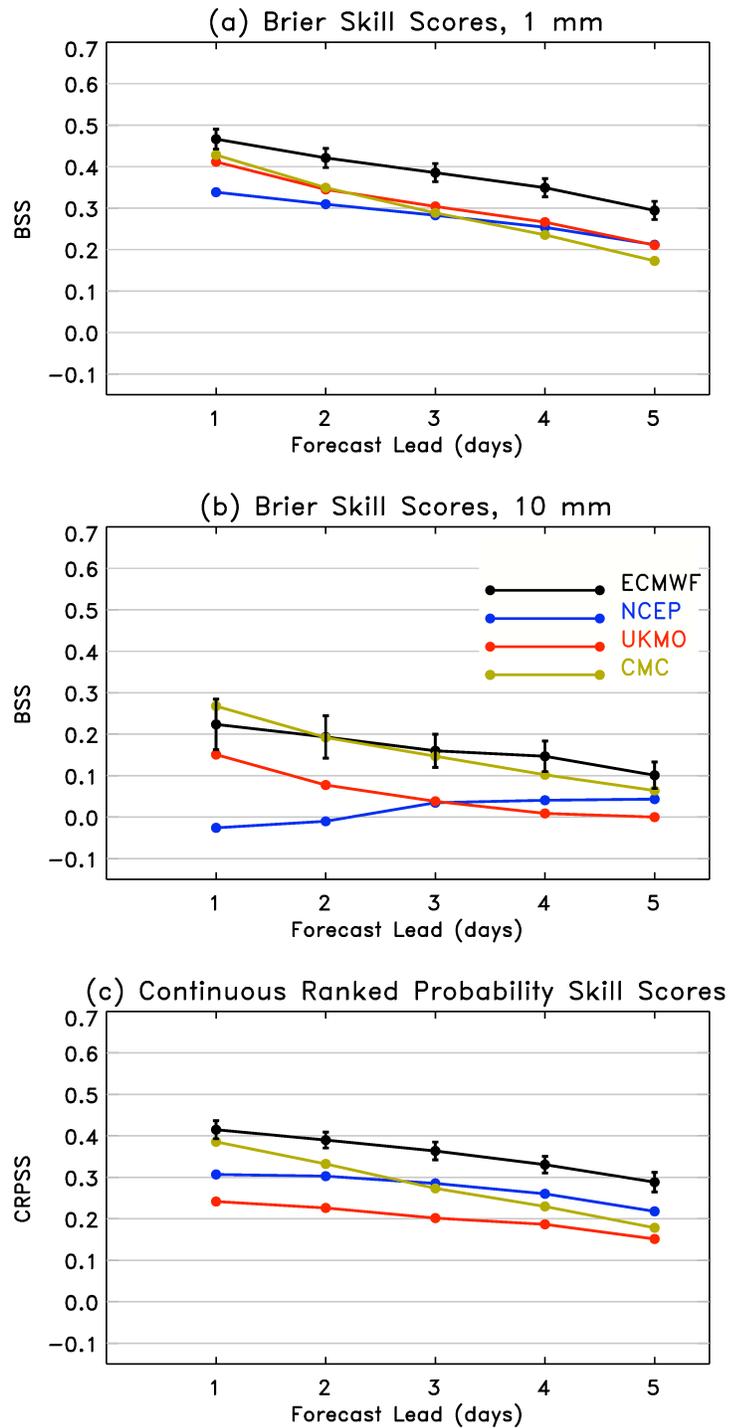
928
929
930



931
932
933
934
935
936
937
938

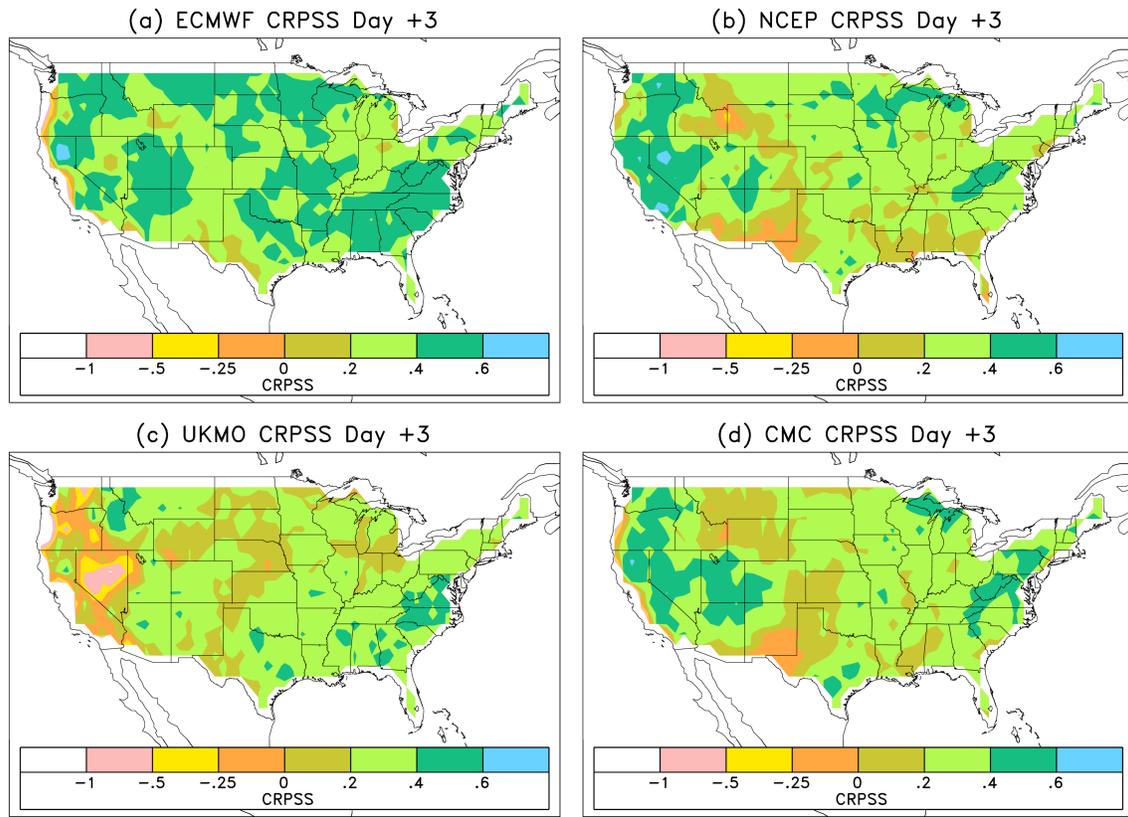
Figure 1: Illustration of the process for determining precipitation classes used in the calculation of BSS . (a) Climatological probability of $> 1\text{-mm } 24\text{h}^{-1}$ precipitation as determined from Stage-IV data for September 2002-2009. (b) Climatological class assigned to each grid point for September, $1\text{-mm } 24\text{ h}^{-1}$ event.

