Engineering & Applied Science
UNIVERSITY OF COLORADO **BOULDER**

Beyond Univariate Calibration

Verification of Spatial Structure in Ensembles of Forecast Fields

Joshuah Jacobson
Department of Applied Mathematics

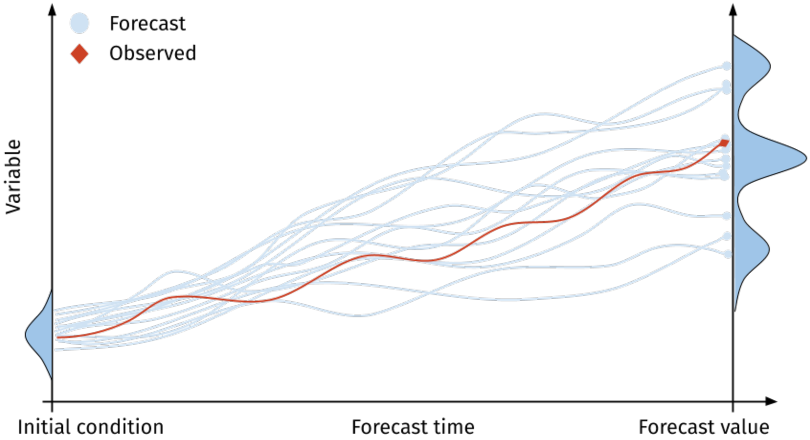
Advisor: William Kleiber
Co-Advisor: Michael Scheuerer

March 20, 2020

Contents

Ensemble forecast verification

Instead of running just a single forecast, the model is run a number of times from slightly different starting conditions to **sample uncertainty** of future conditions.



Univariate calibration: rank histograms

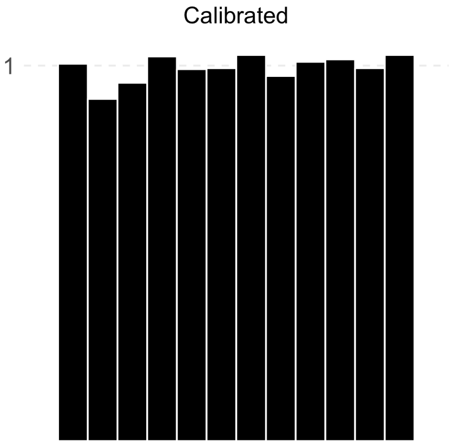
Forecast is calibrated if the truth and the ensemble can be considered samples from the **same probability distribution** (Hamill 2001).

Ensemble: $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$

Observation: V

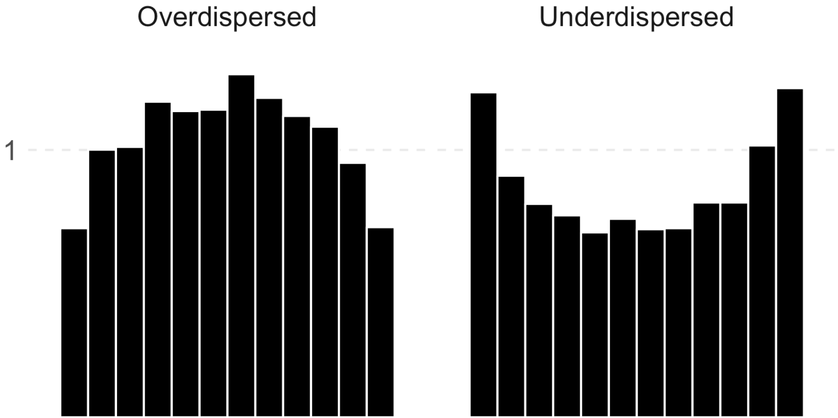
Uniform density:

