The science behind multi-model precipitation probability in the "National Blend of Models"

Tom Hamill and Michael Scheuerer

ESRL, Physical Sciences Division

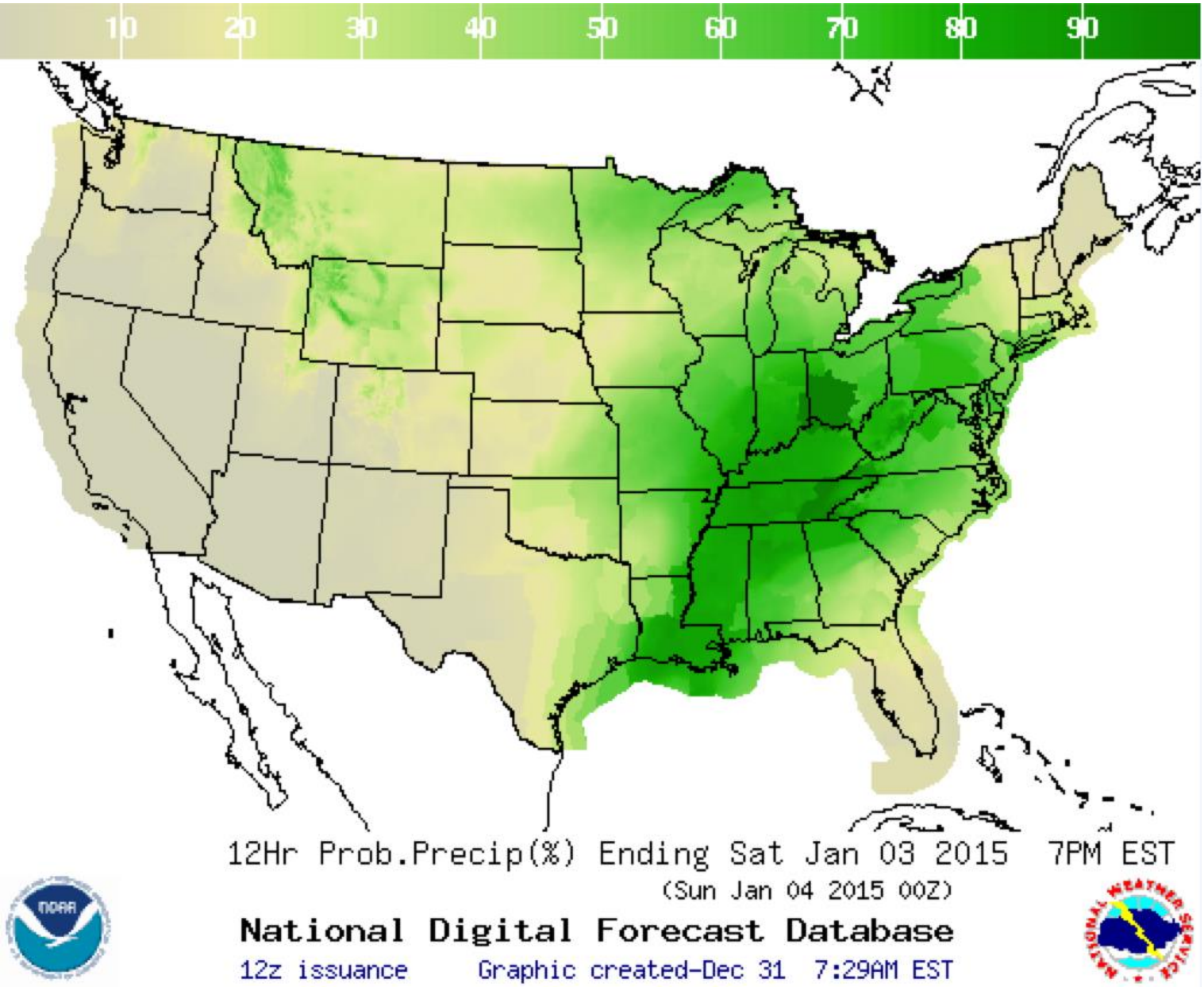
tom.hamill@noaa.gov; michael.scheuerer@noaa.gov

(303) 497-3060; (303) 497-4281

“National Blend of Models” project

- Improve NWS National Digital Forecast Database (NDFD) fields using multi-model ensemble and multiple deterministic forecasts.
  - Reduce need for forecaster manual intervention, free up forecasters for more decision support.
- Eventually “Blend” will cover:
  - all NDFD variables (e.g., T_{max}, T_{min}, T, T_d, wind speed, direction, gust, sky cover, precip amount, snow amount, wave height) on 2.5-km grid.
  - all US regions (CONUS, AK, HI, PR, Guam)
- May expand to PQPF in the future
- PSD’s role: improve multi-model precipitation, and more specifically here, POP. Asked to focus on +3 to +8 day period initially.

NDFD POP product example



One's eye is drawn to major changes in POP at WFO boundaries.

Can we make POP so skillful and reliable that forecasters will need to modify centralized guidance less frequently?

Challenge:

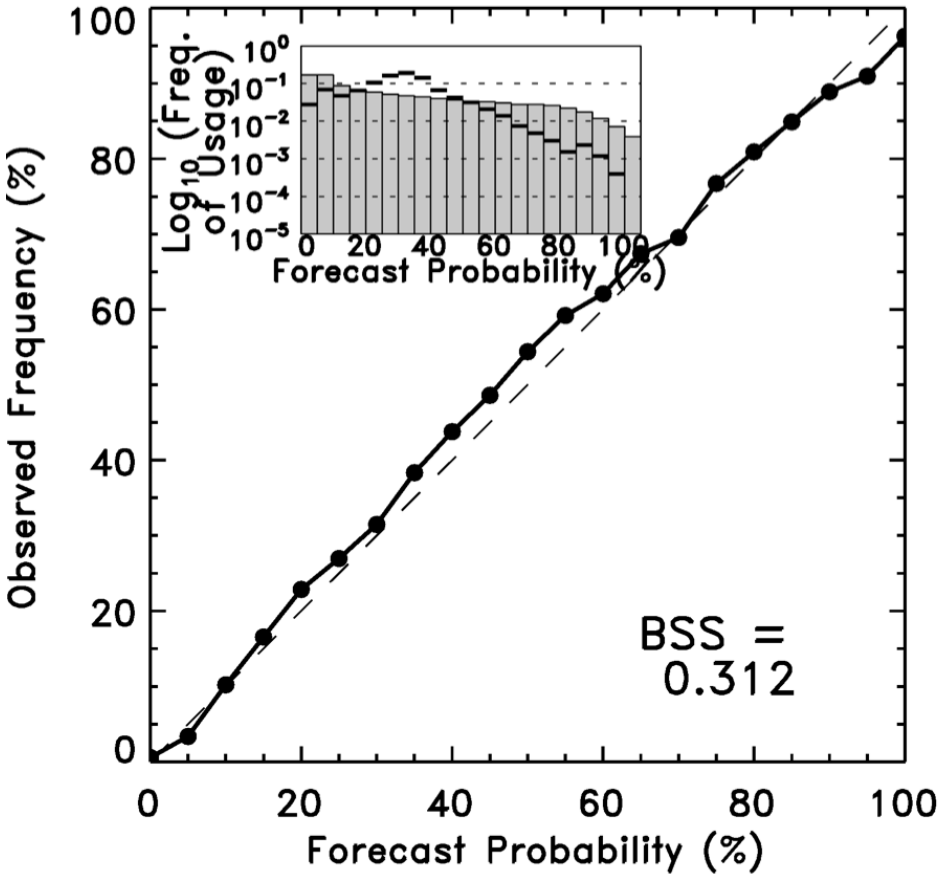
for NWS "National Blend" project, little training data exists, or it's hard to get (and archive).

- **GEFS**: will change soon to new version, obviating the existing 00Z cycle reforecasts.
- **ECMWF**: ~20-year, 1x weekly reforecasts, but have not been made available to us (yet).
- **CMC**: ~ 5 years of reforecasts based on ERA-Interim with near-surface adjustment; not available for this project (yet).
- No reforecasts for control/deterministic of various centers.

This limits the range of post-processing methodologies we feel comfortable trying.

What might we reasonably do to post-process multi-model ensemble guidance?

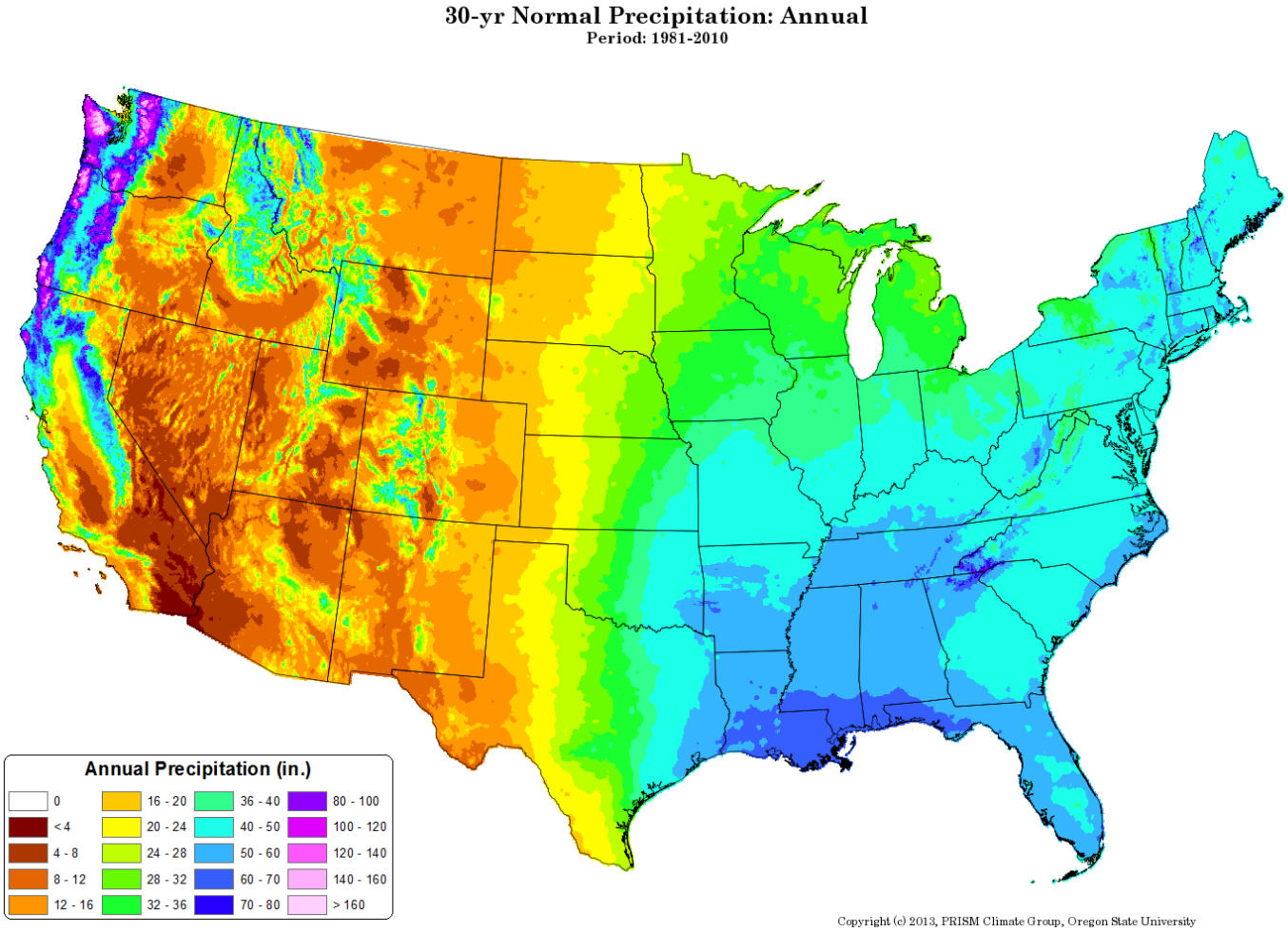
Multi-model reliability, Day +5, > 1 mm



- Previously (Hamill, July 2012 *MWR*) multi-model ensembles for POP (July-Oct 2010) averaged to 1 degree grid verified against 1-degree CCPA.
- This procedure produced reliable and skillful POP forecasts.
- Models were ECMWF, UK Met Office, CMC, and NCEP.

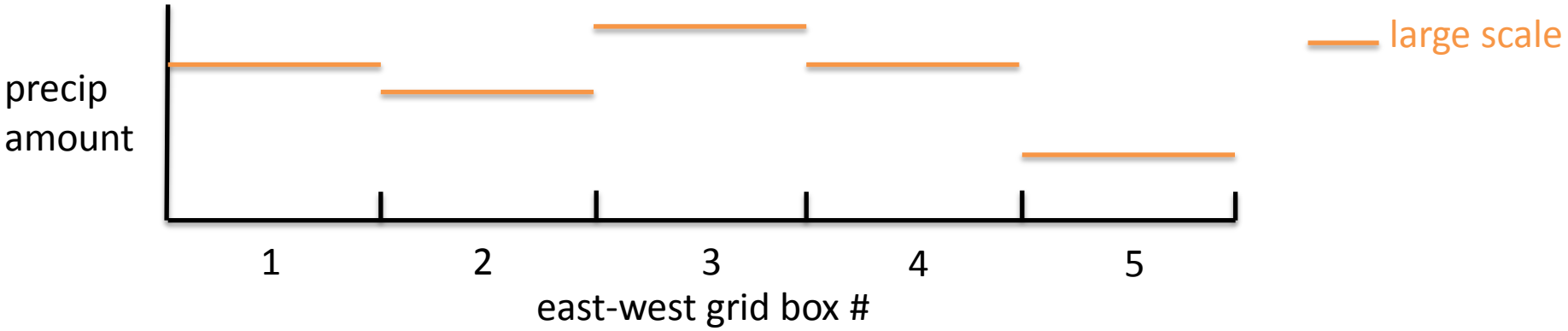
1-degree forecasts are not high-enough resolution for Nat'l Blend; desire on 2.5-km grid.

Why we don't expect >10-km global ensemble forecasts to be reliable when verified against < 10-km analyses.



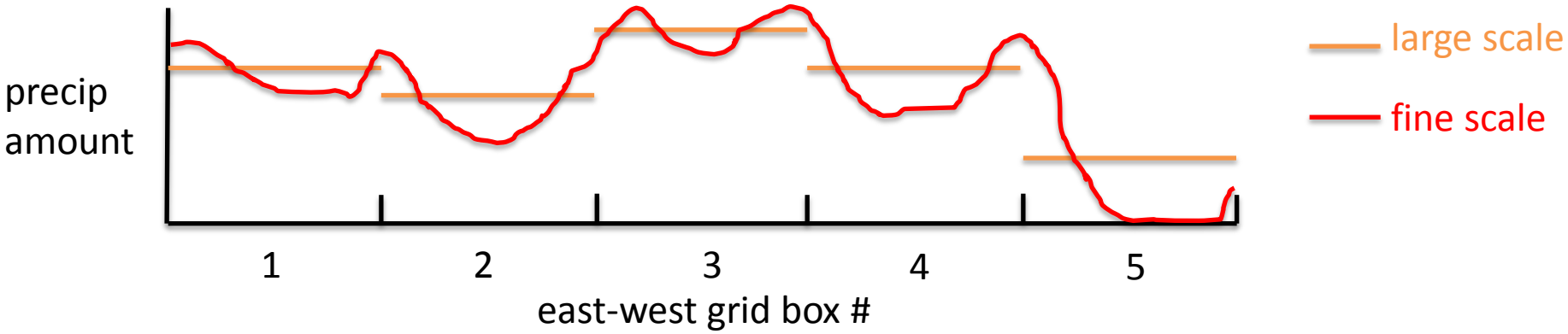
Problem 1: there is coherent climatological detail, especially in the western US at scales < 10 km, below the model-resolvable scales.