939
940



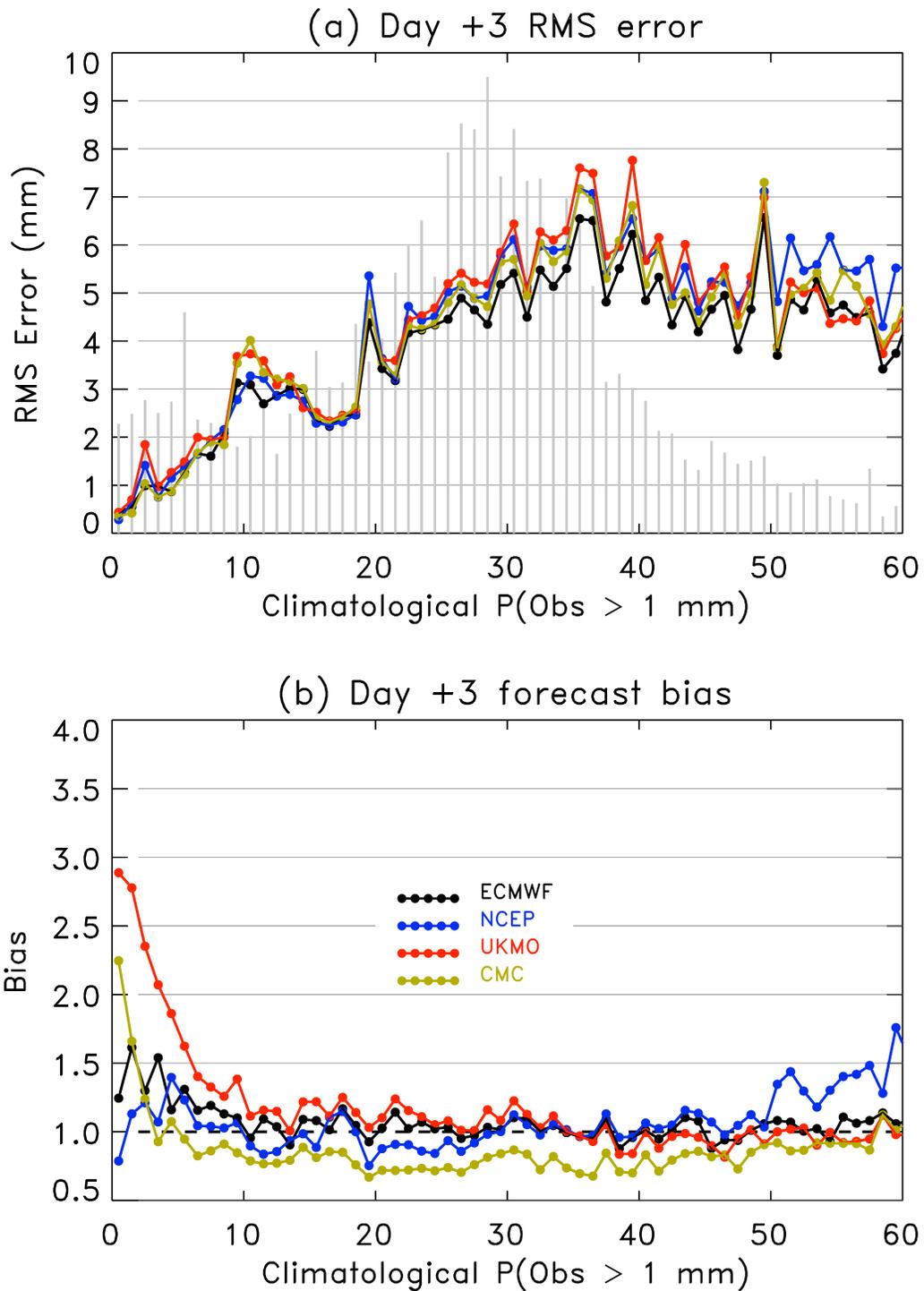
941
942 **Figure 2:** Brier skill scores of various forecasts for the (a) $> 1\text{-mm } 24 \text{ h}^{-1}$ event, (b)
943 $> 10\text{-mm } 24 \text{ h}^{-1}$ event, and (c) continuous ranked probability skill scores, all as a
944 function of forecast lead time. Error bars denote confidence intervals, the 5th and
945 95th percentiles of a paired block bootstrap between ECMWF and NCEP forecasts.
946

947



948
949
950
951
952

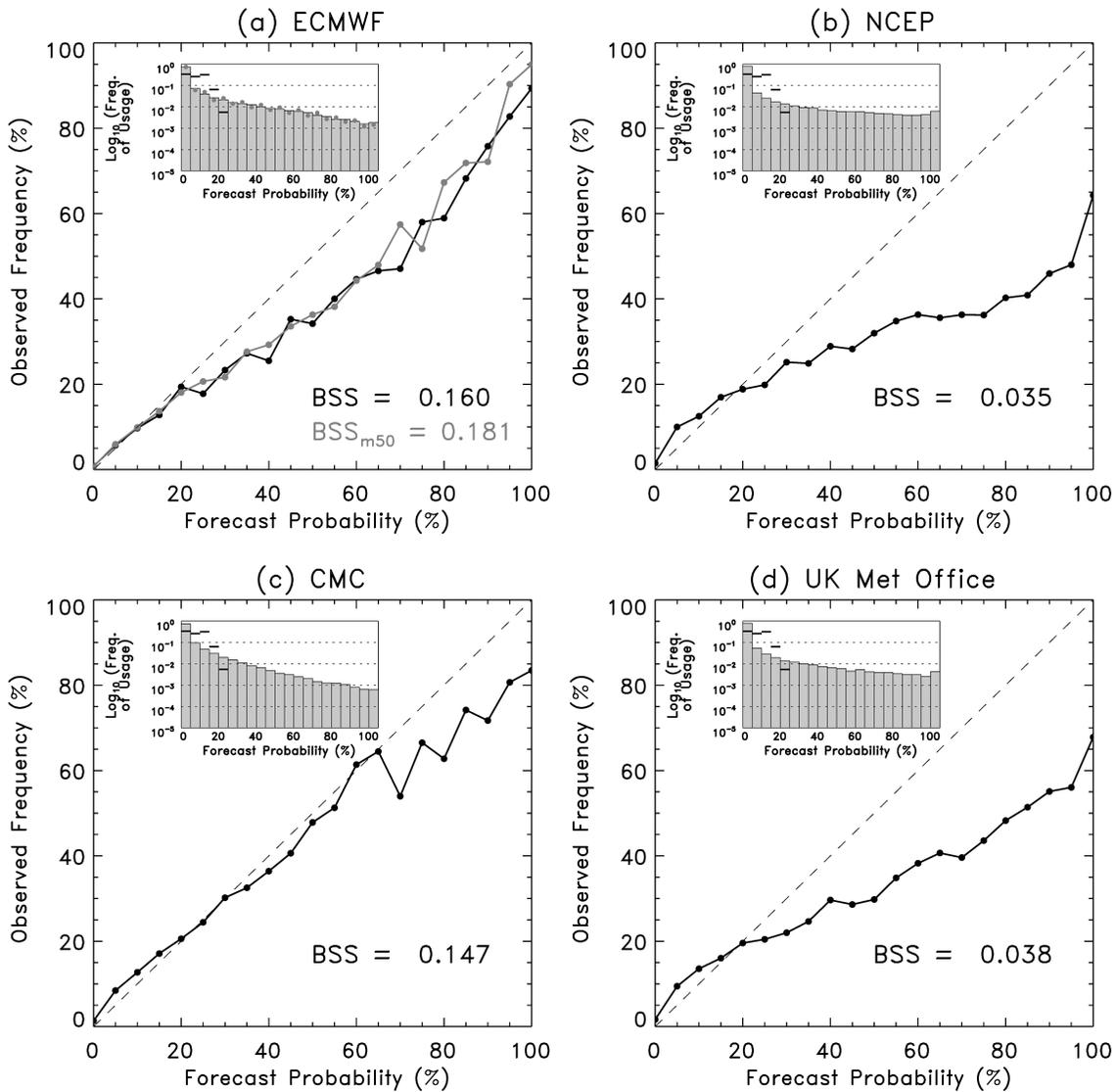
Figure 3: Maps of average *CRPSS* for day +3 forecasts for (a) ECMWF, (b) NCEP, (c) UKMO, and (d) CMC.