$$P(X_{i-1} \leq V < X_i) = \frac{1}{n+1}$$

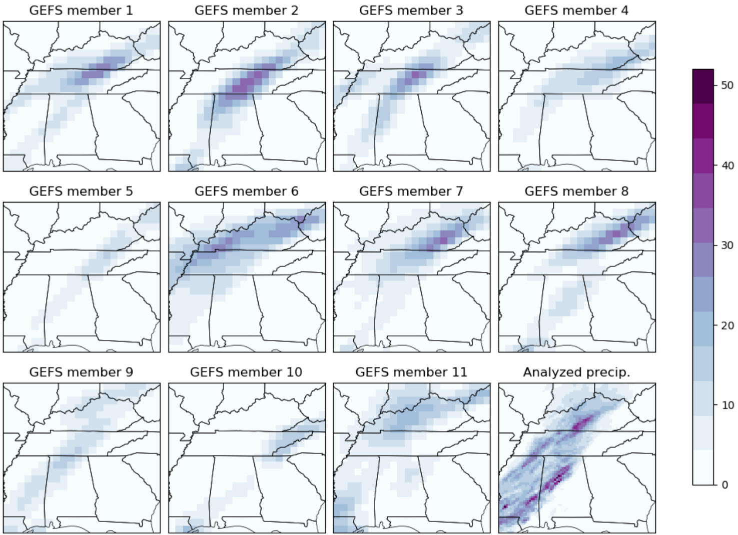


Univariate calibration: signs of miscalibration

Forecast is calibrated if the truth and the ensemble can be considered samples from the **same probability distribution**.

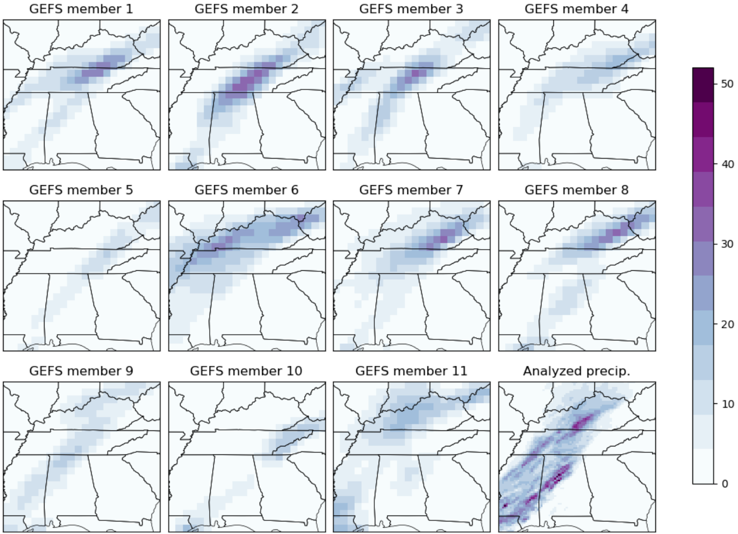


Ensemble forecast fields



If forecasts are studied at each location separately, we can use univariate diagnostic tools like verification rank histograms (Anderson 1996; Hamill 2001).

Ensemble forecast fields

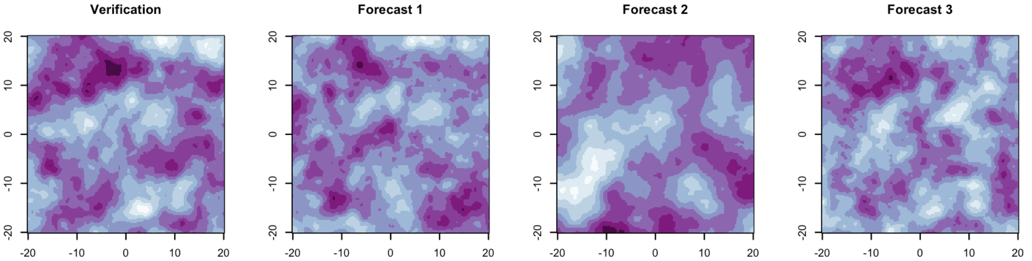


For weather variables like precipitation, it is important that the forecast amounts accumulate correctly over the domain.