Problem 2: sub-grid variability



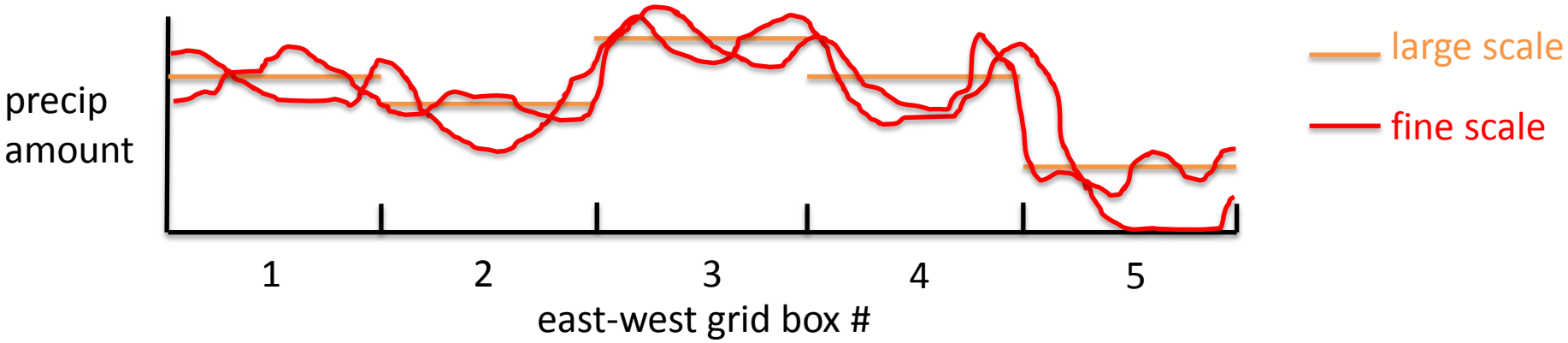
Here we have a precipitation forecast along a segment of a latitude circle.

Problem 2: sub-grid variability



Here's one possible fine-scale analysis that is consistent with the large scale.

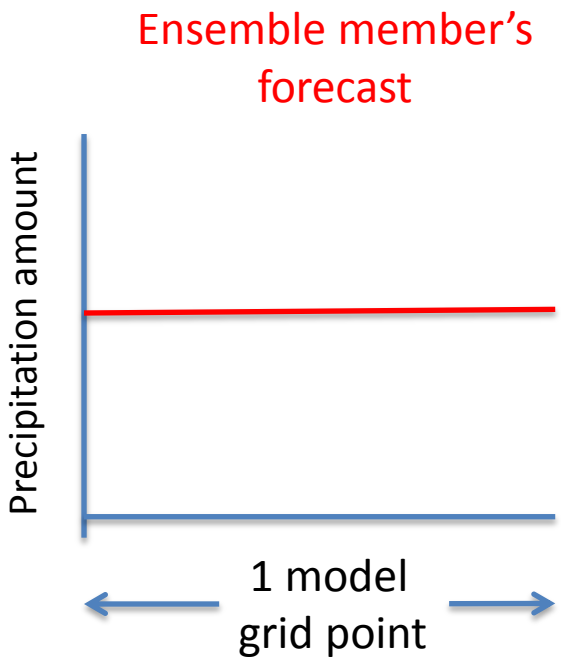
Problem 2: sub-grid variability



But here's yet another. There are infinitely many fine-scale vectors that will spatially average to the correct large-scale vector.

So, when simulating a high-resolution ensemble, one should account for sub-gridscale variability in a realistic manner.

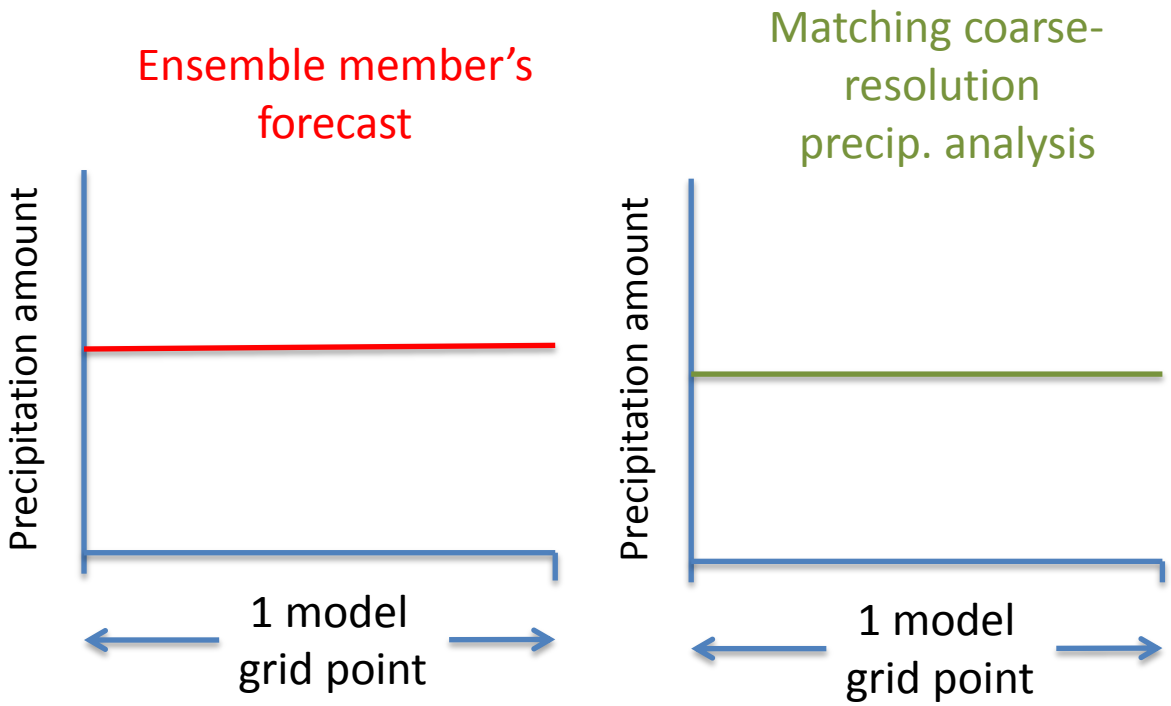
Proposed downscaling methodology



Repeat this process for every member at every grid point:

(1) Consider coarse resolution ensemble member forecast at a given grid point.

Proposed downscaling methodology

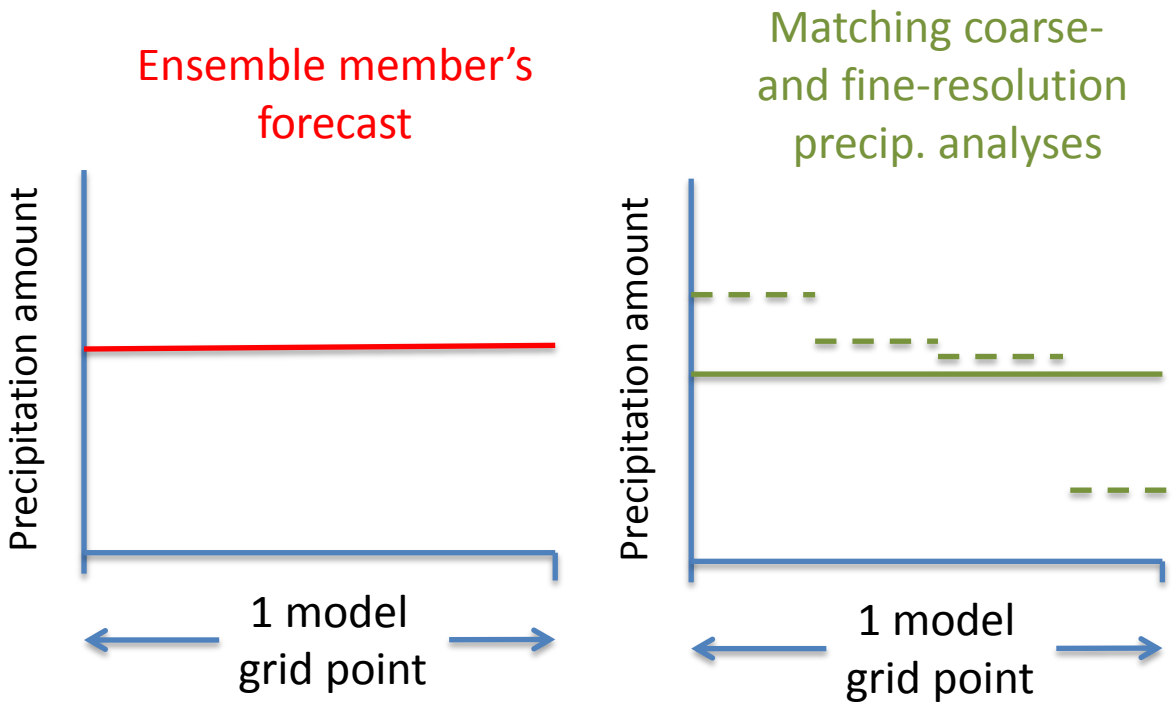


Repeat this process for every member at every grid point:

(1) Consider coarse resolution ensemble member forecast at a given grid point.

(2) Find a past coarse-resolution analysis close in value to this forecast.

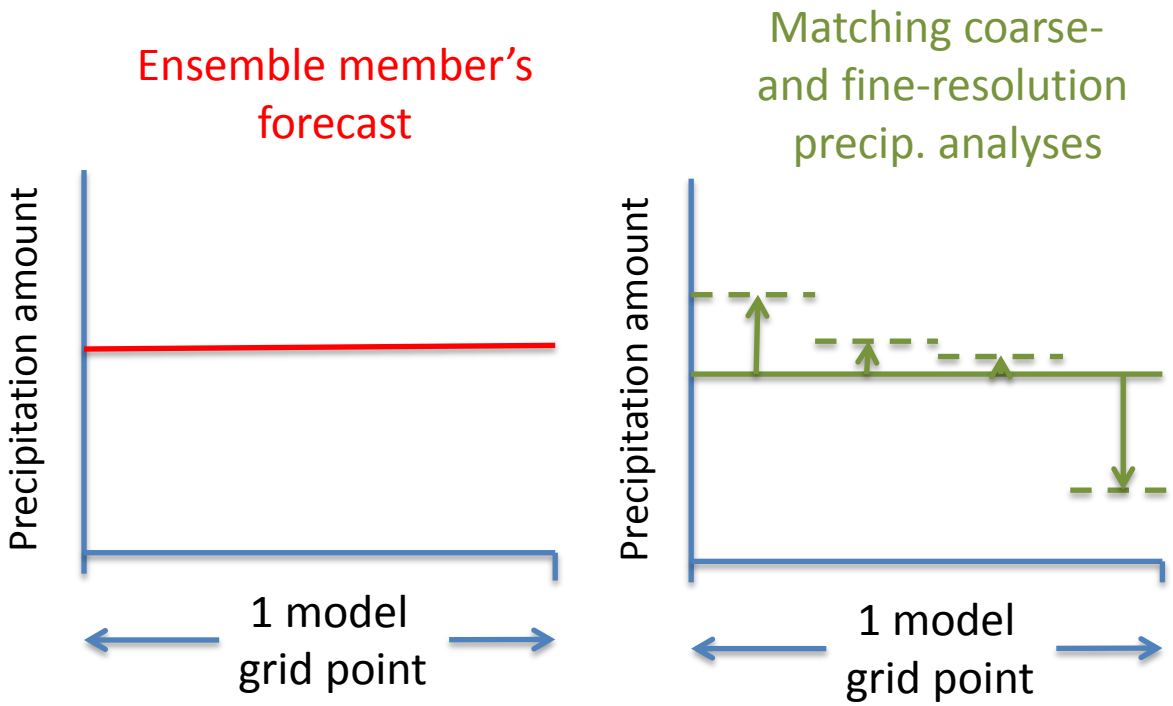
Proposed downscaling methodology



Repeat this process for every member at every grid point:

(1) Consider coarse resolution ensemble member forecast at a given grid point.
(2) Find a past coarse-resolution analysis close in value to this forecast.
(3) Extract the fine-resolution analysis at the same day and location.

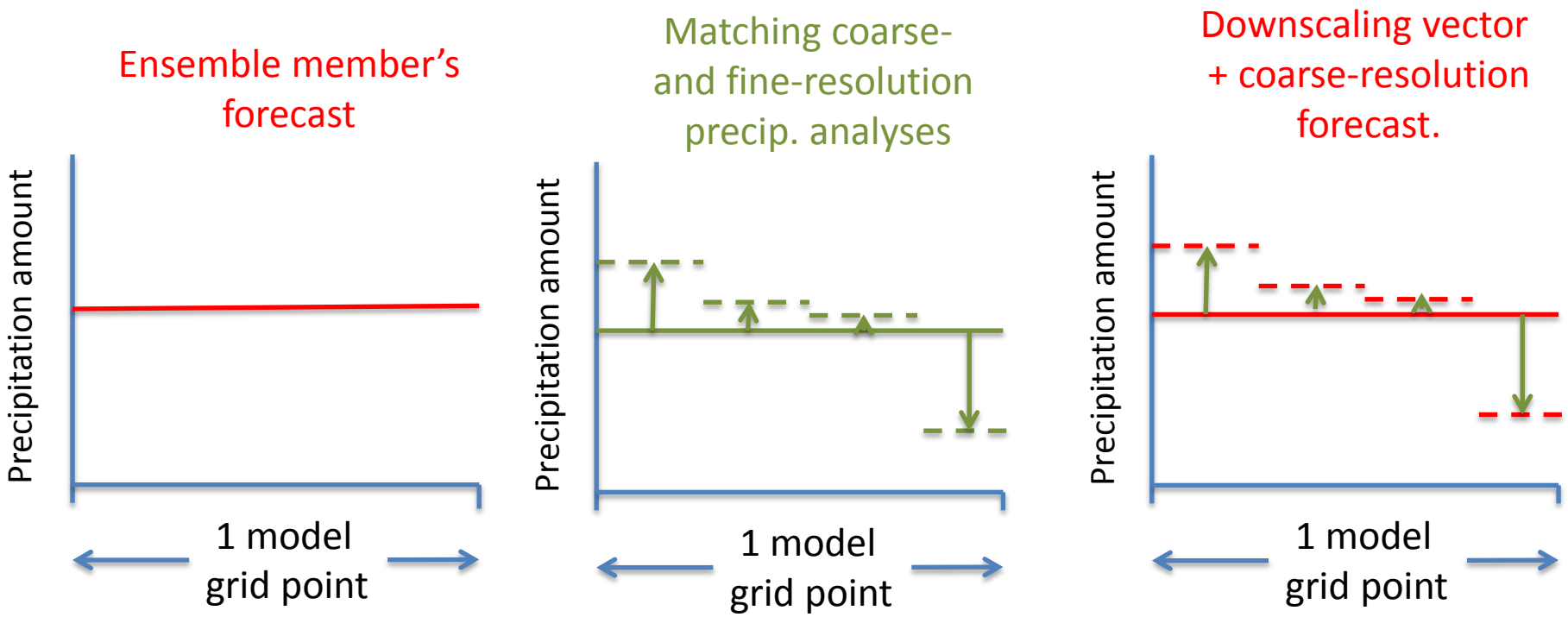
Proposed downscaling methodology



Repeat this process for every member at every grid point:

(1) Consider coarse resolution ensemble member forecast at a given grid point.
(2) Find a past coarse-resolution analysis close in value to this forecast.
(3) Extract the fine-resolution analysis at the same day and location.
(4) Define the downscaling vector.