954
 955
 956
 957
 958
 959
 960

Figure 4: (a) RMS errors, and (b) bias for day +3 forecasts, each as a function of the climatological probability of greater than 1-mm 24 h⁻¹. Light grey bars in panel (a) denote the relative frequency of each climatological probability.

Reliability, Day +3 10.0mm



962

963

964

965 **Figure 5:** Reliability diagrams for day +3 forecasts for the > 10-mm 24 h⁻¹ event. (a)

966 ECMWF, (b) NCEP, (c) CMC, and (d) UKMO. The dark line on each is the 20-member

967 reliability curve. The lighter grey line on panel (a) is the reliability for the full 50-

968 member ensemble. The inset histogram bars show the relative frequency of usage

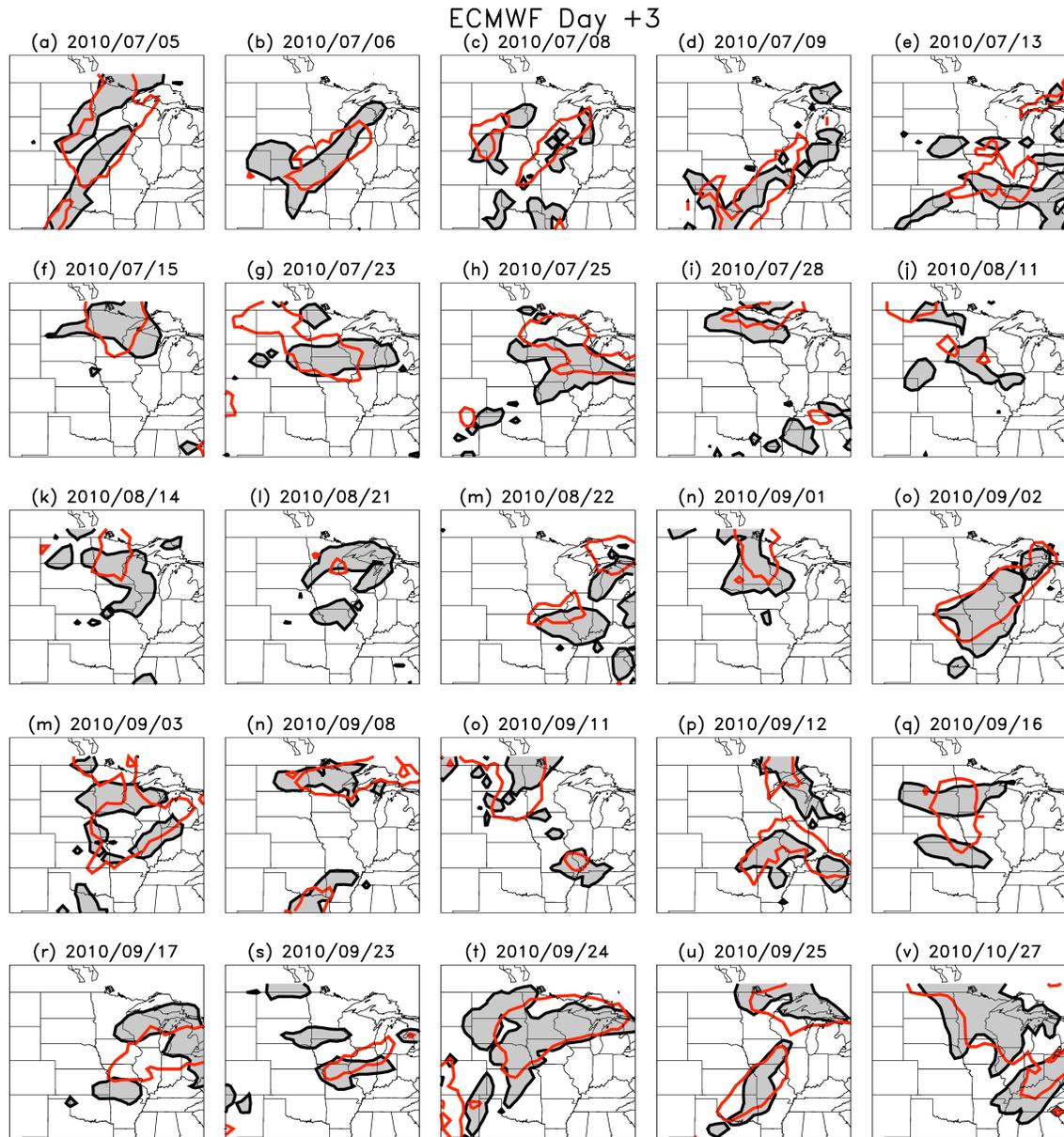
969 for each probability bin. The black lines on the inset are the relative frequency of

970 usage for the climatological distribution across all the sample points. The grey dots

971 on the inset histogram of panel (a) are the relative frequency of usage for the