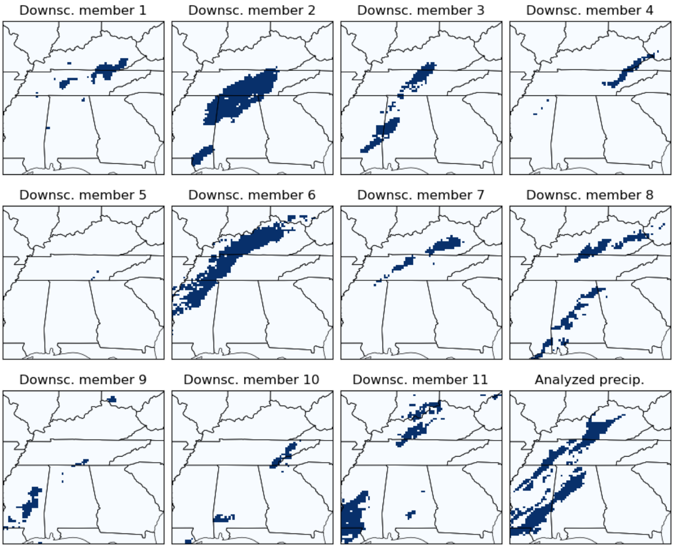
▷ **“correlation length”**

Miscalibration is not always obvious

One of these forecasts has the correct spatial **correlation length**, one is 10% miscalibrated, and one is 50% miscalibrated. Can you tell which is correct?



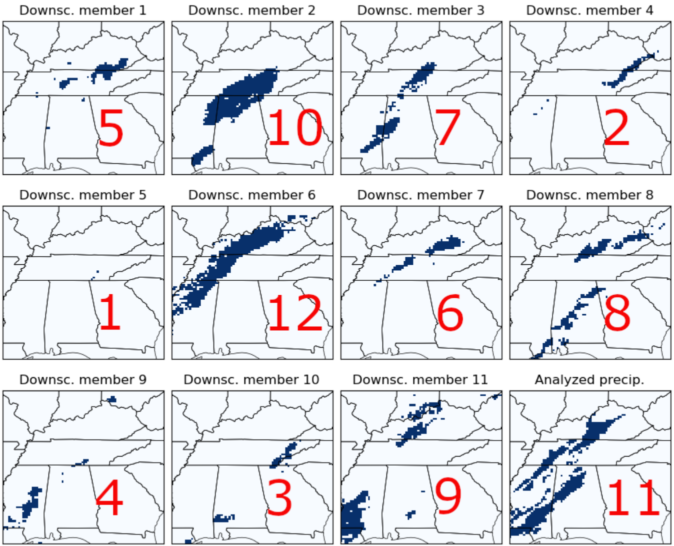
Fraction of threshold exceedance (FTE)



$$\mathrm{FTE}(Z,\tau) = \frac{1}{|D|}\int_D \mathbf{1}_{\{Z(s)>\tau\}}(s)ds$$
$$= \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{Z(s)>\tau\}}(s_j)$$

The FTE is a projection of a multivariate quantity to a univariate quantity that can be evaluated by common univariate verification metrics.

Fraction of threshold exceedance (FTE)

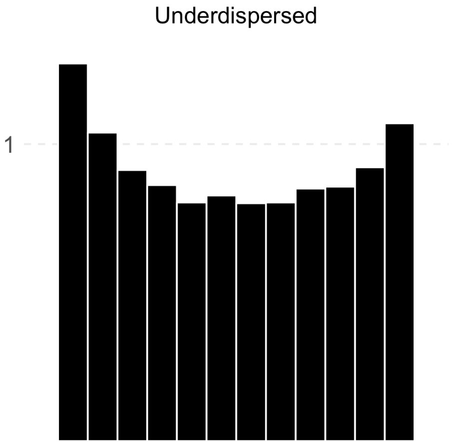


$$\mathrm{FTE}(Z,\tau) = \frac{1}{|D|}\int_D \mathbf{1}_{\{Z(s)>\tau\}}(s)ds$$
$$= \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{Z(s)>\tau\}}(s_j)$$

For reliable ensemble forecasts, the distribution of the analysis FTE should be **interchangeable** with the ensemble FTEs.