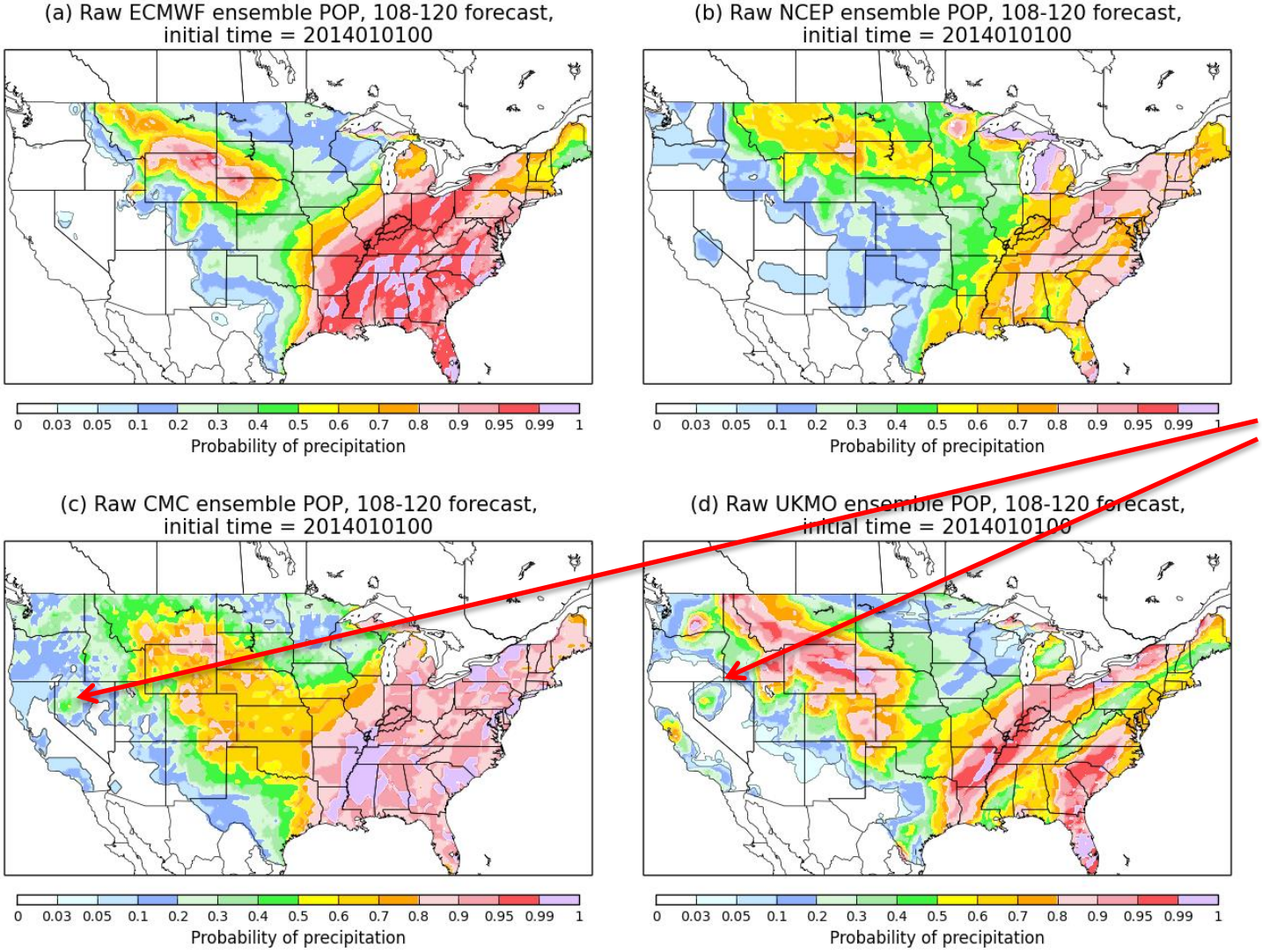
Proposed downscaling methodology



Repeat this process for every member at every grid point:

(1) Consider coarse resolution ensemble member forecast at a given grid point.
(2) Find a past coarse-resolution analysis close in value to this forecast.
(3) Extract the fine-resolution analysis at the same day and location.
(4) Define the downscaling vector.
(5) Add vector to the coarse-res. forecast to define the downscaled forecast.

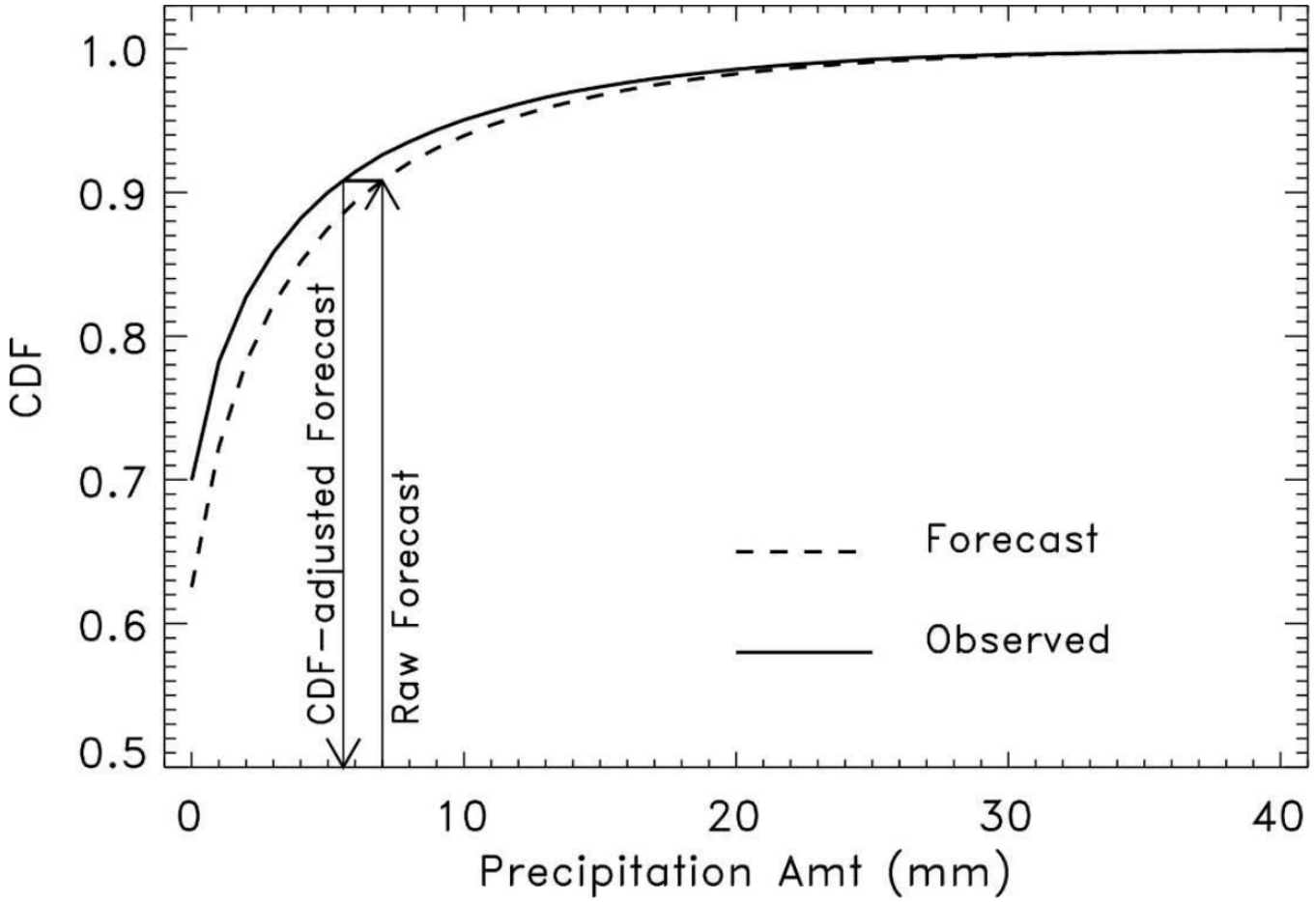
A different problem: systematic biases such as over-forecasting of drizzle.



Some regions, such as in N Nevada, models tend to produce precipitation unrealistically, day after day.

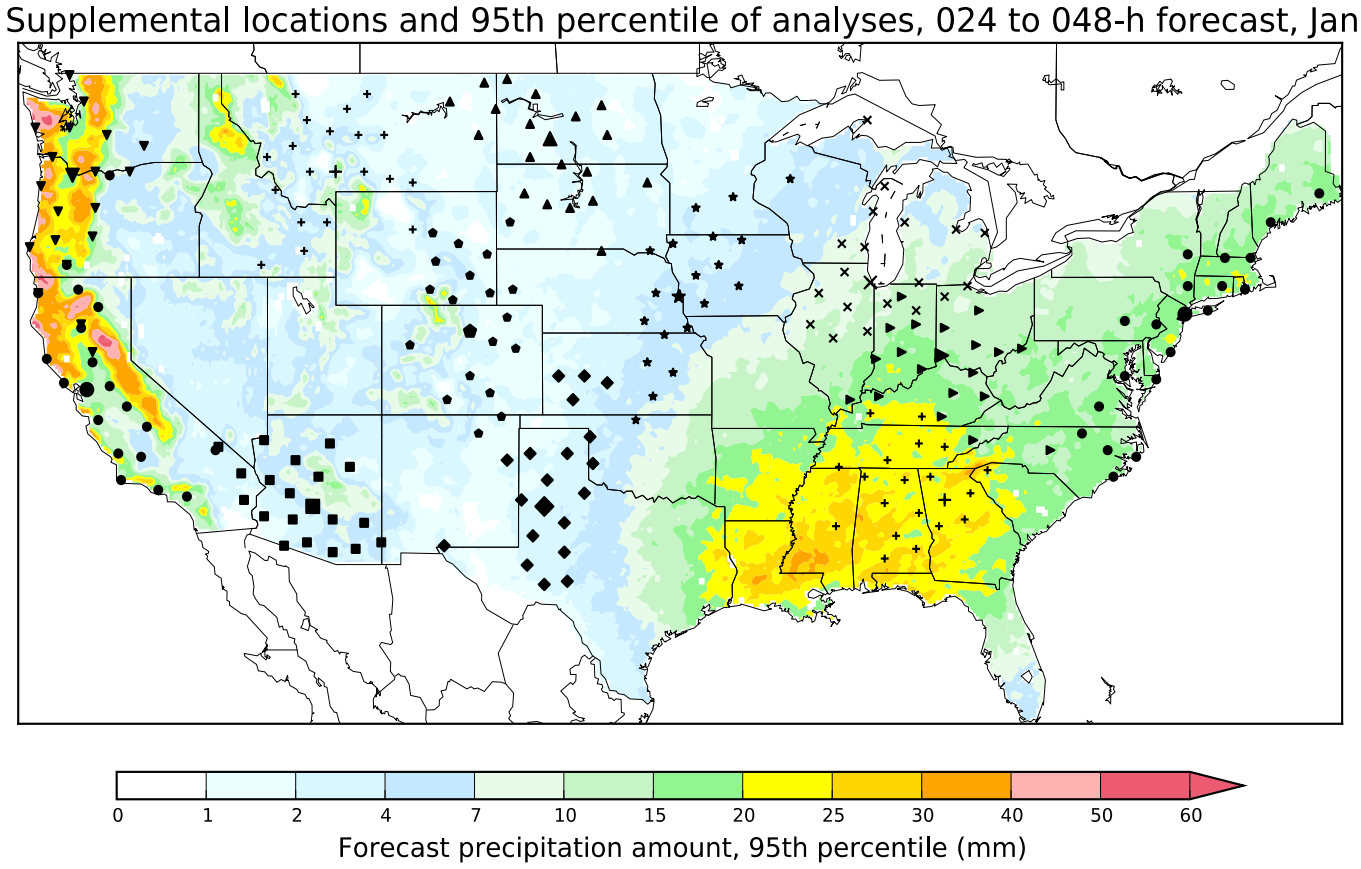
Such systematic errors may affect POP reliability.

CDF-based bias correction, or "quantile mapping"



A challenge is that if the CDFs are generated from small samples (with CDFs computed separately for each grid point) may be noisy if computed using a small training sample size.

Potential remedy for small sample size: use supplemental data locations



Idea is to supplement training data at each grid point using n extra grid points that have similar observed climatology and similar forecast-observed relationships, as determined from GEFS reforecasts.
In this plot, big symbol is where we're training, smaller symbol where we're supplementing training data.

Smoothing the POP forecasts

- Savitzky-Golay (S-G) filter used. For more details, see:
  - http://research.microsoft.com/en-us/um/people/jckrumm/SavGol/SavGol.htm
  - http://www.wire.tu-bs.de/OLDWEB/mameyer/cmr/savgol.pdf (good tutorial)
  - *Numerical Recipes* text.
- I coded up the S-G filter with a window size of 9 and using 3rd-order polynomial fitting.

Savitzky-Golay filter

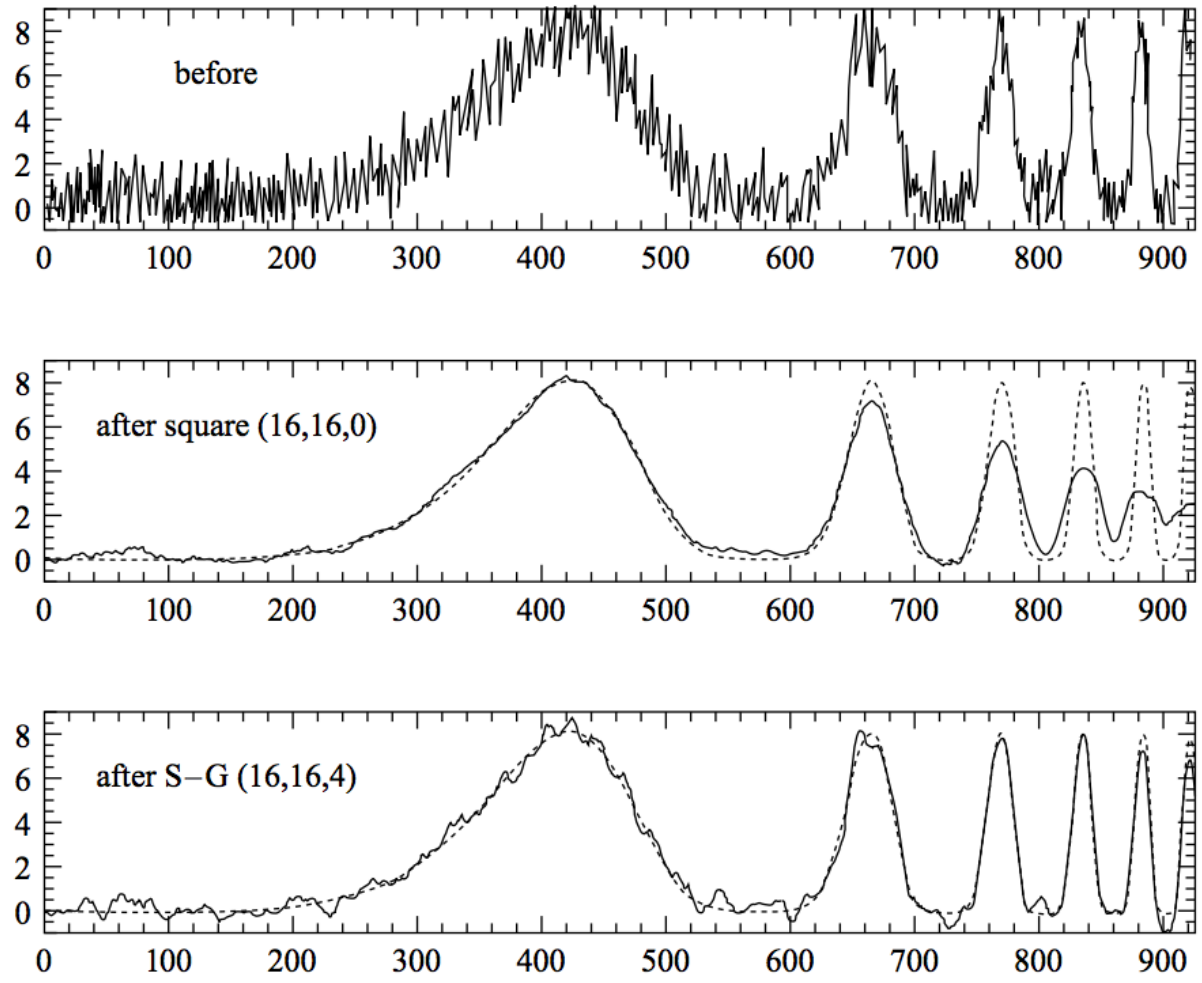
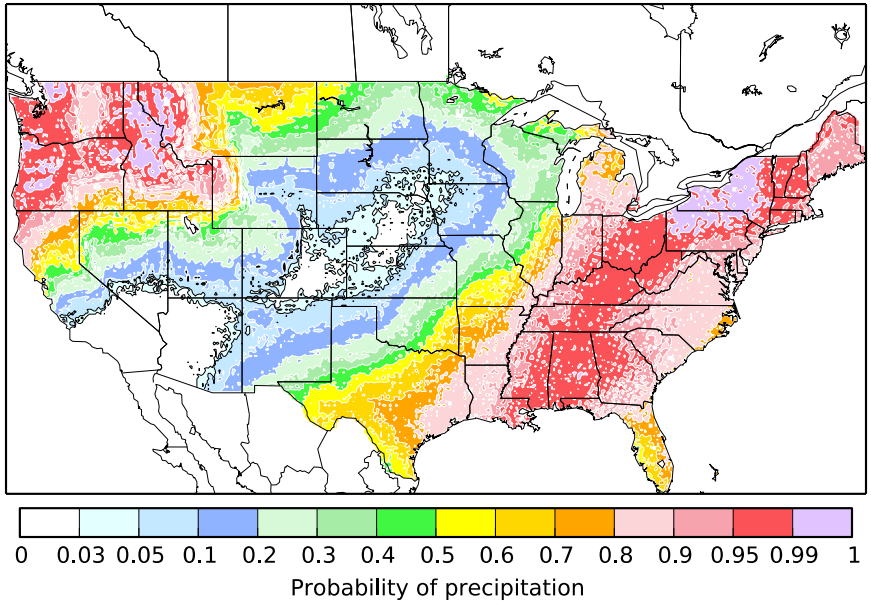


Figure 14.8.1. Top: Synthetic noisy data consisting of a sequence of progressively narrower bumps, and additive Gaussian white noise. Center: Result of smoothing the data by a simple moving window average. The window extends 16 points leftward and rightward, for a total of 33 points. Note that narrow features are broadened and suffer corresponding loss of amplitude. The dotted curve is the underlying function used to generate the synthetic data. Bottom: Result of smoothing the data by a Savitzky-Golay smoothing filter (of degree 4) using the same 33 points. While there is less smoothing of the broadest feature, narrower features have their heights and widths preserved.

