Assessing multivariate forecast calibration

E-values and pre-rank functions

Sam Allen David Ginsbourger Johanna Ziegel

UNIVERSITÄT BERN

OESCHGER CENTRE
CLIMATE CHANGE RESEARCH

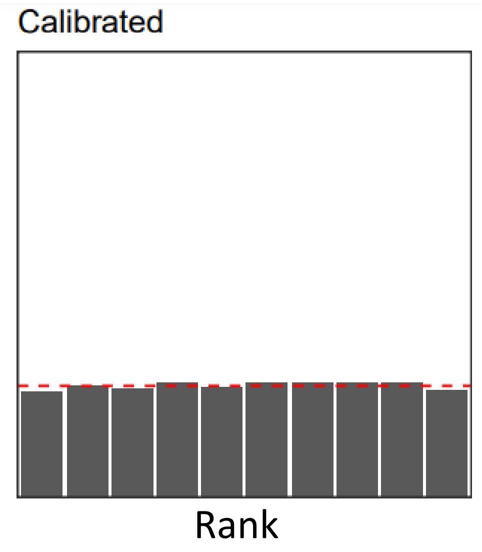
MeteoSwiss

Motivation

- Forecast verification allows us to understand how our forecasts behave
- Probabilistic forecasts should be as sharp as possible, subject to being calibrated (Gneiting et al., 2007)
- Calibration is a minimum requirement for forecasts to be used optimally
- The calibration of multivariate forecasts is rarely assessed in practice

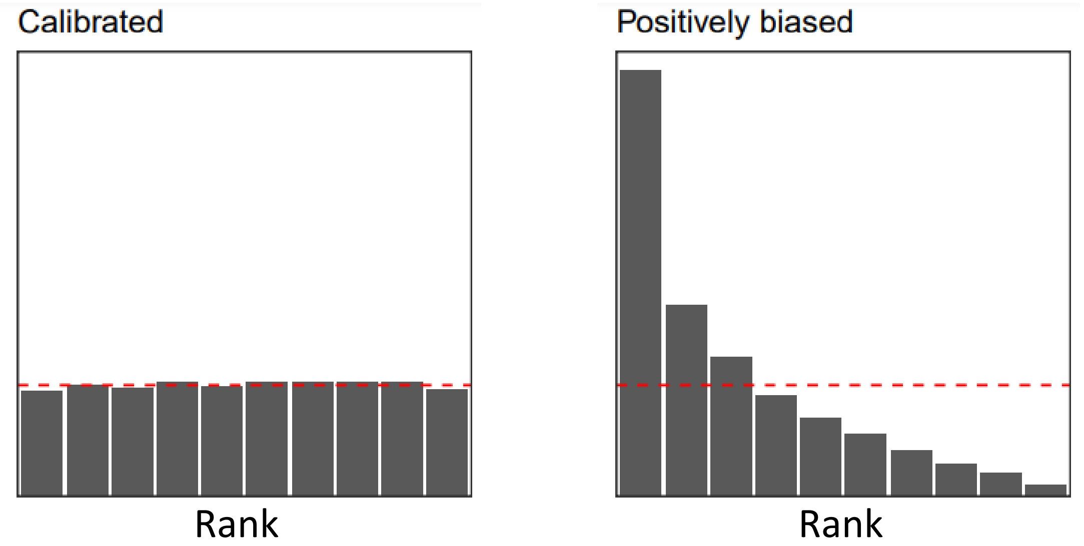
Rank histograms

- Univariate calibration is typically assessed using rank histograms
- An ensemble prediction system is calibrated if its rank histogram is flat



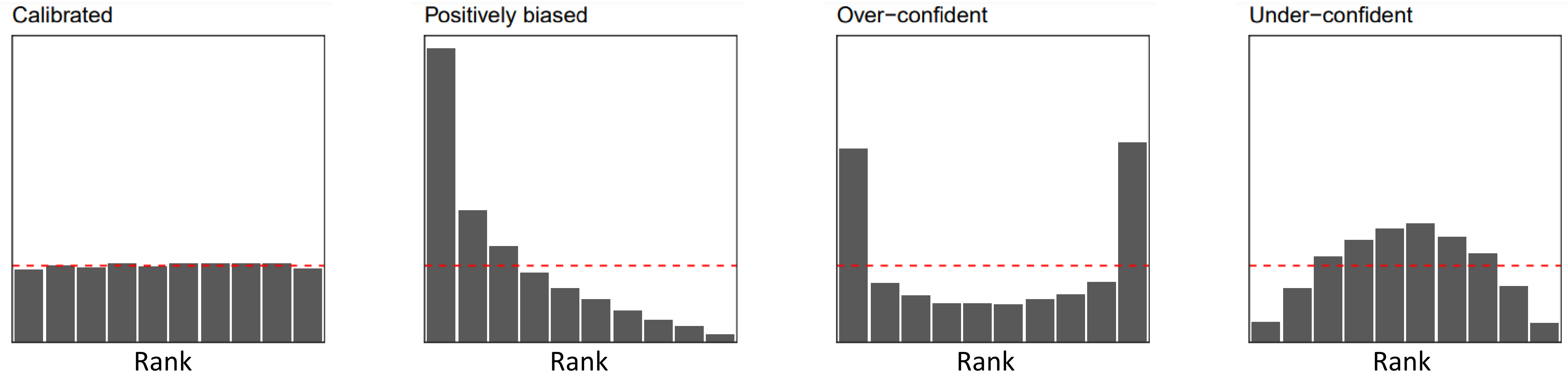
Rank histograms

- Univariate calibration is typically assessed using rank histograms
- An ensemble prediction system is calibrated if its rank histogram is flat
- A \ or /-shaped histogram suggests the ensemble forecasts are biased