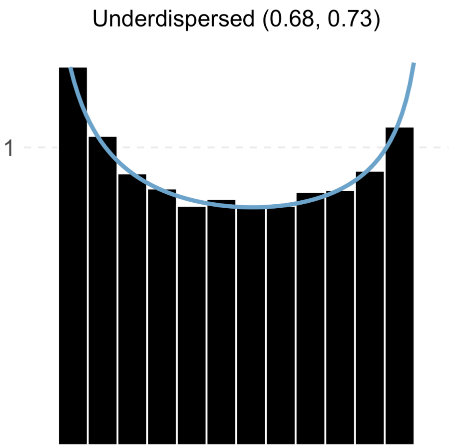
FTE histogram

- Use a verification rank histogram as a univariate metric.
- Repeatedly tally the rank of the analysis FTE among the set of the analysis and ensemble forecast FTEs, and study the histogram of these ranks.
- Interpretation is similar to standard rank histograms.



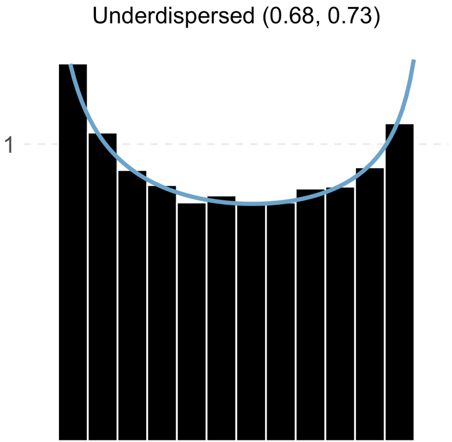
Summary statistics

- In cases where uniformity of the histogram is not obvious, it can be useful to have an objective summary statistic.
- Fitting a **beta distribution** to standardized, disaggregated rank data by maximum likelihood estimation yields a set of summary statistics in the form of distribution parameters.



Summary interpretation

Parameter Relationship	Histogram Interpretation
$a = b = 1$	Uniform
$a, b < 1$	$\cup$–shaped
$a, b > 1$	$\cap$–shaped
$a < b$	Right-skewed
$a > b$	Left-skewed



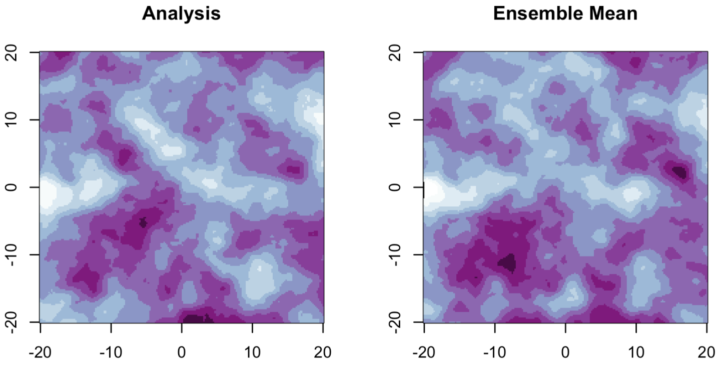
Research question:

Does the FTE accurately identify miscalibration of ensemble correlation length?