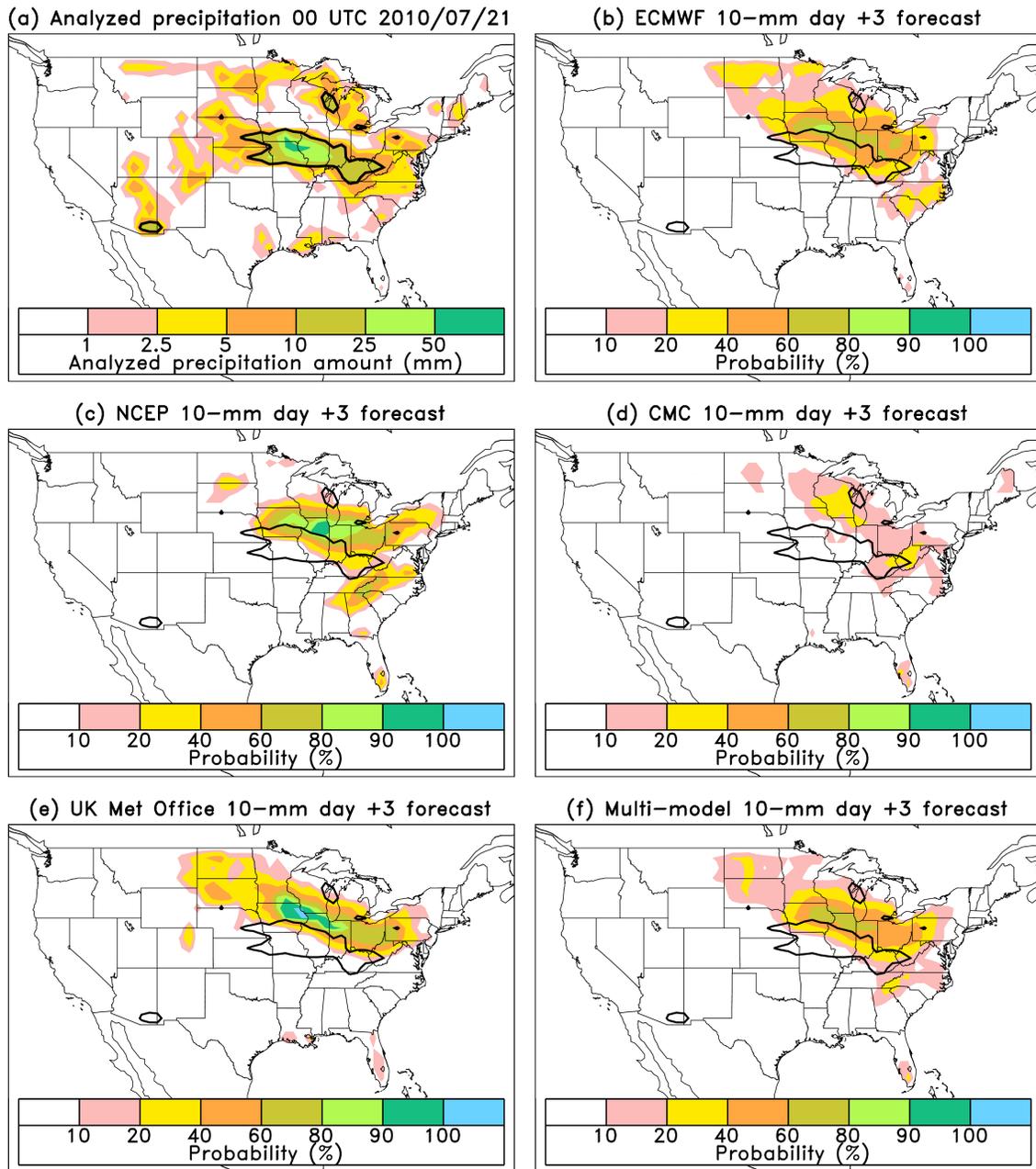
ECMWF full 50-member ensemble.

972
973



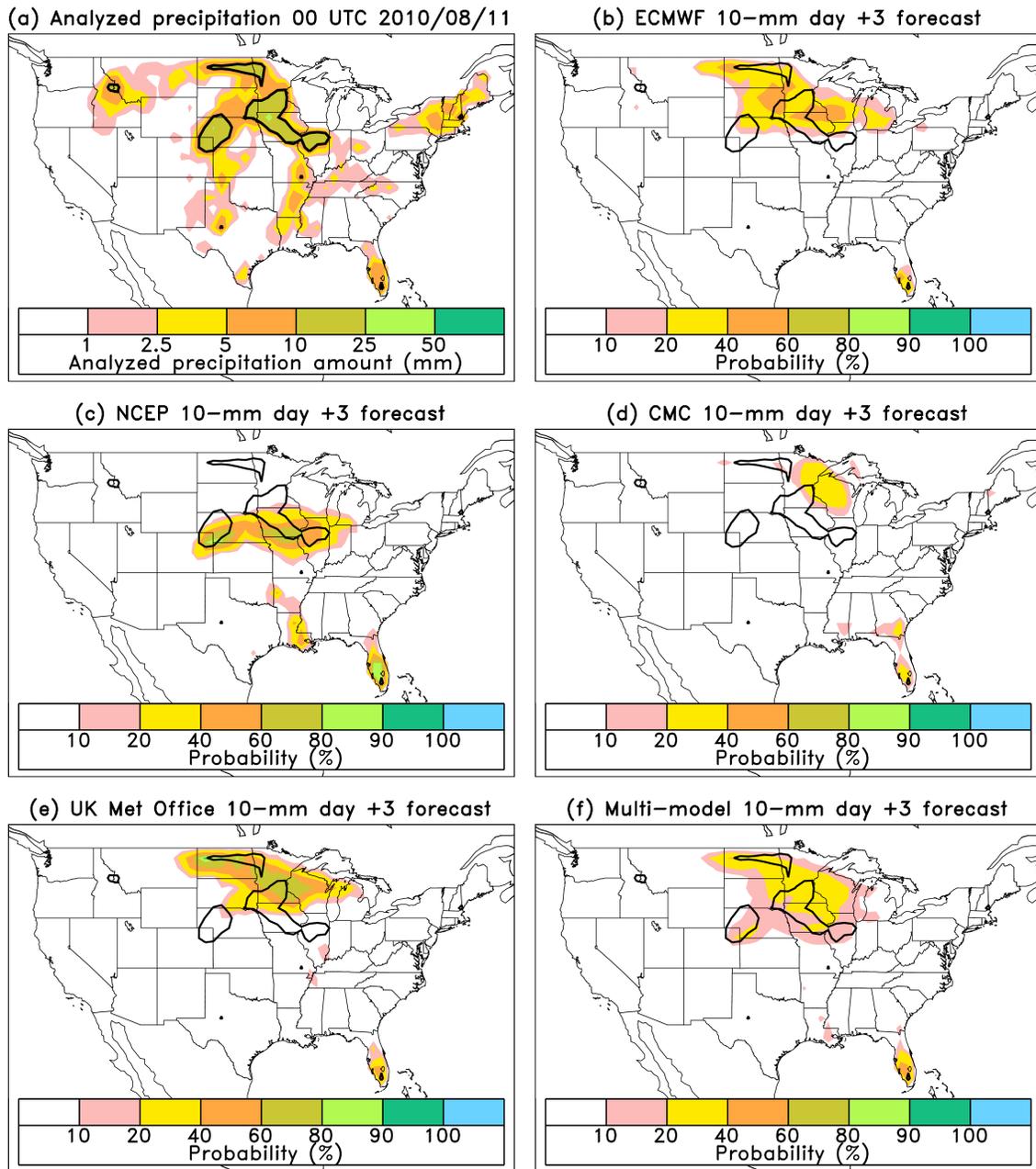
974
975
976
977
978
979
980

Figure 6: Analyzed $> 10\text{-mm } 24 \text{ h}^{-1}$ precipitation boundary (black line) and area exceeding 10 mm (grey shading) for 25 cases with the largest areal coverage of greater than 10 mm in the upper Midwest US. Red lines indicate the 0.5 probability contour from the ECMWF ensemble for the day +3 forecasts of $> 10 \text{ mm } 24 \text{ h}^{-1}$.