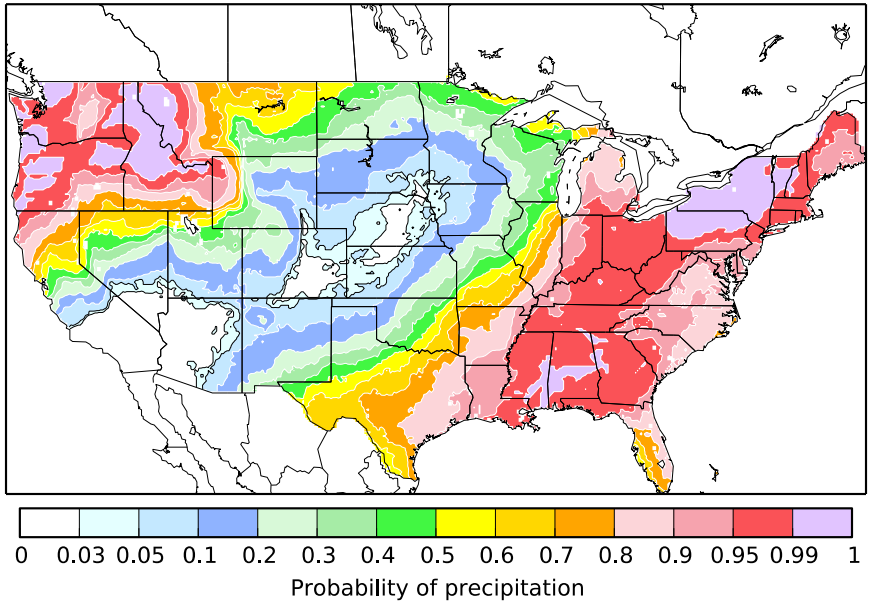
- designed to preserve amplitudes, unlike more common nxn block smoothers

from Press et al. Numerical Recipes text

(a) Downscaled POP 108-120 forecast,
initial time = 2014010700

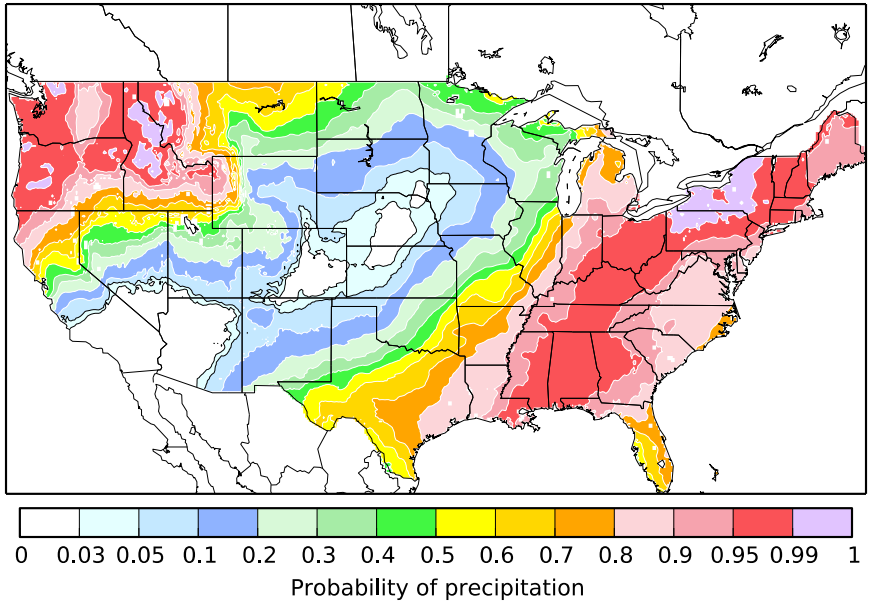


(b) Raw multi-model ensemble POP 108-120 forecast,
initial time = 2014010700

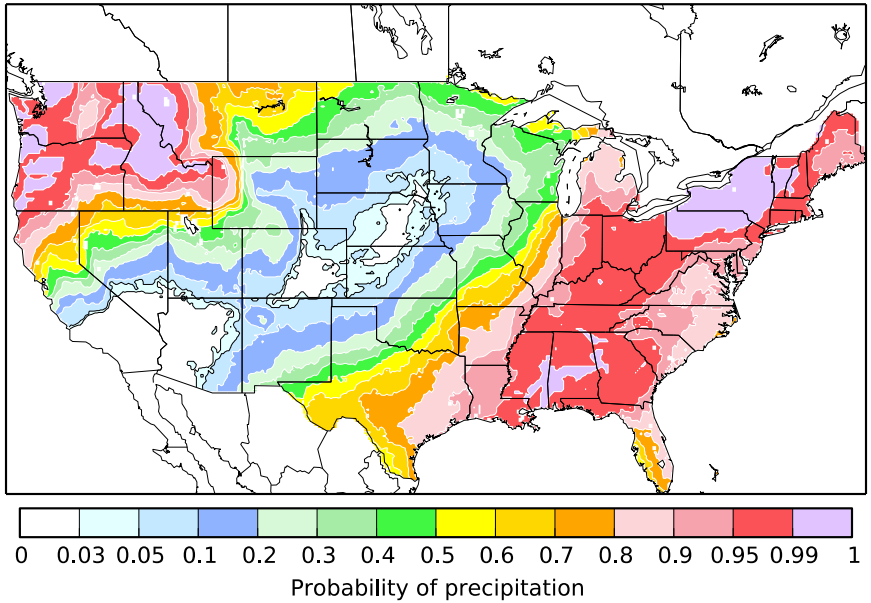


Multi-model POP with statistical downscaling (but w/o Savitzky-Golay)

(a) Downscaled POP 108-120 forecast,
initial time = 2014010700



(b) Raw multi-model ensemble POP 108-120 forecast,
initial time = 2014010700



Multi-model POP with statistical downscaling (but WITH Savitzky-Golay)

In this implementation, more smoothing is applied in less varied terrain of eastern US than in more varied terrain of western US.

Strategy for POP forecasting implemented here.

- POPs via multi-model ensemble, but with two modifications:
  - Adjust each member with a CDF-based bias correction.
  - Apply statistical downscaling as described previously.
- Data used in subsequent demo:
  - 108-120 h coarse res. precip. forecasts, ensembles and deterministic for Jan 2014 from ECMWF, NCEP GEFS, UK Met, CMC.
  - CDF bias correction trained using Nov-Dec 2013 108-120 h forecasts
  - Long time series of both hi-res. and coarse-res. CCPA precip. analyses (2002-2013) for statistical downscaling.
- Technique:

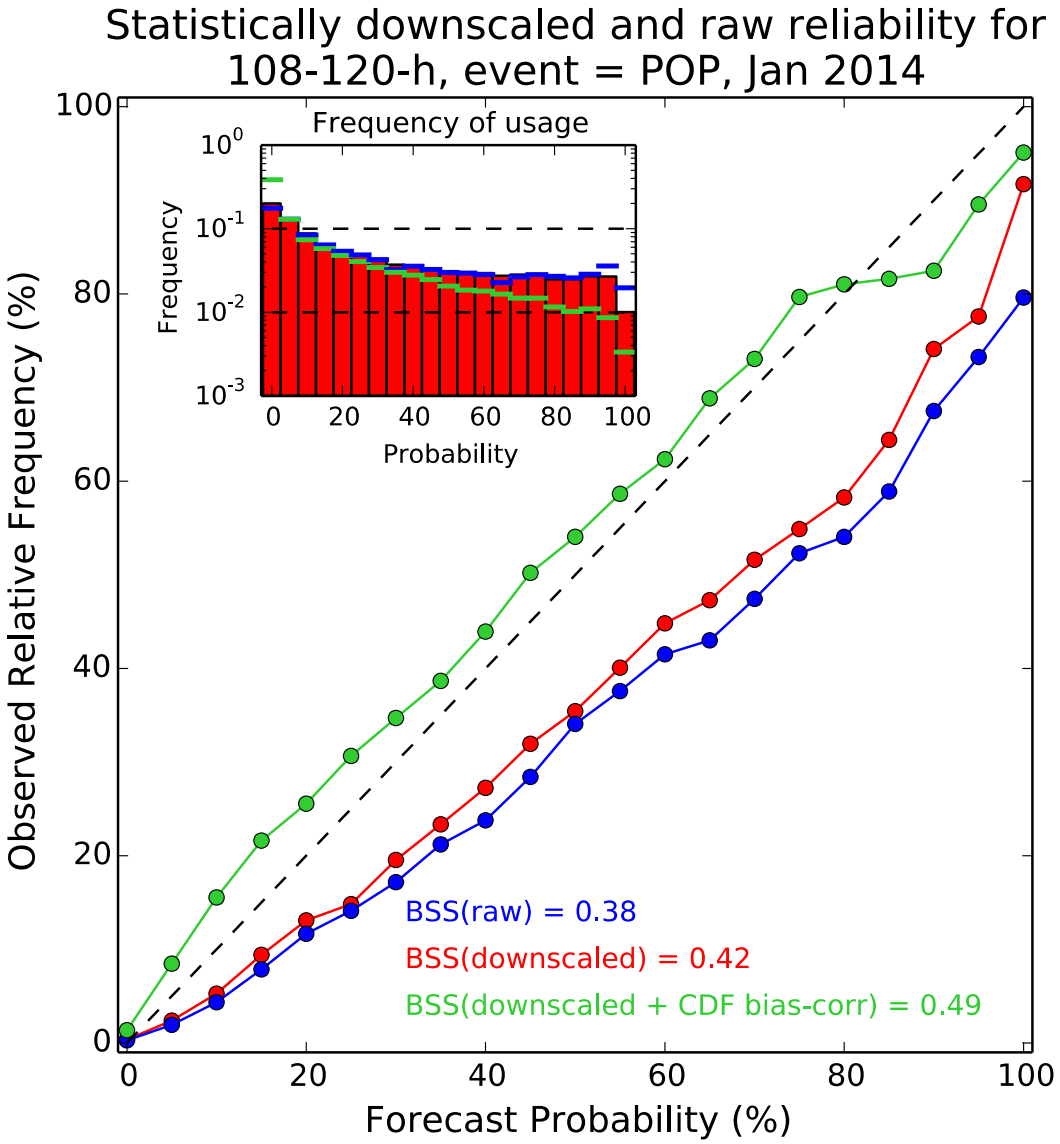
  (1) CDF bias correct the forecasts using the past 60 days of forecasts and observations.

  (2) Statistically downscale each member.

  (3) Compute probability from ensemble relative frequency.

  (4) Apply Savitzky-Golay smoothing to mimimize noise from sampling error.

POP Reliability and skill, before & after

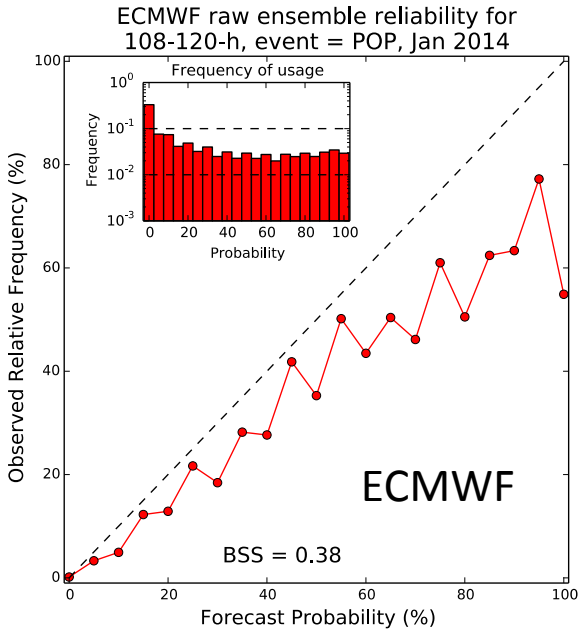


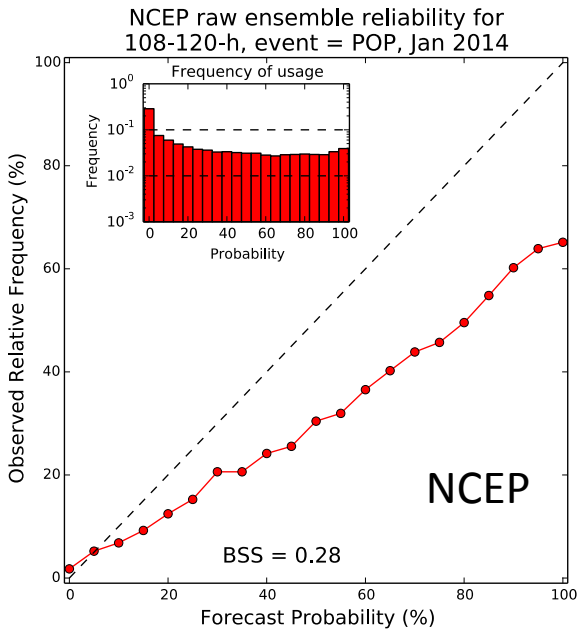
The BSS numbers look rather high, even for the raw ensembles (next slide). I think this is because:

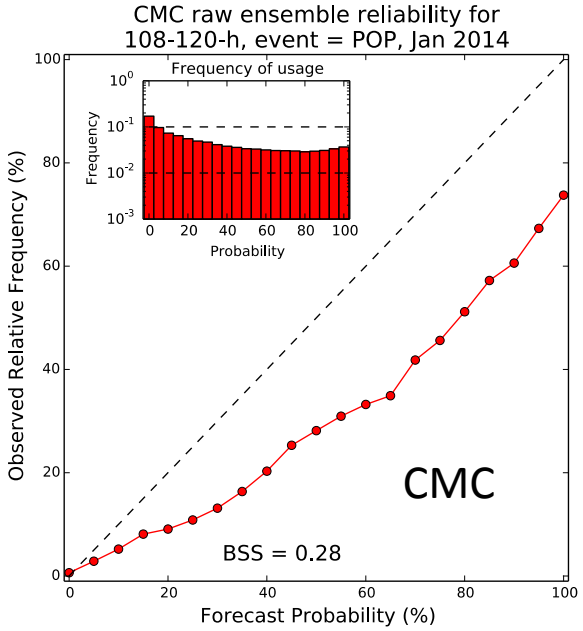
(a) no big "dropouts" this month?