Simulation algorithm

We generate synthetic exceedance fields as follows:

1. Simulate verification Z_0 and ensemble mean Z_m as cross-correlated Gaussian random fields according to the bivariate Whittle-Matérn model (Gneiting, Kleiber, and Schlather 2010)



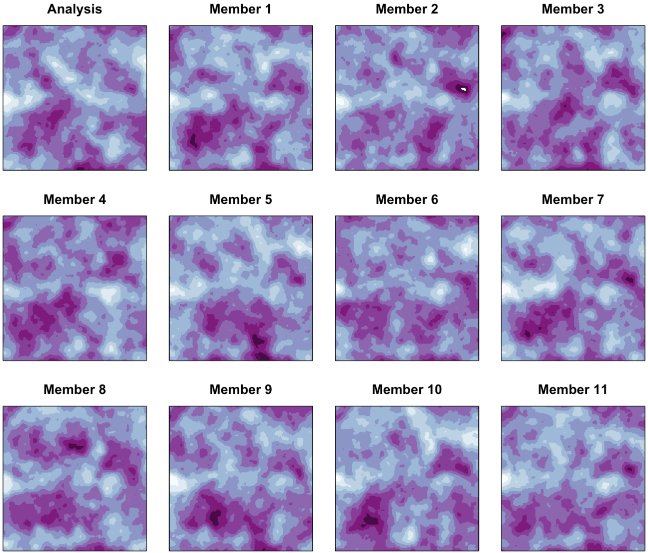
Simulation algorithm

2. Simulate 11 independent **perturbation fields** W_i under the univariate Whittle-Matérn model with the **same spatial parameters** as the ensemble mean, and standard Gaussian marginal distributions. Construct the ensemble under the following weighted average:

$$Z_i(s) = \omega Z_M(s) + \sqrt{1-\omega^2} W_i(s), \qquad i = 1, \ldots, 11.$$

The resulting ensemble fields have standard Gaussian marginal distributions with "skill" controlled by parameter ω.