982
 983
 984
 985
 986
 987
 988

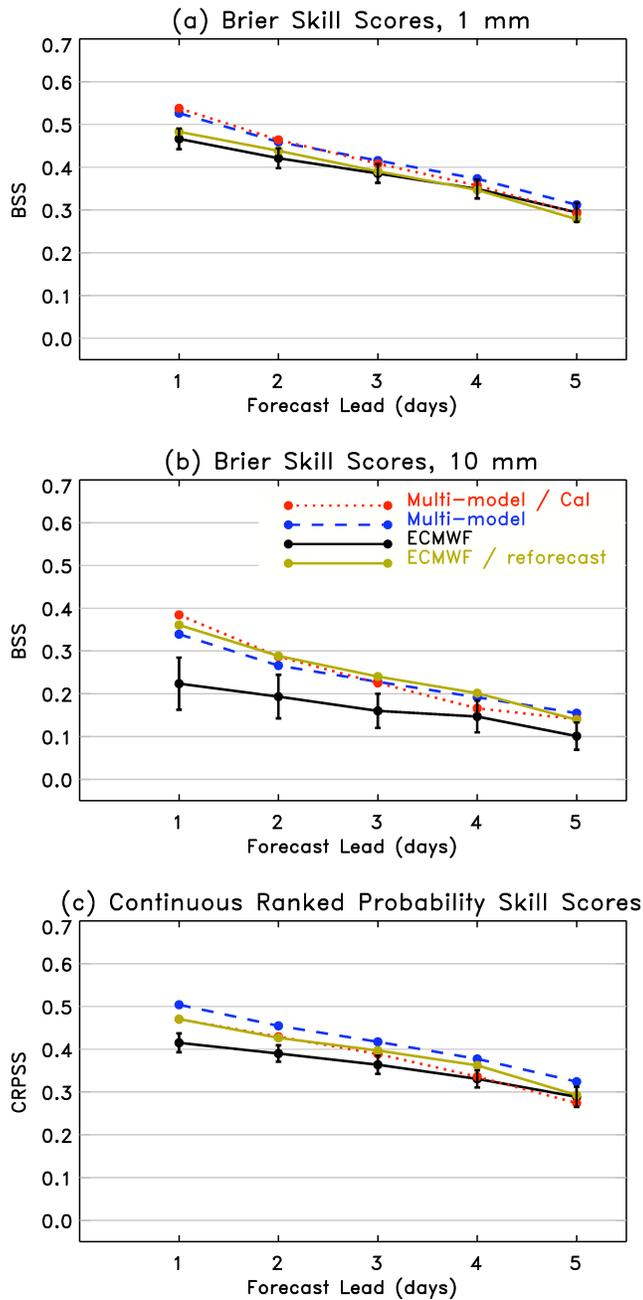
Figure 7: (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010. 10-mm 24 h⁻¹ contour is denoted by the thick black line. (b) Probability of greater than 10 mm 24 h⁻¹ for day +3 forecast from the ECMWF ensemble for the same period. The analyzed 10-mm contour from panel (a) is repeated. (c) as in (b), but for NCEP. (d) CMC, (e) UK Met Office, and (f) multi-model combination.



990
991
992
993

Figure 8: As in Fig. 7, but for 24-h period ending 00 UTC 8 August 2010.

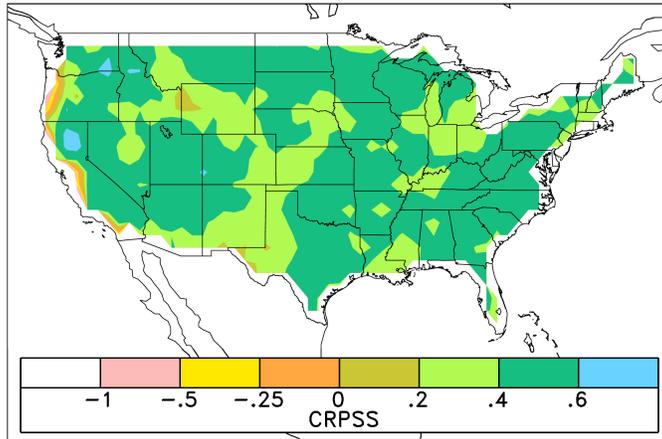
994
995



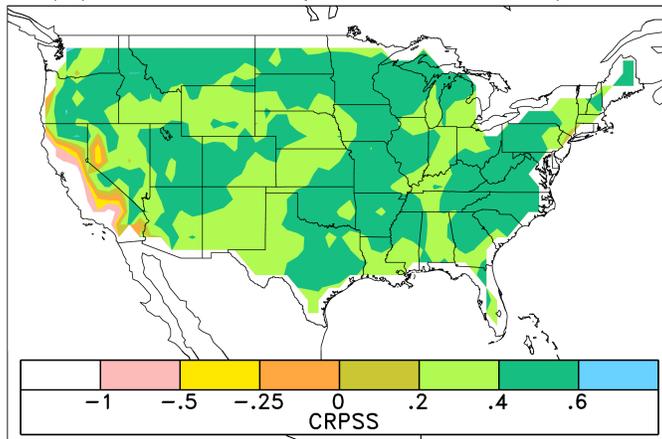
996
997
998
999
1000
1001
1002
1003
1004

Figure 9: Brier skill scores of various forecasts for (a) > 1-mm 24 h^{-1} event, and (b) > 10-mm 24 h^{-1} event, and (c) continuous ranked probability skill scores, all as a function of forecast lead time. “Multi-model/cal” refers to forecasts from the multi-model, calibrated using ELR. “ECMWF/reforecast” refers to ECMWF forecasts calibrated using ELR and the reforecast data set. Error bars denote confidence intervals, the 5th and 95th percentiles of a paired block bootstrap between ECMWF and NCEP forecasts.