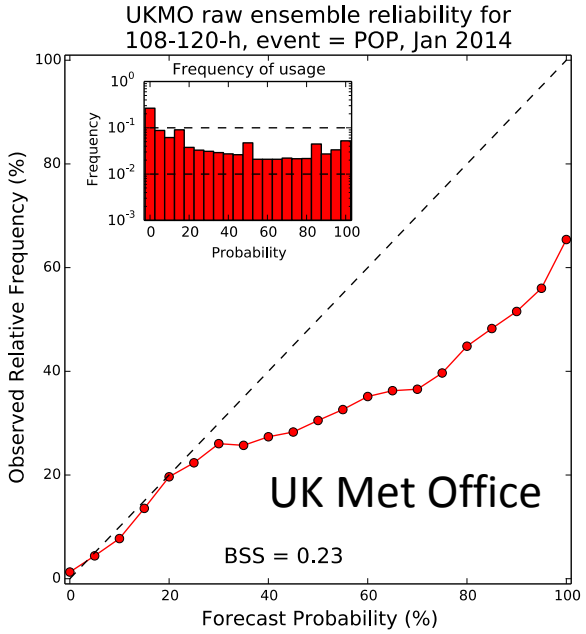
(b) forecasts were very different than the long-term climatology, so skill relative to climatology was larger (the month was abnormally dry in many locations).

Reliability and skill of raw ensembles





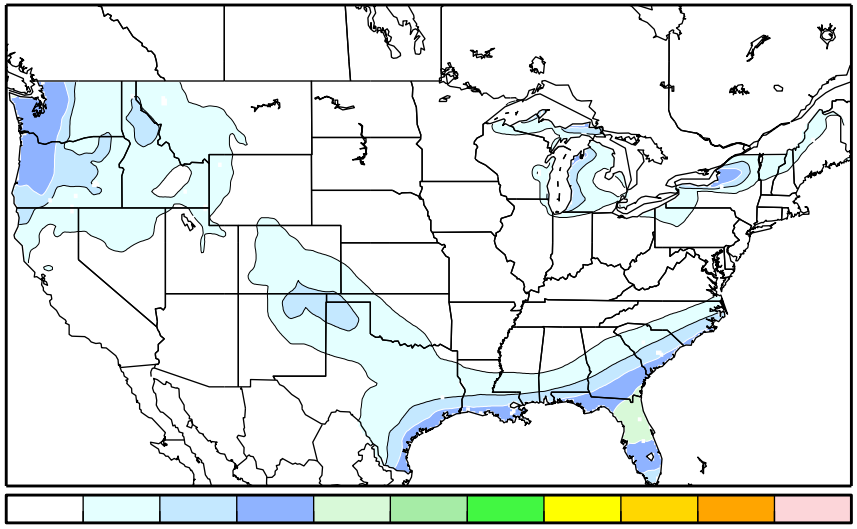




Over this sample, the multi-model raw ensemble skill wasn't better than that of the best forecast, from ECMWF.

Case study

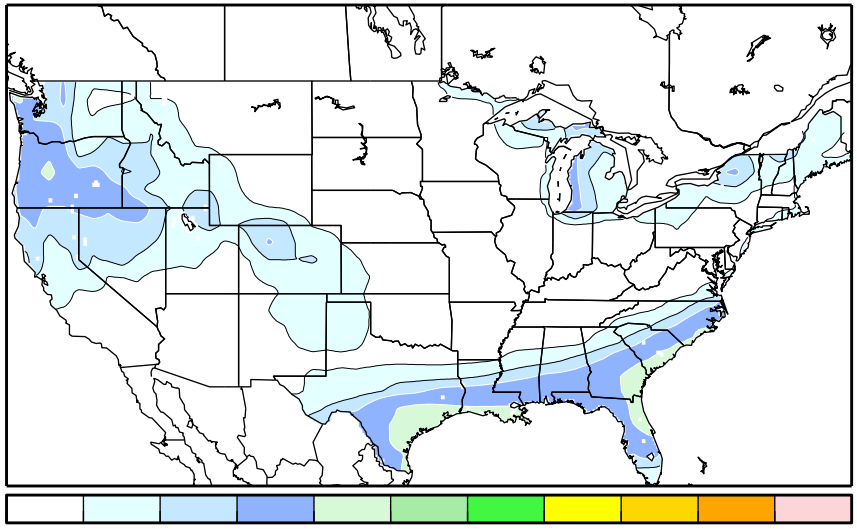
(a) ECMWF ensemble mean 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

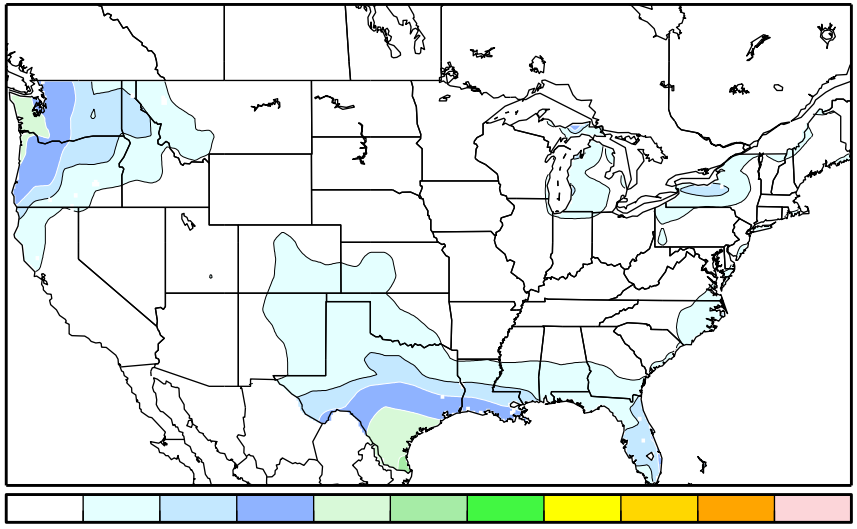
(b) NCEP ensemble mean 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

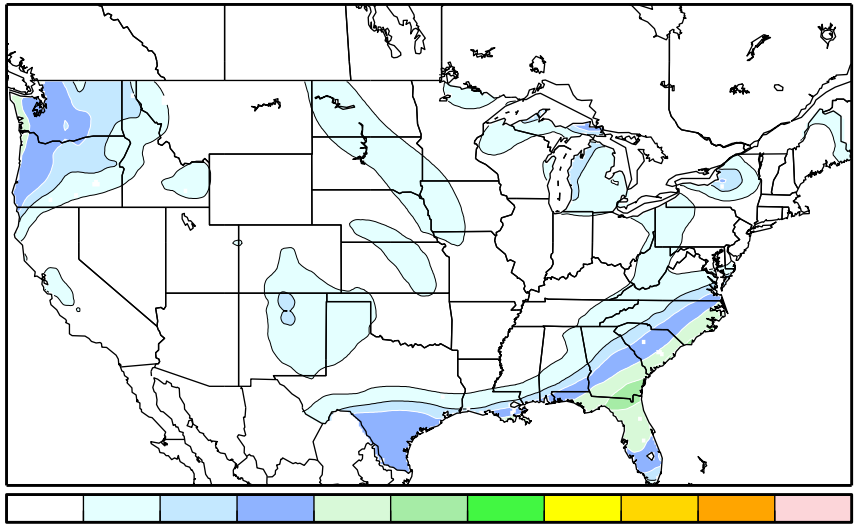
(c) CMC ensemble mean 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

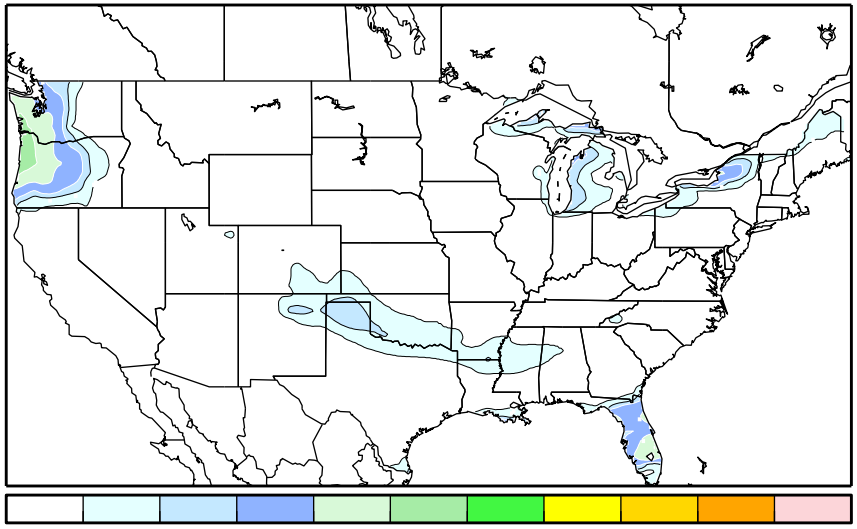
(d) UKMO ensemble mean 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

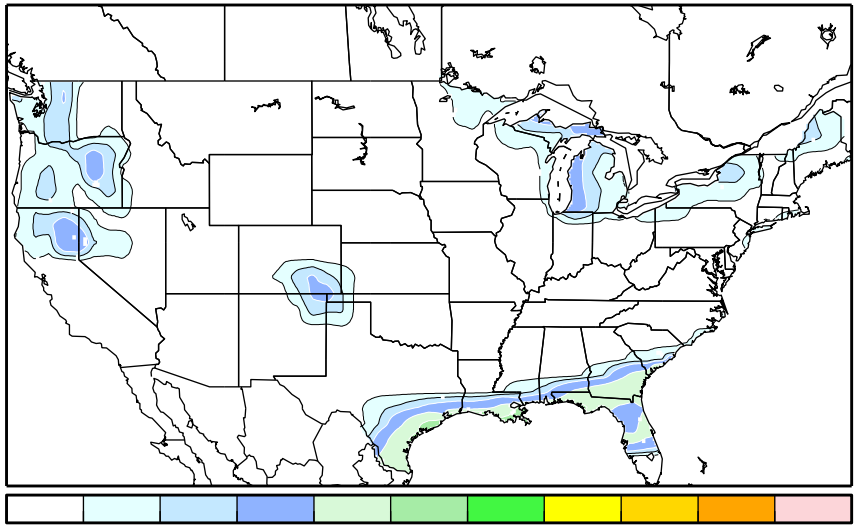
(a) ECMWF deterministic 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

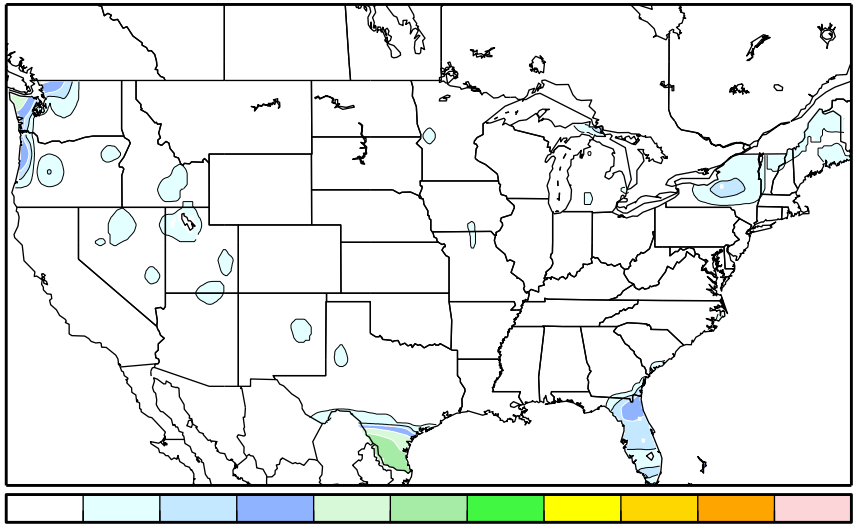
(b) NCEP control 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

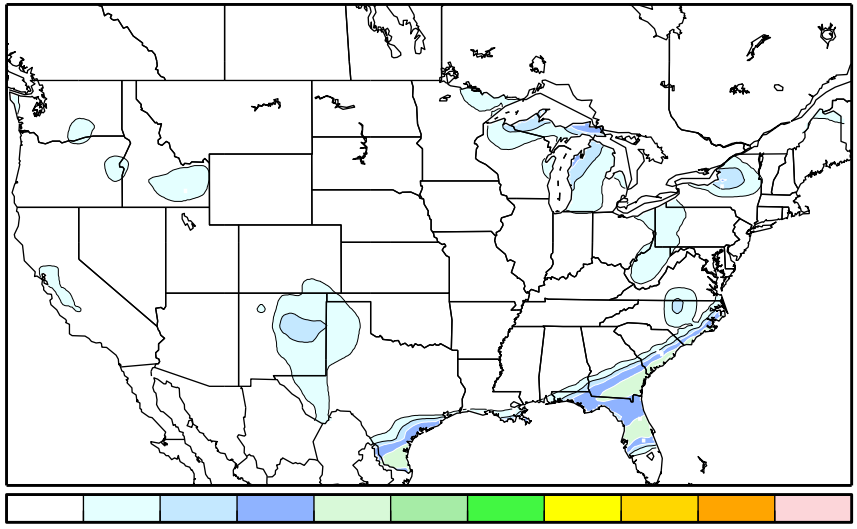
(c) CMC control 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

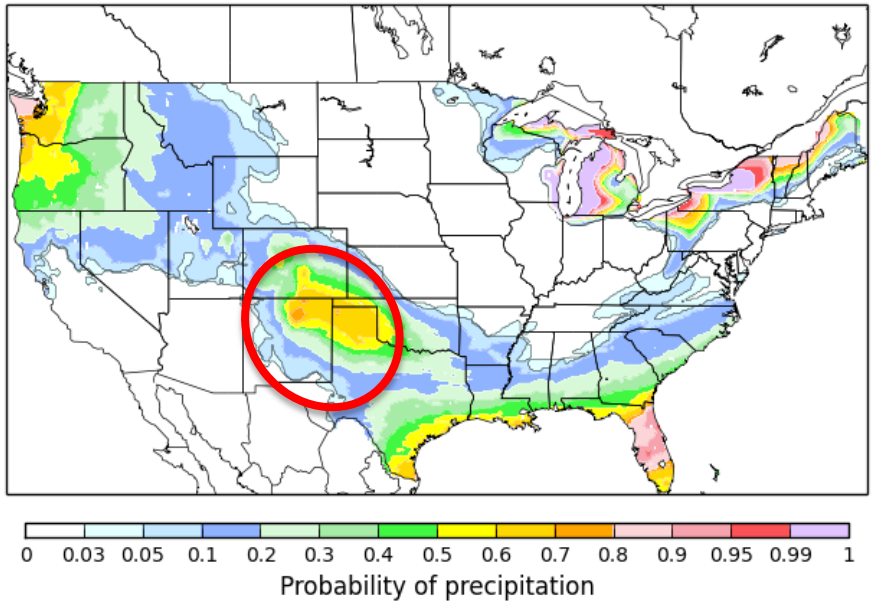
(d) UKMO control 108-120 forecast,
initial time = 2014012400



0 0.1 0.5 1 2.5 5 10 25 50 100 150 200

Precipitation amount (mm)

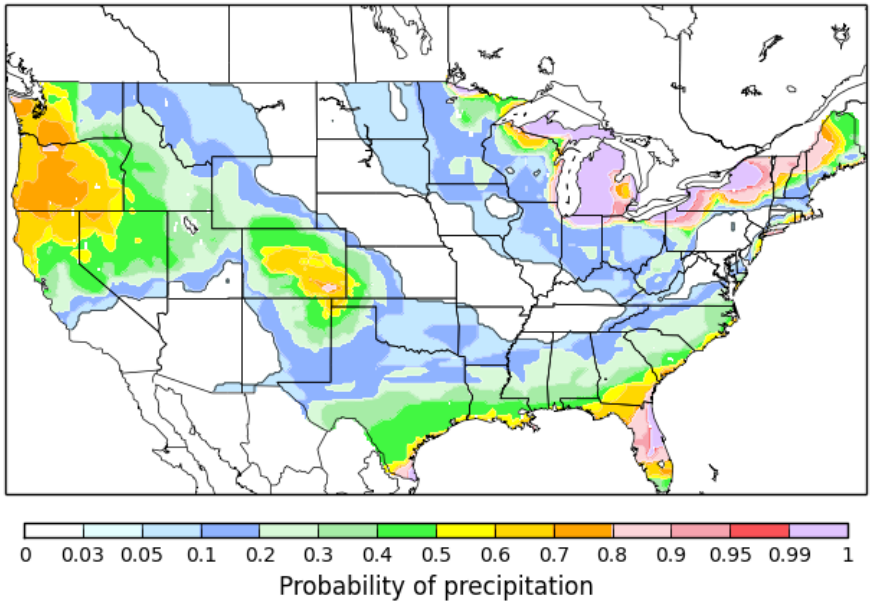
(a) Raw ECMWF ensemble POP, 108-120 forecast, initial time = 2014012400