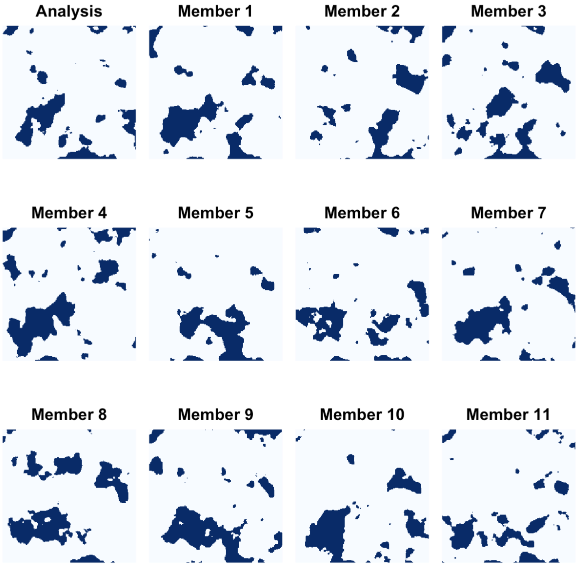
Simulation algorithm



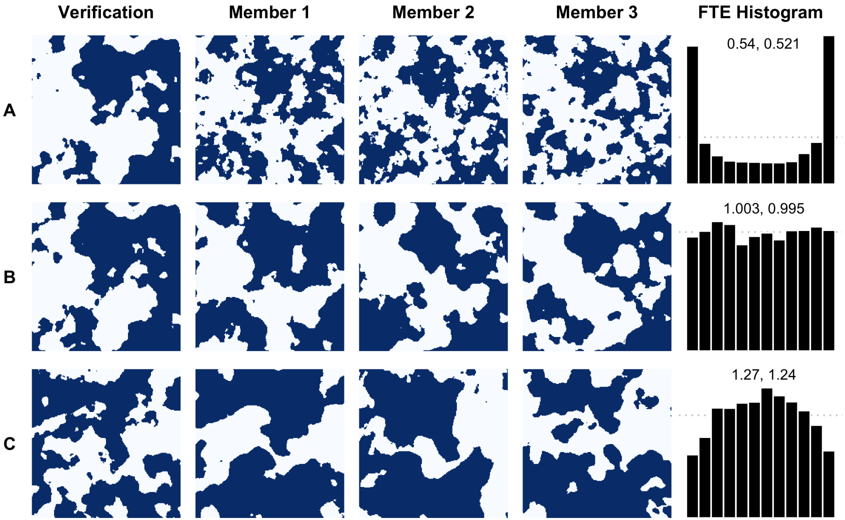
Simulation algorithm

3. Binary exceedance fields are obtained by thresholding at τ.

Here: $\tau = 1$

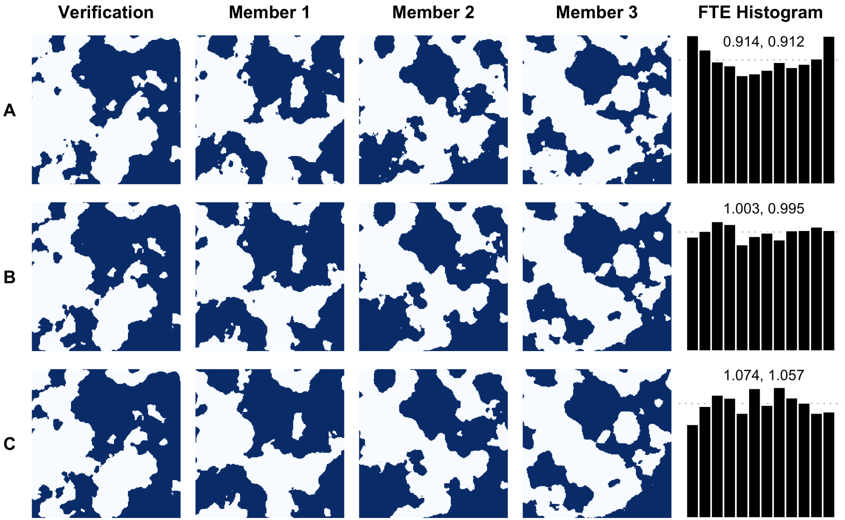


Analysis: does the FTE detect obvious miscalibration?



FTE histograms detect type of miscalibration accurately.

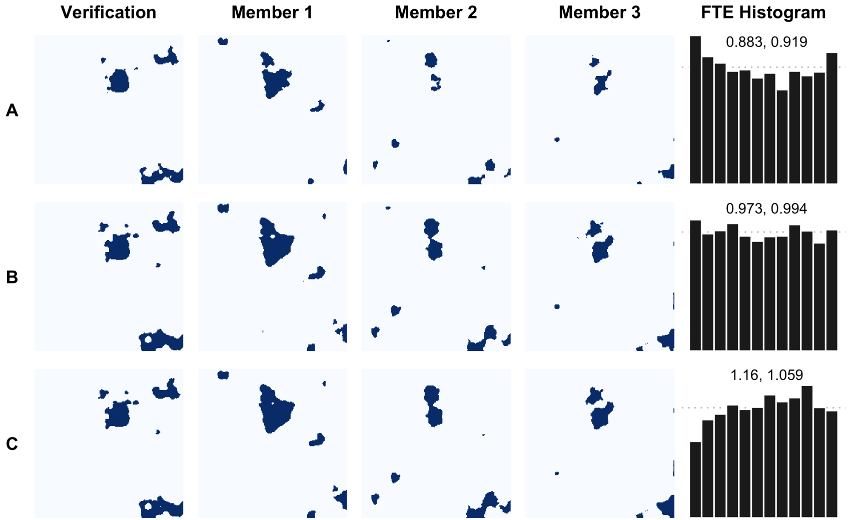
Analysis: non-obvious miscalibration



FTE histograms can detect even minor issues with calibration.