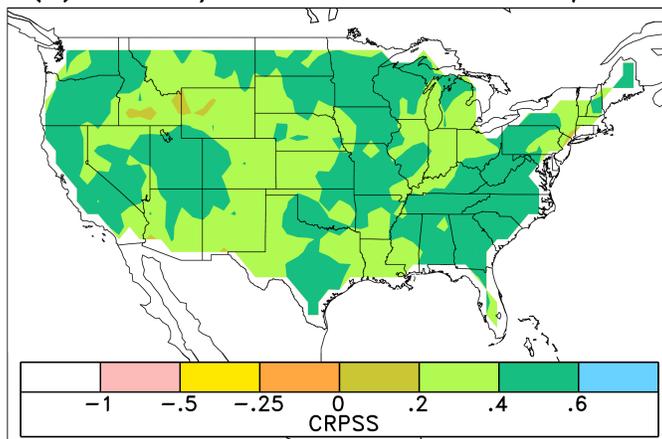
(a) Multi-model CRPSS, day +3



(b) Multi-model/calibrated, day +3

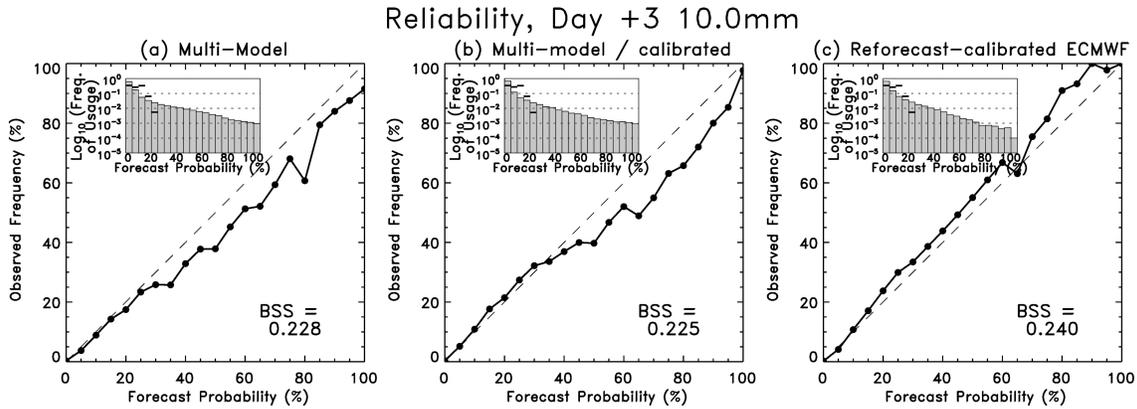


(c) ECMWF/reforecast CRPSS, day +3



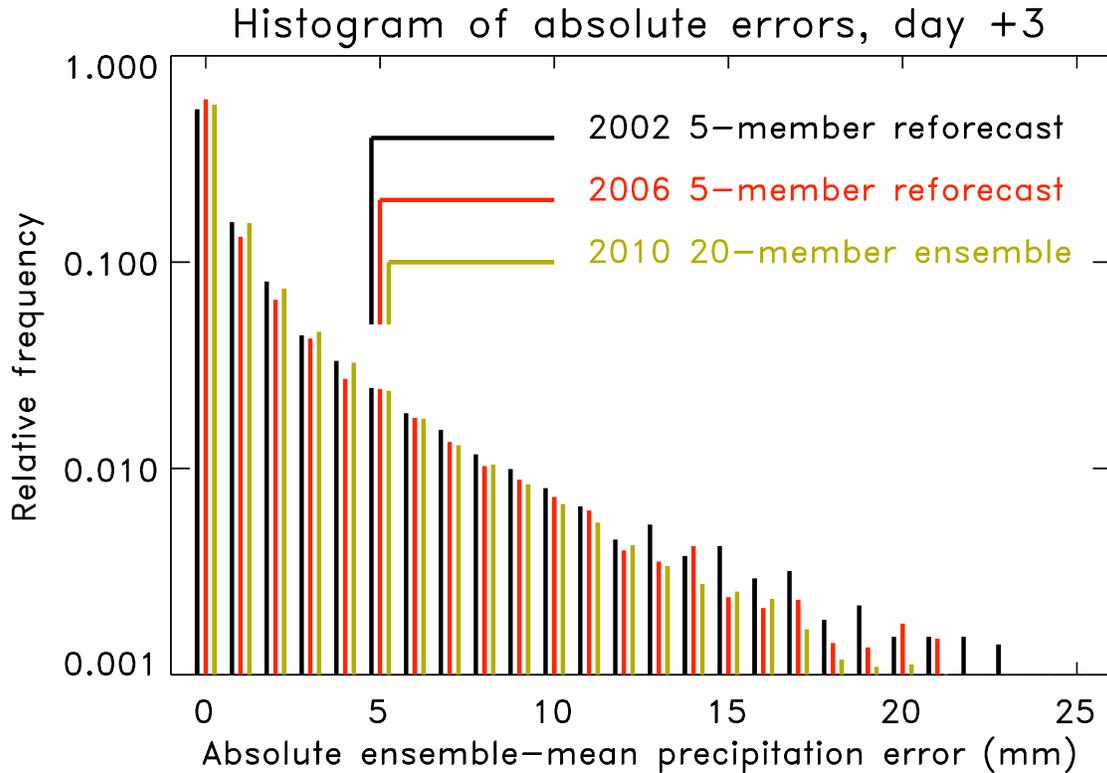
1005
1006
1007
1008

Figure 10: Maps of *CRPSS* for day +3 forecasts for (a) multi-model, (b) multi-model with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.