0 0.03 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95 0.99 1

Probability of precipitation

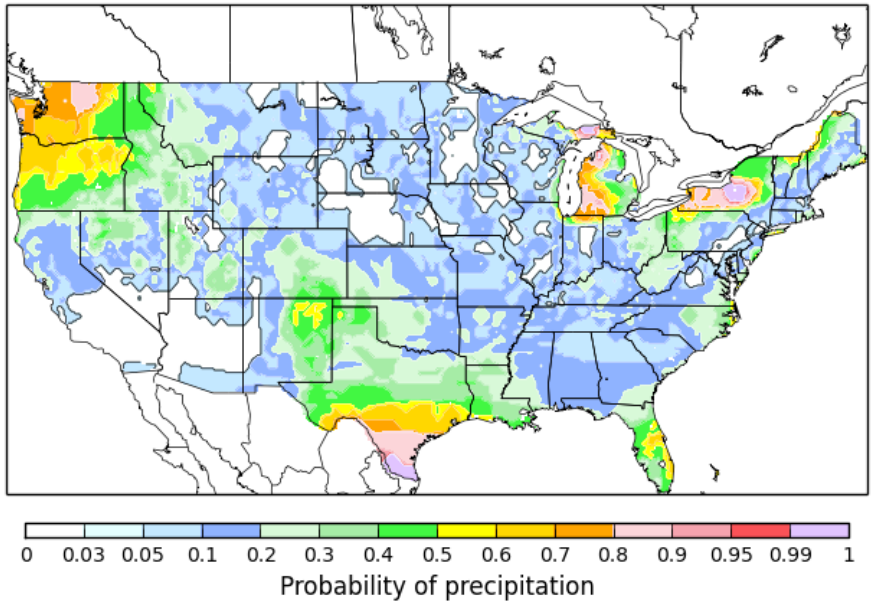
(b) Raw NCEP ensemble POP, 108-120 forecast, initial time = 2014012400



0 0.03 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95 0.99 1

Probability of precipitation

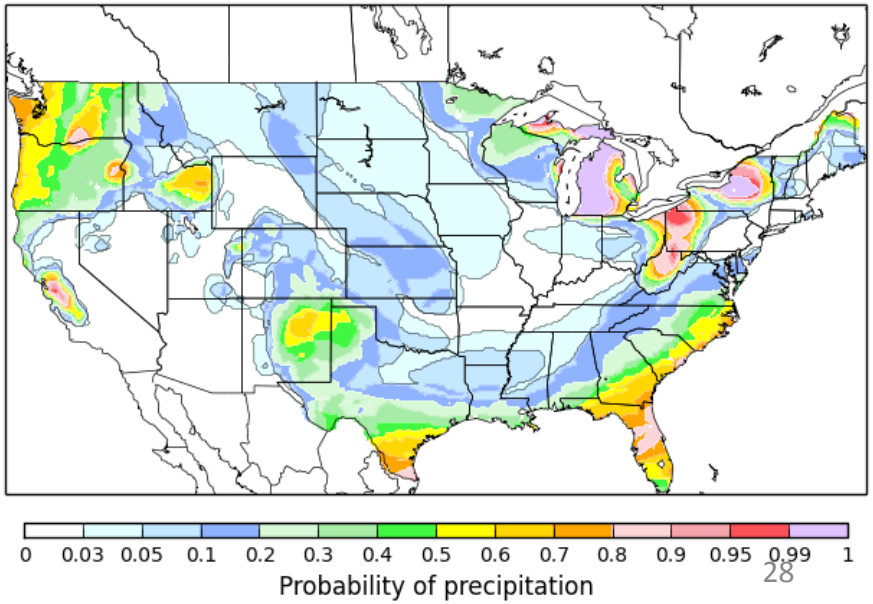
(c) Raw CMC ensemble POP, 108-120 forecast, initial time = 2014012400



0 0.03 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95 0.99 1

Probability of precipitation

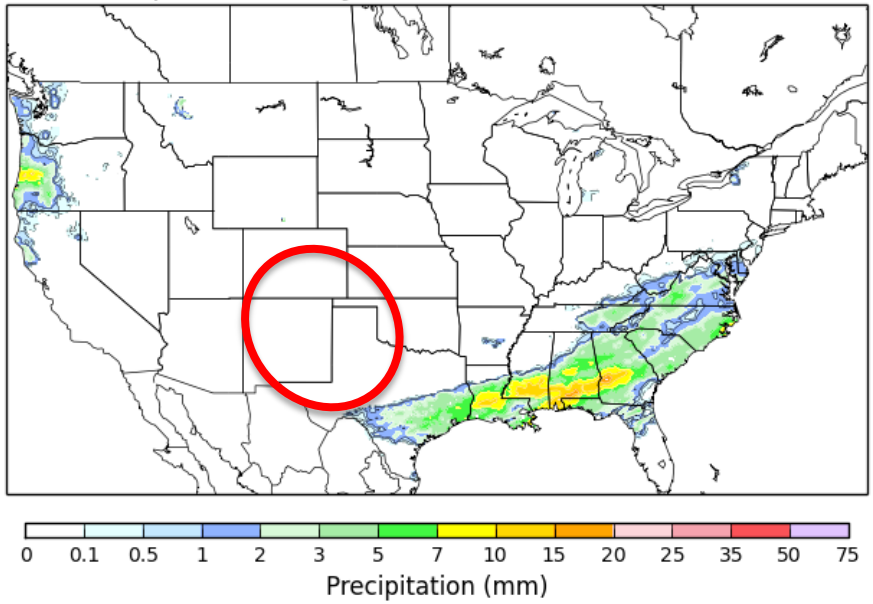
(d) Raw UKMO ensemble POP, 108-120 forecast, initial time = 2014012400



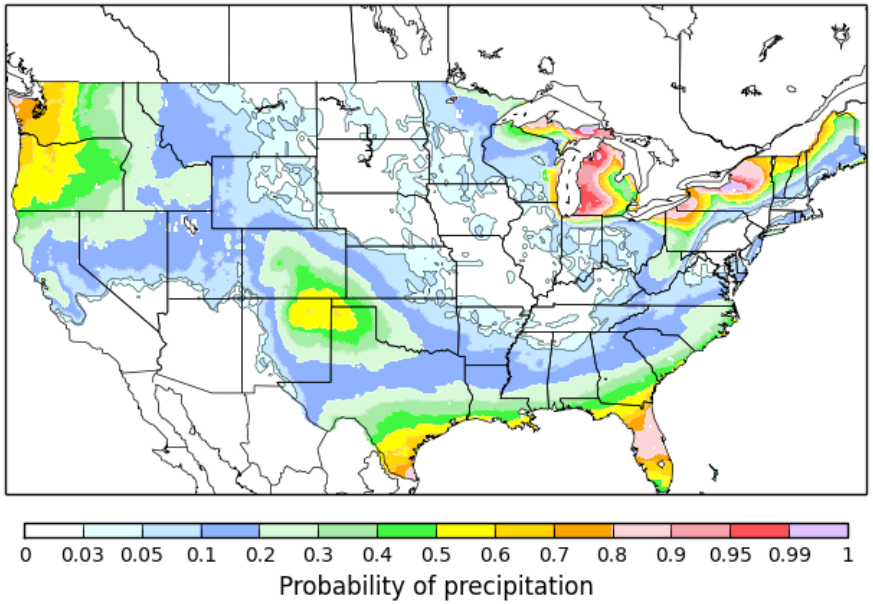
0 0.03 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95 0.99 1

Probability of precipitation

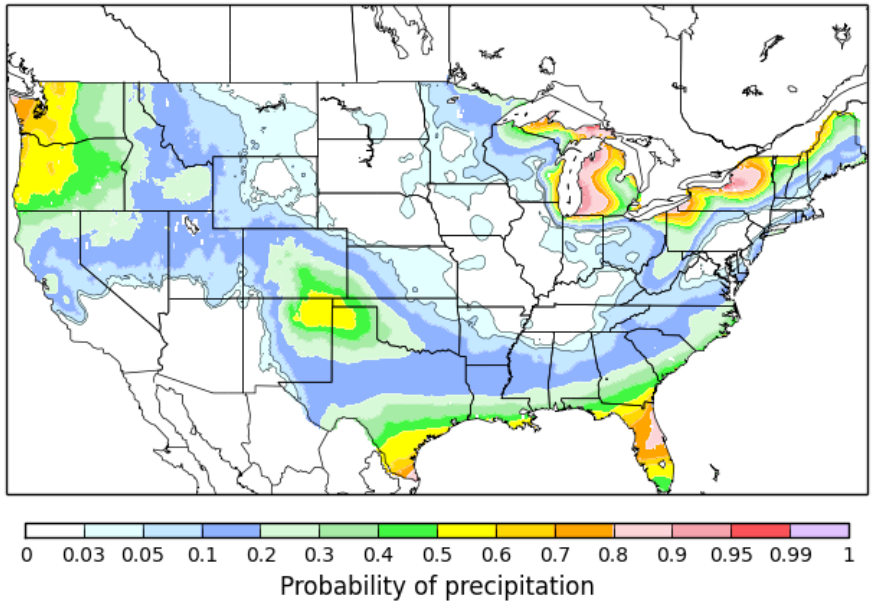
(a) Precipitation analysis, 2014012812 to 2014012900



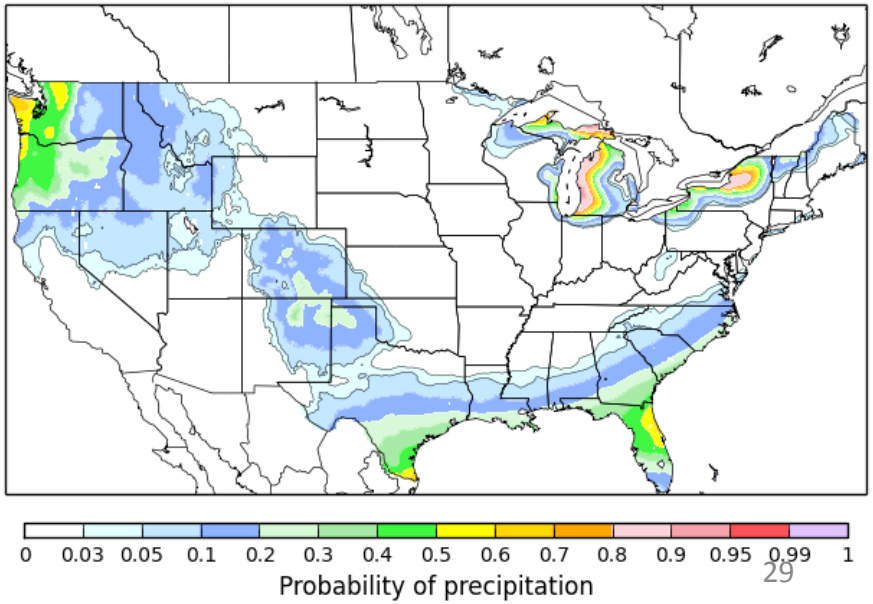
(b) Raw multi-model ensemble POP 108-120 forecast, initial time = 2014012400



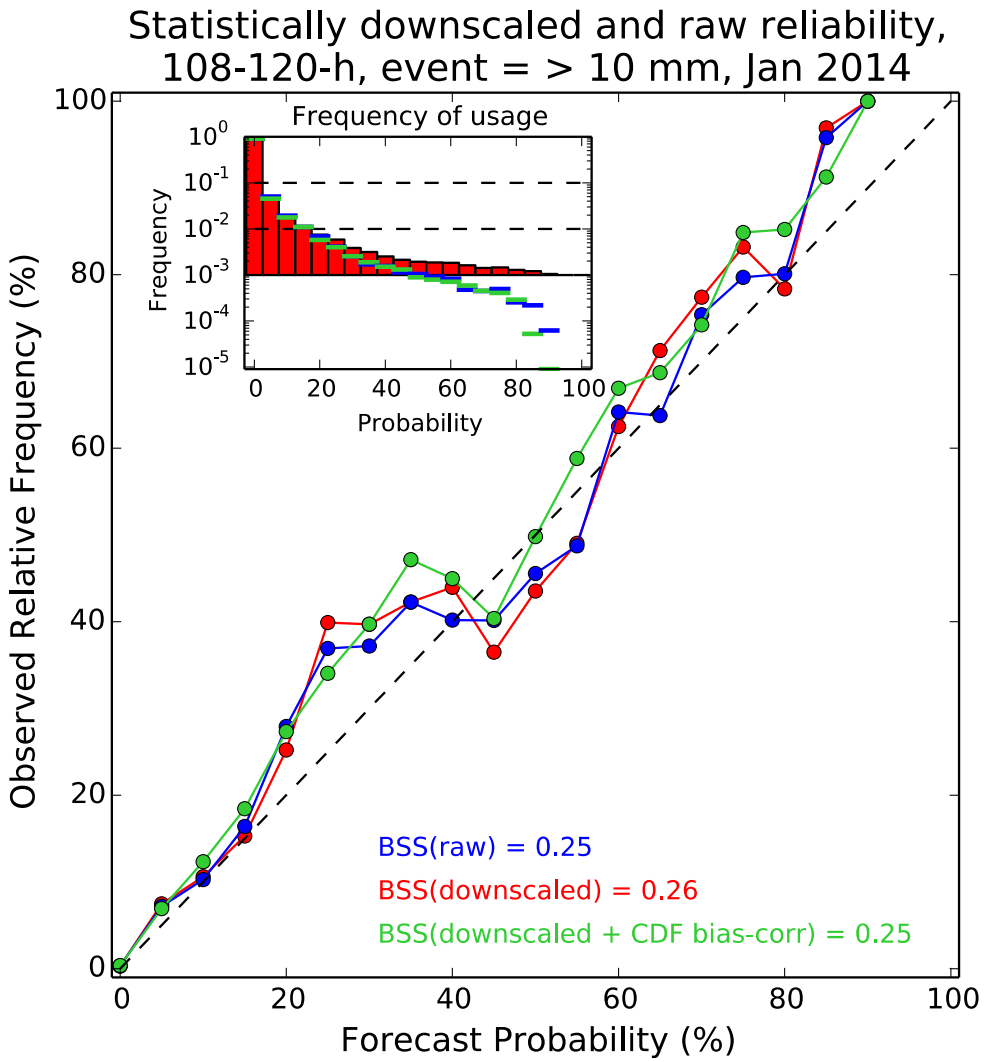
(c) Downscaled POP 108-120 forecast, initial time = 2014012400



(d) Downscaled POP & CDF bias corr.,108-120 forecast, initial time = 2014012400



Benefits of CDF correction may not generalize to higher precipitation amounts



Applying the same methodology to > 10 mm 12 h^{-1} forecasts, we see little improvement as a result of statistical downscaling or CDF-based bias corrections.

It's likely that 60 days, even with supplemental locations, is not enough samples to properly adjust > 10 mm forecasts (reforecasts and/or other post-processing approach needed)

Thinking ahead: How to improve PQPF for rare events

- Next-Generation Global Prediction System priorities include:
  - Development of reanalysis/reforecast data sets to provide large training samples needed to improve probabilistic forecasts of rare events like heavy precipitation.
  - Development and deployment of advanced post-processing methods.

Where are we with these developments?