Here: 10%

Analysis: miscalibration at high thresholds

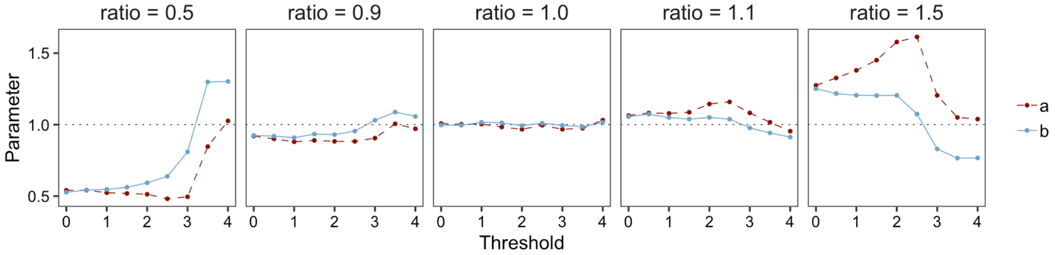


Even at high thresholds, miscalibration is still identified accurately.

Here: $\tau = 2$

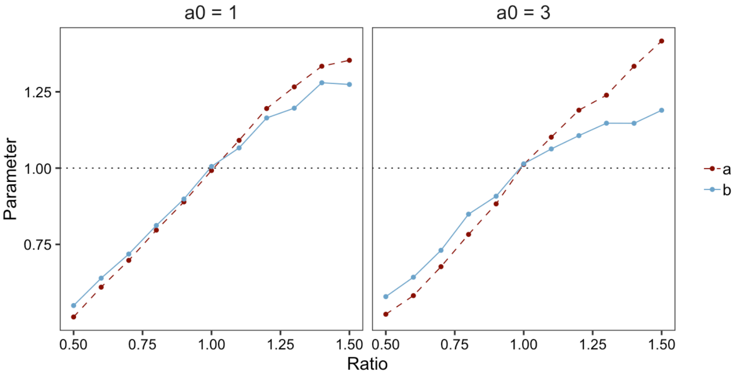
Summary statistics are “sufficient”

Summary statistics alone can accurately characterize the type of (mis)calibaration in the ensemble.



Sensitivity to domain size

Steep slopes around a correlation length ratio of 1.0 indicate that the FTE metric maintains good discrimination ability regardless of domain size.



Verification of downscaled forecast fields

- Hydrological models often require inputs at a relatively high resolution, but for longer lead times, only forecasts from global ensemble forecasts systems are available.
- These come at a coarse range and need to be downscaled to a high-resolution grid.
- Here we use a combination of Scheuerer and Hamill (2015) and Gagnon et al. (2012) to obtain high-res forecasts based on NOAA's GEFS data.

Verification of downscaled forecast fields

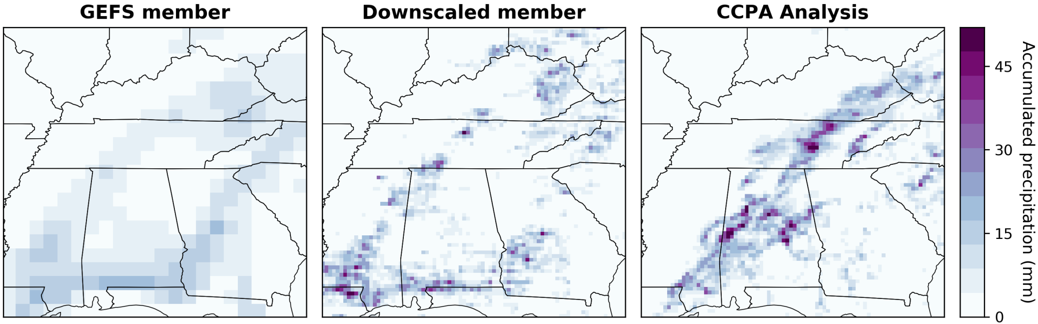
- Hydrological models often require inputs at a relatively high resolution, but for longer lead times, only forecasts from global ensemble forecasts systems are available.