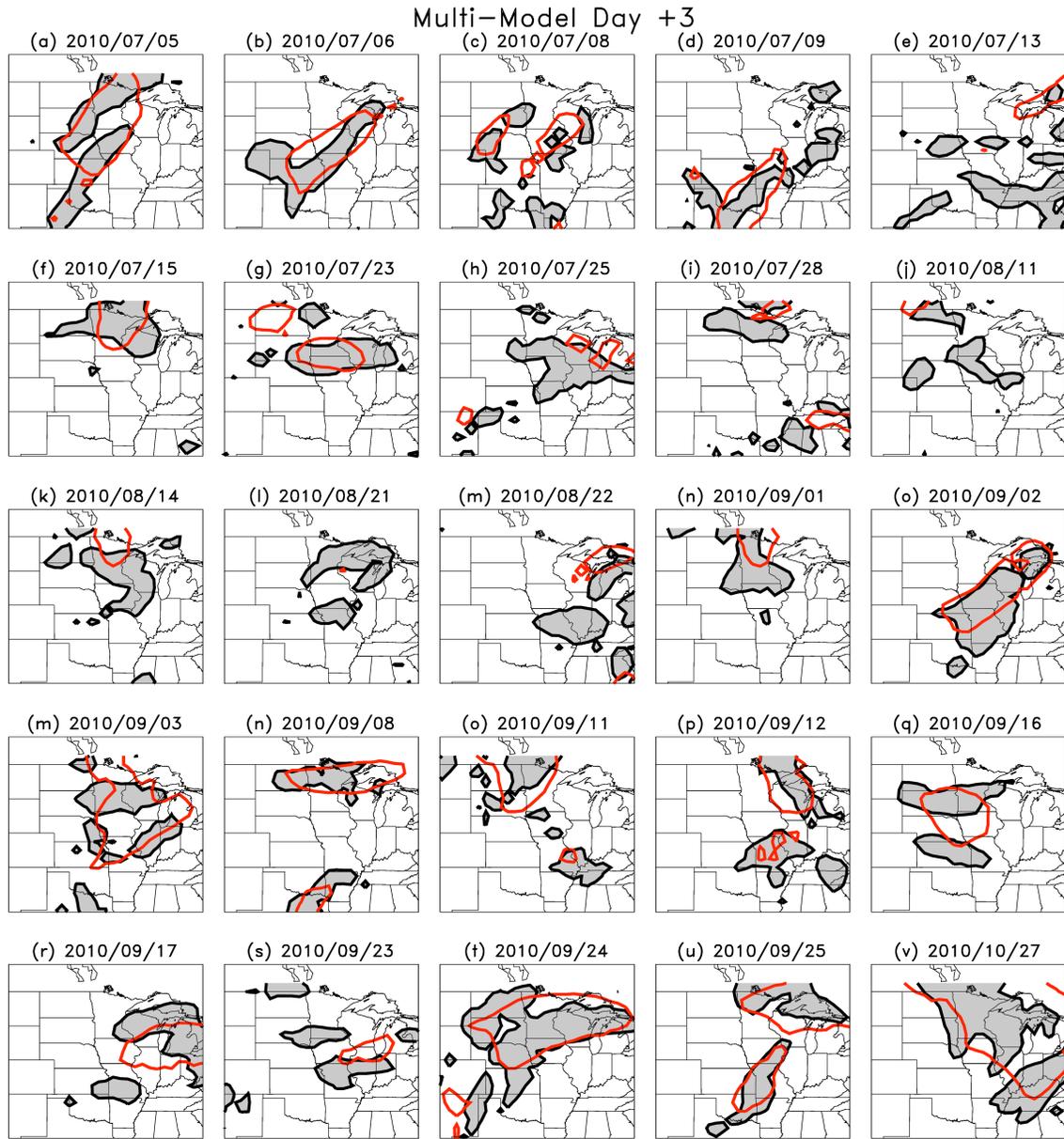
1009
1010
1011
1012
1013
1014
1015

Figure 11: As in Fig. 3, reliability diagrams for day +3 forecasts at the > 10-mm 24 h⁻¹ event. (a) Multi-model forecasts, (b) multi-model with ELR calibration, and (c) ECMWF with ELR calibration using reforecasts.



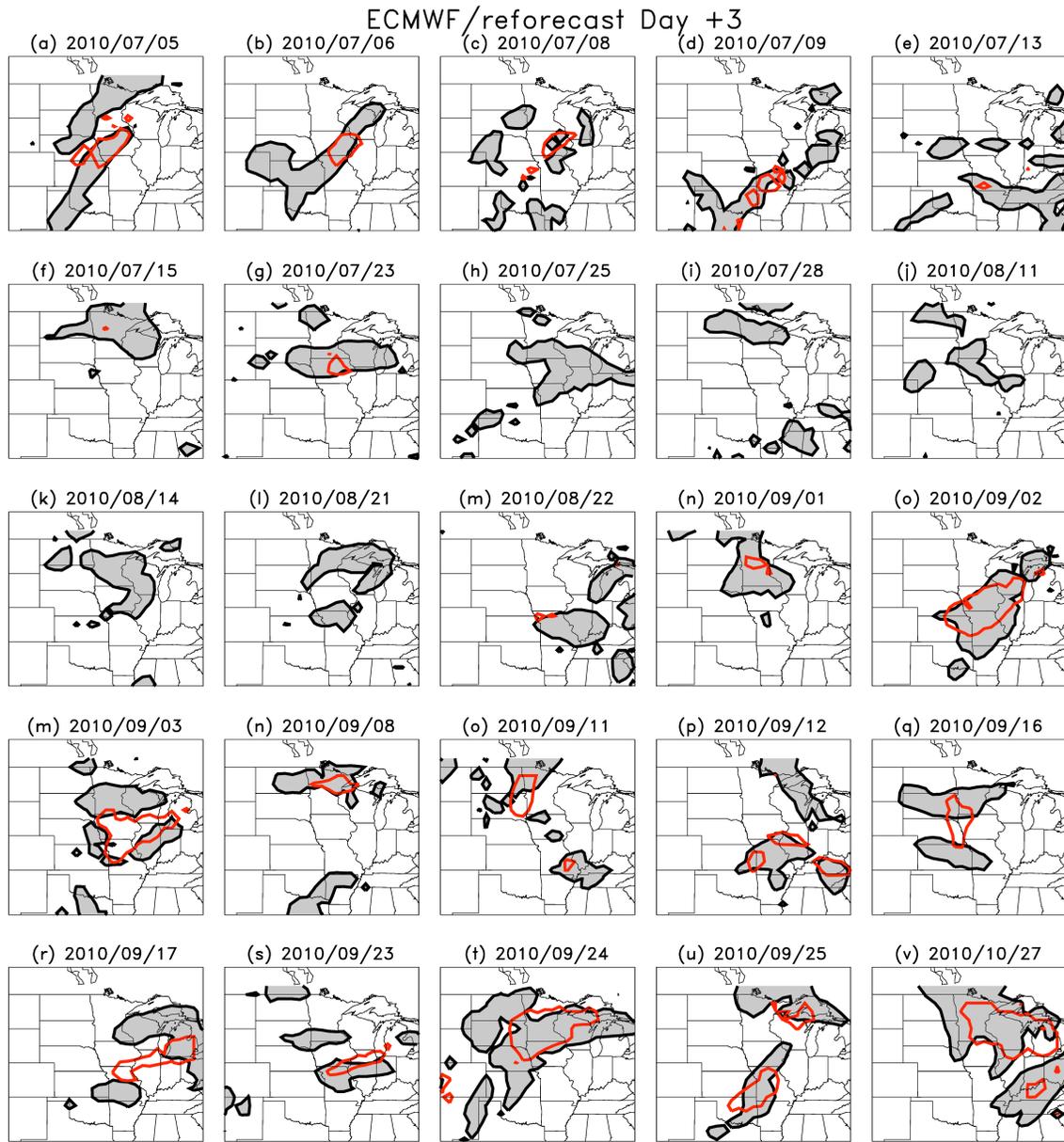
1016
1017
1018
1019
1020
1021
1022

Figure 12: A histogram of the absolute errors of day +3 ensemble-mean precipitation forecasts for the 2002 and 2006 reforecasts and for the 2010, 20-member real-time ensemble.