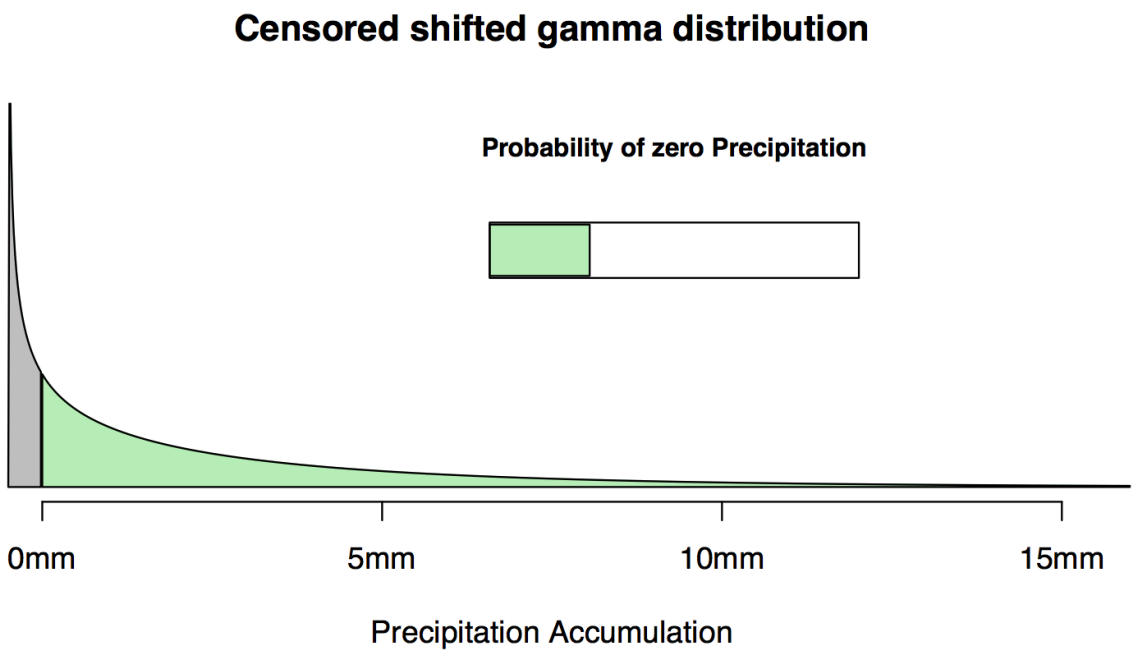
Reanalysis / reforecast: where are we?

- What's important is to have reforecasts be statistically consistent with real-time forecasts.
- To do this, probably need reanalysis using same model, same data assimilation procedure as used for real-time GFS analysis.
- Existing reanalysis is now ~ 7 years old, T382 3D-Var; new analyses soon to be T1534 4D-En-Var. So we need new reanalysis.
- ESRL/PSD funded to begin reanalysis work, expect funding to continue it in FY2016 with EMC, CPC.
- Reforecast generation is straightforward (if computationally expensive) once reanalysis is generated.

Advanced post-processing methods: where are we?

- A distribution family for precipitation

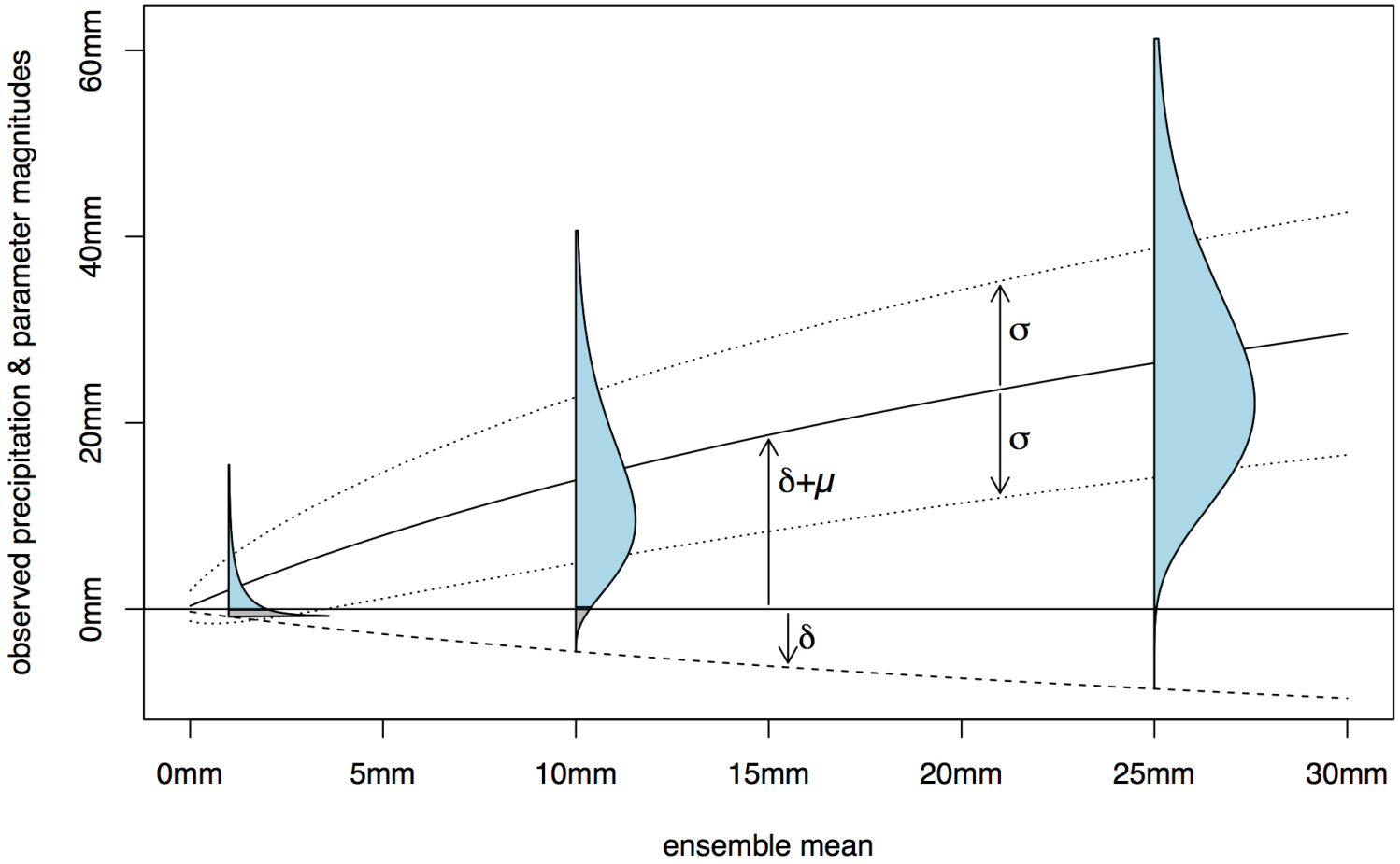
We model precipitation accumulations by censored, shifted gamma distributions (CSGDs):



Scheuerer and Hamill, 2015 MWR, conditionally accepted.

Gamma distribution's shift, location, and scale parameters set with ensemble data

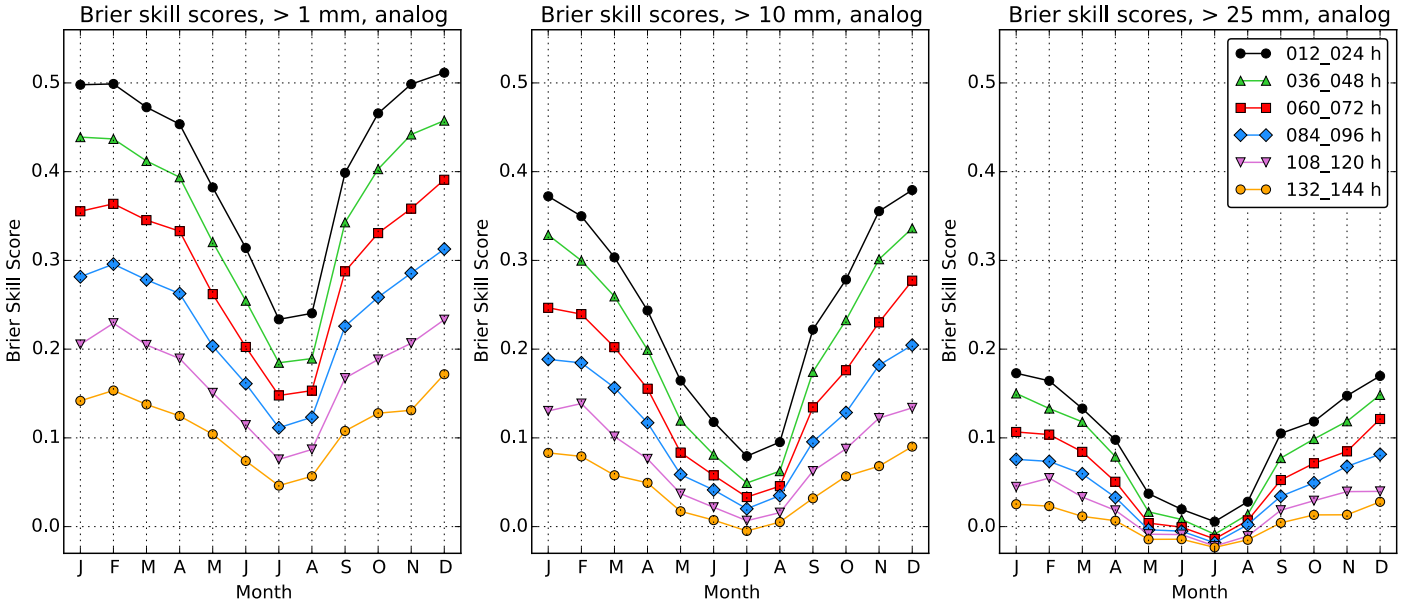
This produces predictive distributions with a variety of shapes depending on the ensemble mean, spread, and other predictors



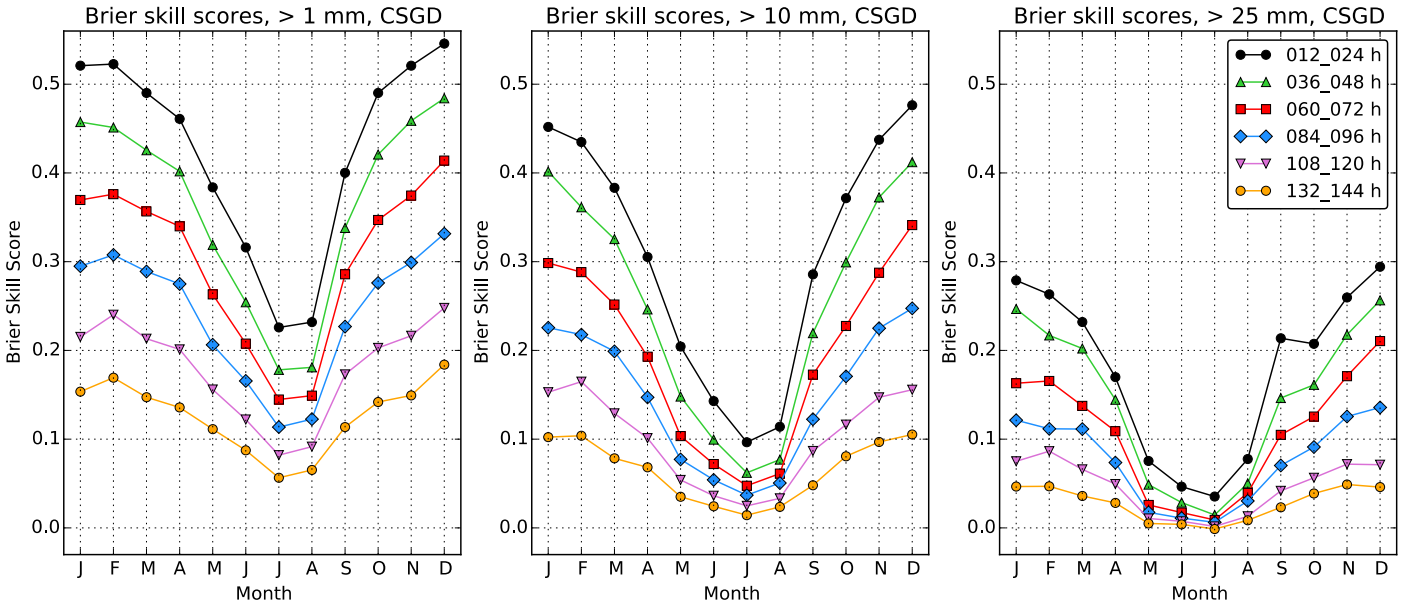
Ref: ibid.

Brier Skill Scores

Rank analog (no supplemental locations)



Censored, shifted gamma distributions

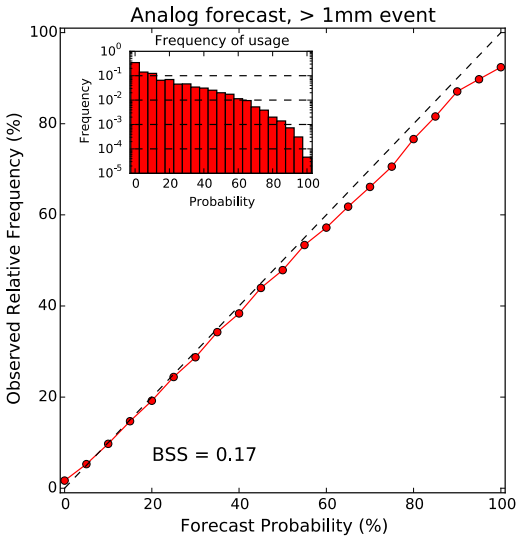


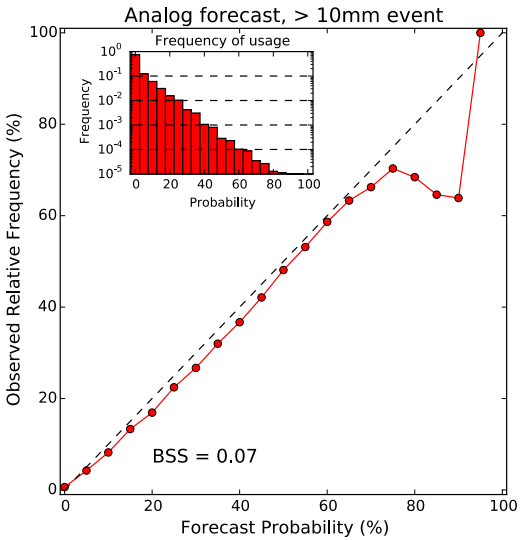
At the higher thresholds, the CSGD forecasts significantly outperform the rank analog.

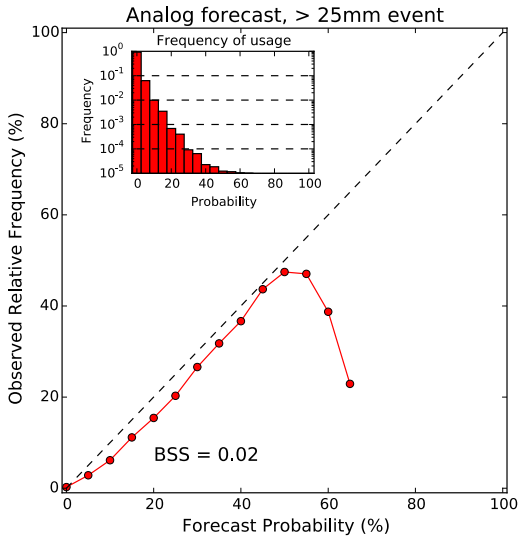
No data from supplemental training locations used in either method here.