- These come at a coarse range and need to be downscaled to a high-resolution grid.

- Here we use a combination of Scheuerer and Hamill (2015) and Gagnon et al. (2012) to obtain high-res forecasts based on NOAA's GEFS data.

Question: Does the method produce fields with appropriate fine-scale variability?

Daily precipitation

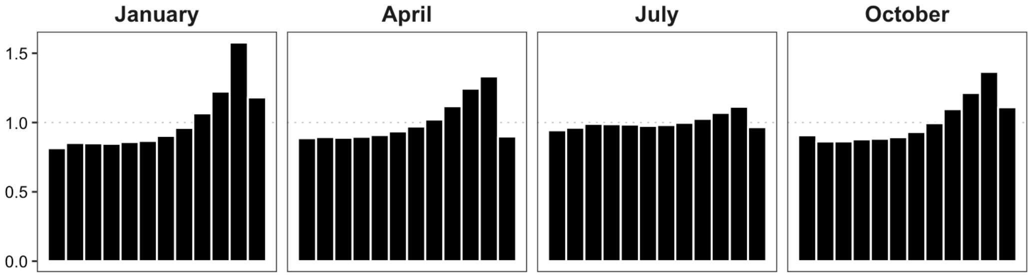


- South-Eastern US
- 2002 - 2016
- 66h lead time, 6h accumulation
- Forecasts: GEFS (0.5°)
- Analyses: CCPA (0.125°)

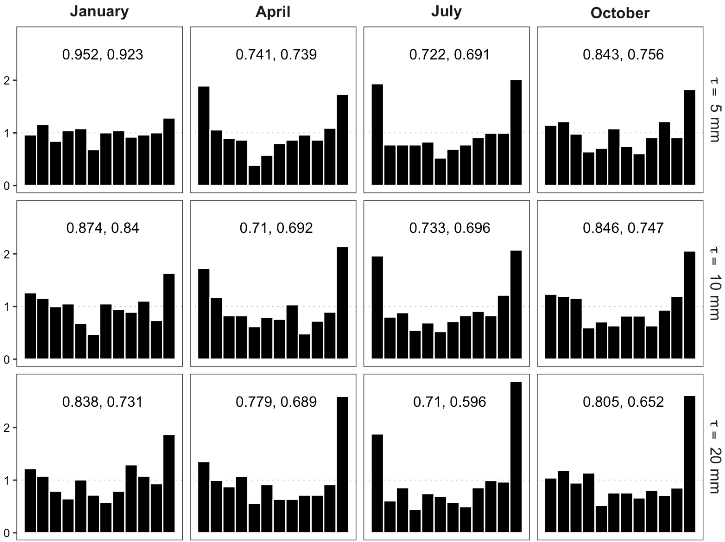
Seasonal univariate verification

Checking for issues with marginal calibration upfront:

- Consistent bias toward underestimation of accumulation levels
- Otherwise fairly uniform; no sign of over- or under-dispersion



Verification of spatial structure



- Fixed thresholds at 5mm, 10mm, 20mm
- More uniform in fall/winter
- Under-dispersion in spring/summer
- Similar across all thresholds

Conclusions

- The FTE is a verification metric to assess whether ensemble forecasts correctly represent the **spatial dependence structure** in forecast fields.
- The interpretation is **familiar and intuitive**, and is related to quantities that are relevant in practice.
- FTE histograms are surprisingly **sensitive to miscalibration** of spatial dependence and have proven useful in identifying issues with statistical downscaling algorithms.

Jacobson, J., Kleiber, W., Scheuerer, M., and Bellier, J. (2020). Beyond univariate calibration: Verifying spatial structure in ensembles of forecast fields. *Nonlinear Processes in Geophysics Discussions*, 2020:1–20