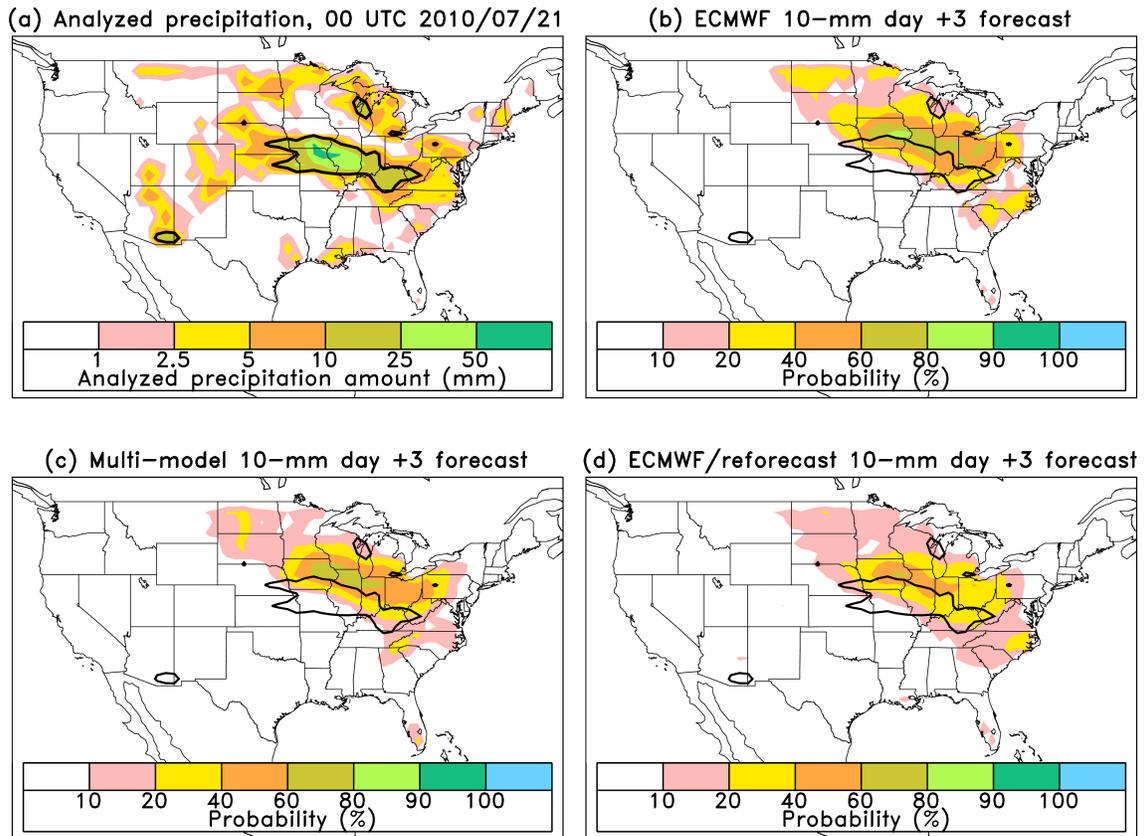
1024
1025
1026
1027

Figure 13: As in Fig. 6, but for multi-model forecasts.



1029
1030
1031
1032

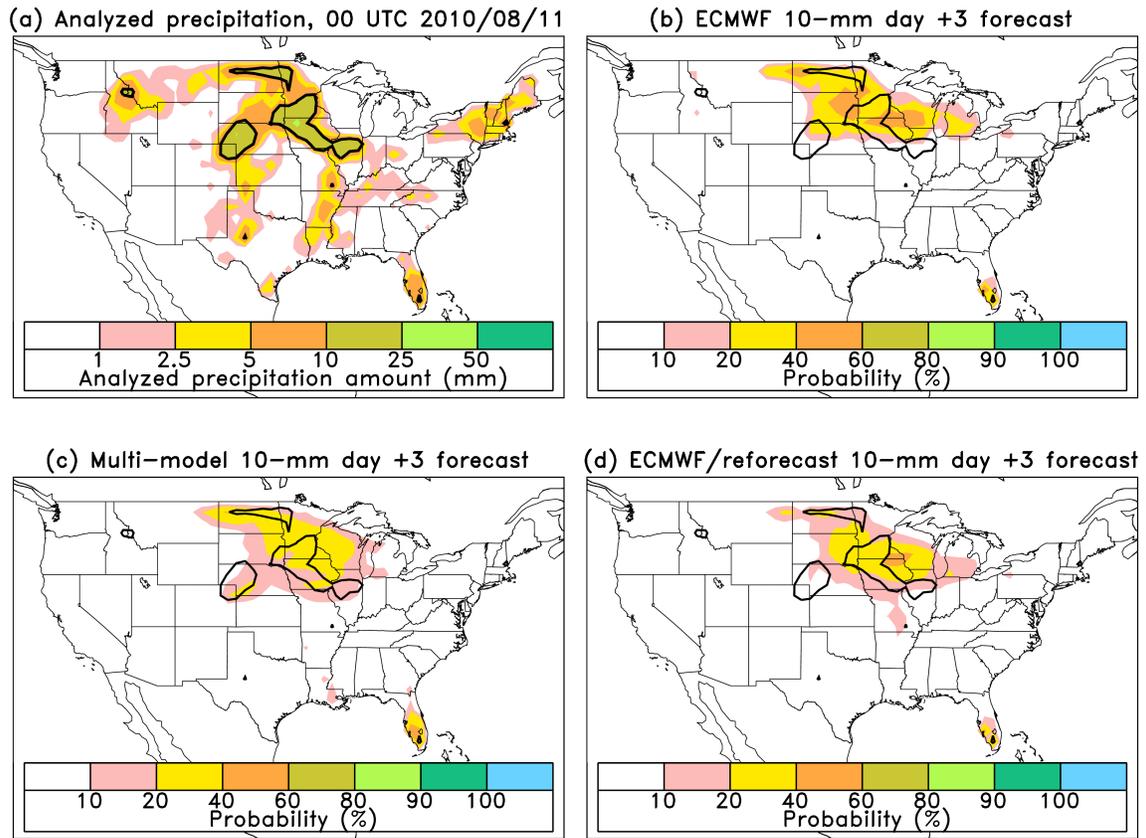
Figure 14: As in Fig. 6, but for reforecast-calibrated ECMWF forecasts.



1034
 1035
 1036
 1037
 1038
 1039
 1040
 1041
 1042

Figure 15: (a) Analyzed precipitation for the 24-h period ending 00 UTC 21 July 2010. 10-mm contour is denoted by the thick black line. (b) Probability of greater than 10 mm 24 h⁻¹ for day +3 forecast from the ECMWF ensemble for the same period. (c) as in (b), but for multi-model ensemble, and (d) as in (b), but for reforecast-calibrated ECMWF ensemble.

1043



1044
1045
1046
1047

Figure 16: As in Fig. 15, but for the 24-h period ending 00 UTC 8 August 2010.