Reliability diagrams, +108 to +120 h

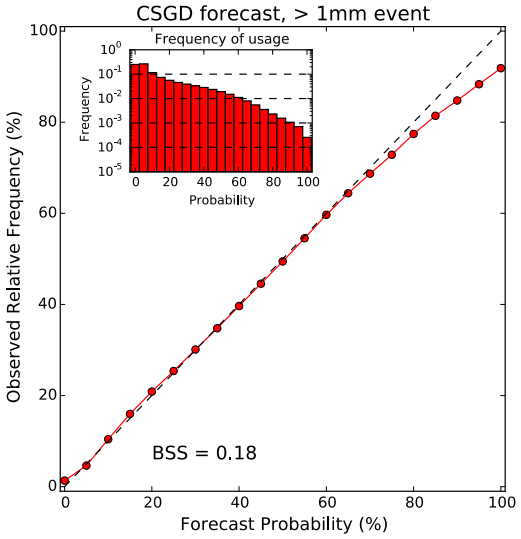
Rank analog, no supplemental locations

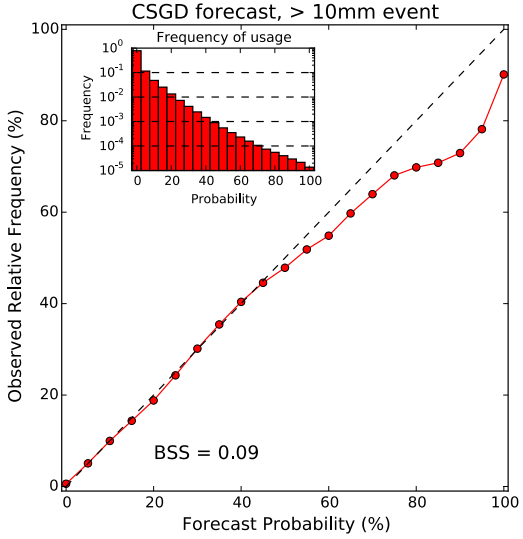


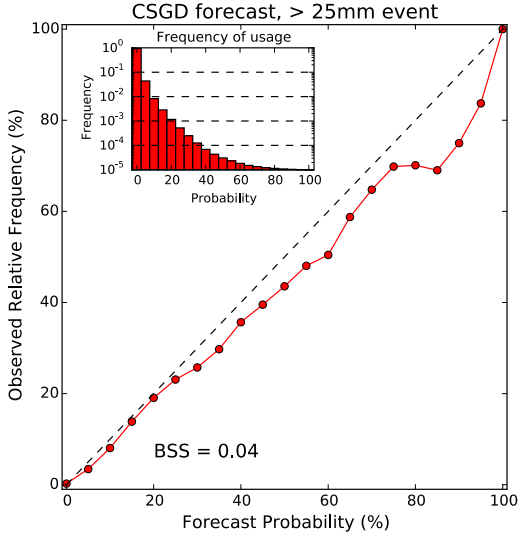




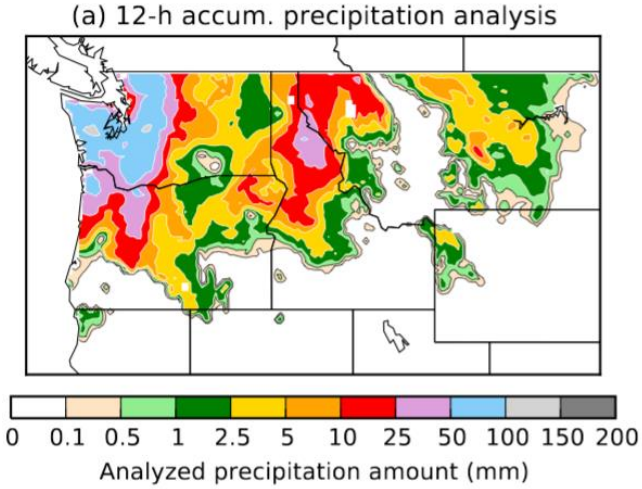
Censored, shifted gamma distribution fitting

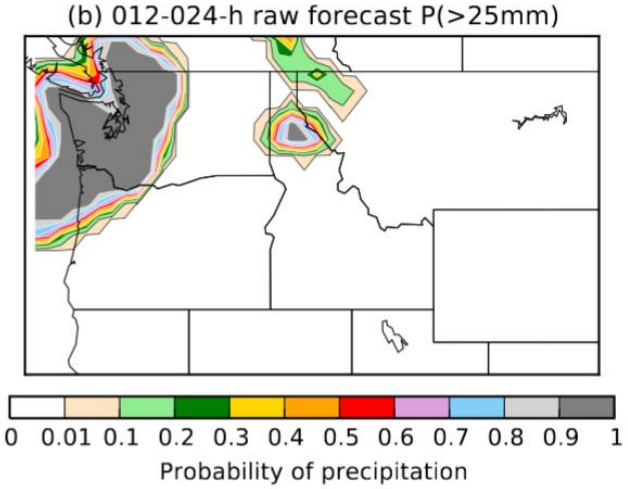


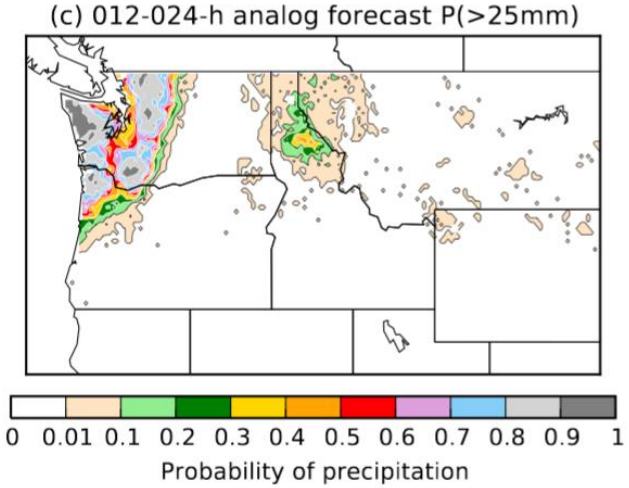


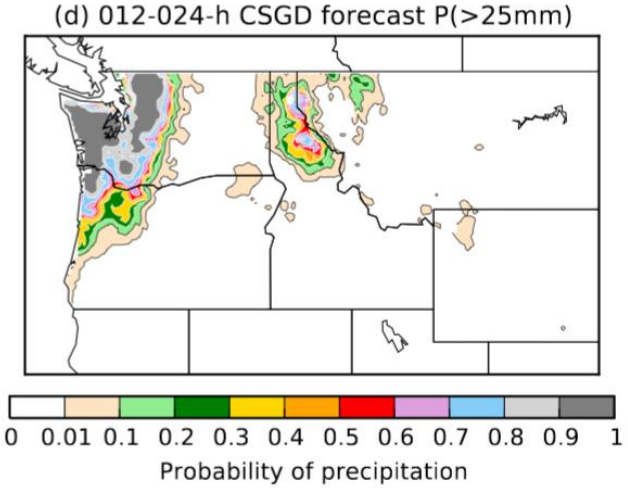


Improved high-end sharpness



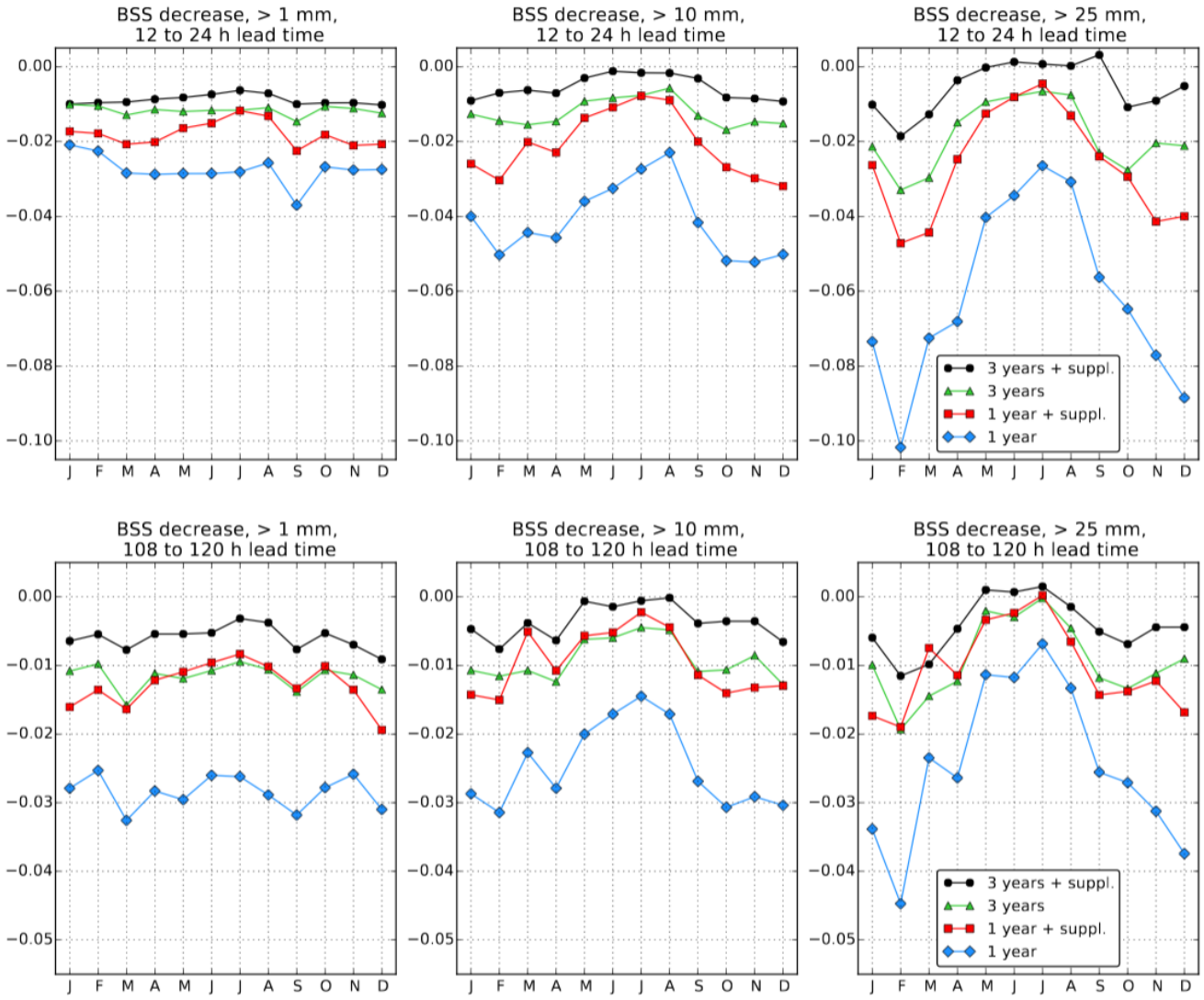






In this case (one where the raw ensemble guidance was quite good), one can see that the CGSD approach produced much higher probabilities of heavy precipitation than the analog.

How skillful is CSGD when using smaller training sample sizes?



There is some drop-off in skill, shown here, and reliability (not shown) when decreasing samples to 3- or 1 year of training data from the original 2002-2014 training data. Use of supplemental locations to increase the sample size does have a positive benefit.

Further tests and comparisons underway.

Conclusion / next steps

- It appears to be possible to add skill, improve reliability of multi-model ensemble POP with statistical downscaling,CDF quantile mapping, and judicious smoothing.
- Improvements to probabilities for higher event thresholds more marginal with this method, probably because CDFs used in bias correction are noisy with small sample size at higher quantiles.
- Our group is working on more advanced methods, including censored, shifted Gamma distributions, which show promise for post-processing an individual model.
- This method is being evaluated in real time by forecasters.
  - Pending evaluation, will migrate this technology to downscale using high-resolution Stage-IV precipitation analyses.

Supplementary slides

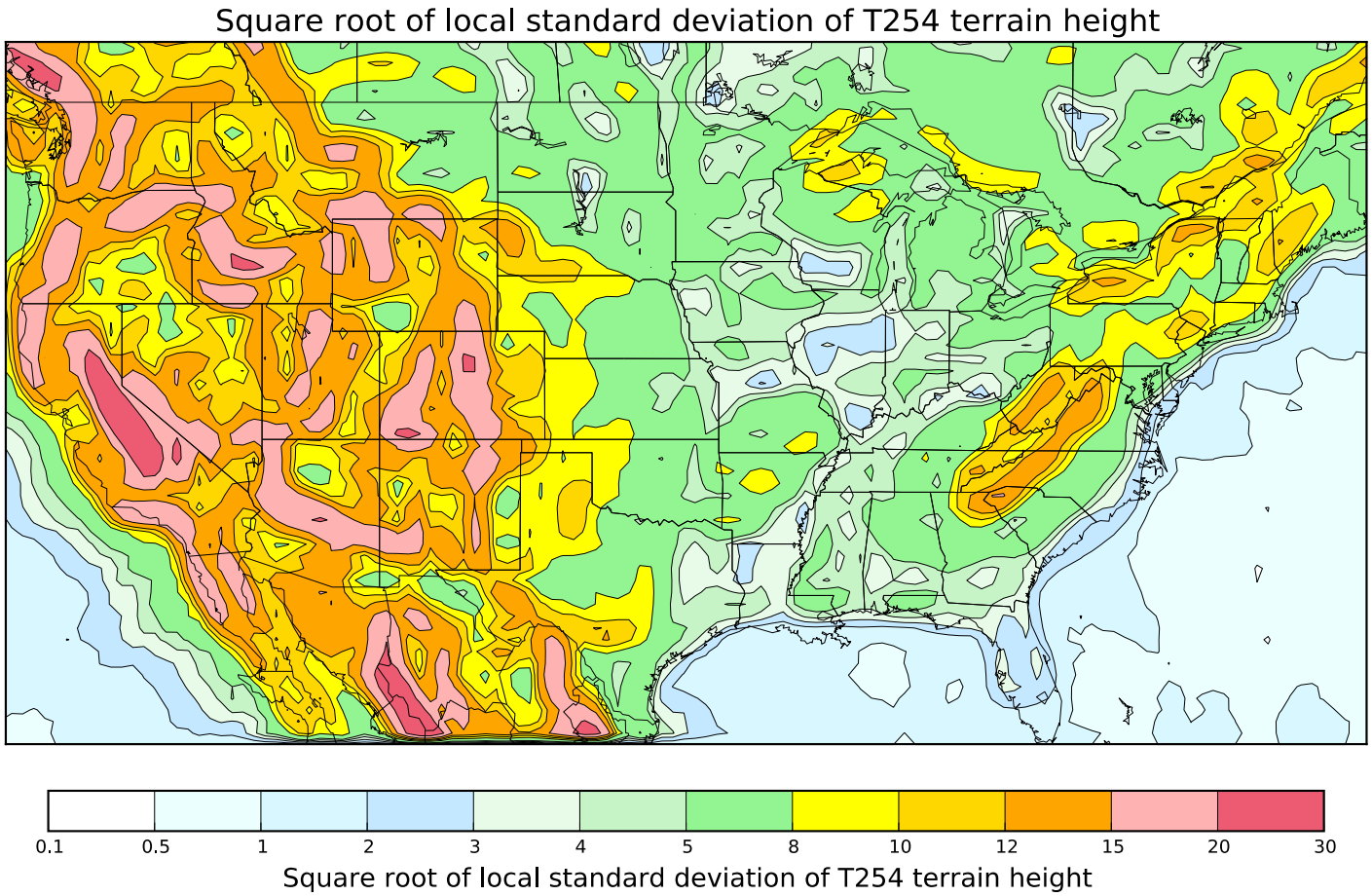
Methodology for weighting smoothed vs. original ensemble probabilities.

At each ~ ½-degree forecast grid point, the local mean terrain height was determined for a 3x3 grid-point box centered on the box of interest. The standard deviation $\sigma_{i,j}^{T}$ for that 3x3 box was then calculated with respect to this mean value. A plot of the standard deviations are shown in Fig. A4. Finally, the weight $w_{i,j}$ to apply to the raw forecast was calculated as:

$$w_{i,j} = \begin{cases} 0.2 & if\ {\sigma_{i,j}^{T}}^{1/2} < 8 \\ 0.2 + \dfrac{\left({\sigma_{i,j}^{T}}^{1/2} - 8\right)}{16.66667} & if\ 8 \le {\sigma_{i,j}^{T}}^{1/2} < 108 \\ 0.8 & if\ 108 \le {\sigma_{i,j}^{T}}^{1/2} \end{cases}, \quad \text{(A5)}$$

and the weight applied to the smooth forecast was 1- $w_{i,j}$. A map of the weight applied to the raw forecast is shown in Fig. A5.

Terrain-height variations



Weight applied adjusted ensemble probabilities

Weight applied to CDF bias-corrected and downscaled multi-model ensemble probabilities