Rank histograms

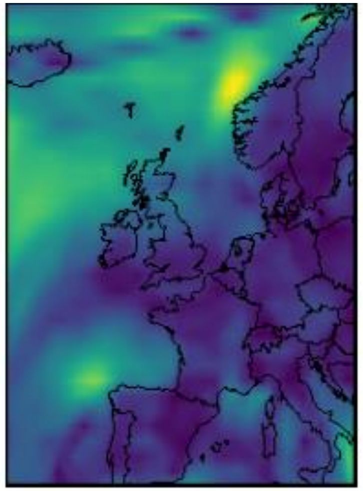
- Univariate calibration is typically assessed using rank histograms
- An ensemble prediction system is calibrated if its rank histogram is flat
- A \ or /-shaped histogram suggests the ensemble forecasts are biased
- A ∪ or ∩ -shaped histogram suggests the ensemble forecasts are over/under-confident



Multivariate rank histograms

- Multivariate calibration can be assessed using multivariate rank histograms

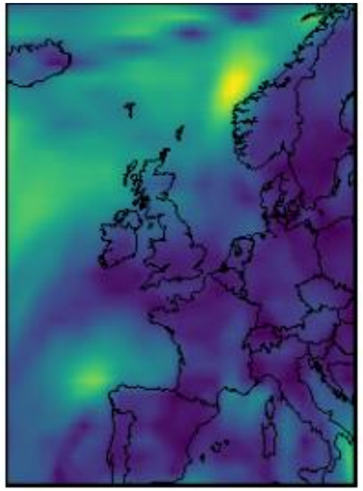
$\boldsymbol{x} = (x_1, \dots, x_d)$



Multivariate rank histograms

- Multivariate calibration can be assessed using multivariate rank histograms
- Multivariate rank histograms transform the multivariate observations into univariate objects

$\boldsymbol{x} = (x_1, \ldots, x_d)$

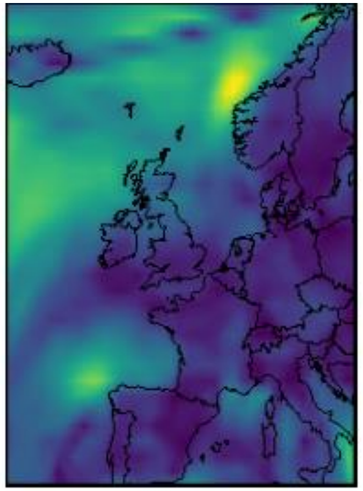


Pre-rank function
$\rho: \mathbb{R}^d \rightarrow \mathbb{R}$

Multivariate rank histograms

- Multivariate calibration can be assessed using multivariate rank histograms
- Multivariate rank histograms transform the multivariate observations into univariate objects
- A univariate rank histogram can be constructed from the transformed observations

$\boldsymbol{x} = (x_1, \dots, x_d)$



Pre-rank function
$\rho: \mathbb{R}^d \to \mathbb{R}$

Pre-rank $\rho(\boldsymbol{x}) \in \mathbb{R}$

Pre-rank functions

- Several pre-rank functions have been proposed in the literature

Pre-rank functions

- Several pre-rank functions have been proposed in the literature, based on:
  - Minimum spanning trees (Smith and Hansen, 2004; Wilks, 2004)
  - Multivariate ranks (Gneiting et al., 2008)
  - Averages of univariate ranks (Thorarinsdottir et al., 2016)
  - Band-depths (Thorarinsdottir et al., 2016)
  - Threshold exceedances (Scheuerer and Hamill, 2018)
  - Multivariate scoring rules (Knüppel et al., 2022)

Pre-rank functions

- Several pre-rank functions have been proposed in the literature, based on:
  - Minimum spanning trees (Smith and Hansen, 2004; Wilks, 2004)
  - Multivariate ranks (Gneiting et al., 2008)
  - Averages of univariate ranks (Thorarinsdottir et al., 2016)
  - Band-depths (Thorarinsdottir et al., 2016)
  - Threshold exceedances (Scheuerer and Hamill, 2018)
  - Multivariate scoring rules (Knüppel et al., 2022)

- Studies recommend employing a variety of pre-rank functions (Wilks, 2017)

Pre-rank functions

- Several pre-rank functions have been proposed in the literature, based on:
  - Minimum spanning trees (Smith and Hansen, 2004; Wilks, 2004)
  - Multivariate ranks (Gneiting et al., 2008)
  - Averages of univariate ranks (Thorarinsdottir et al., 2016)
  - Band-depths (Thorarinsdottir et al., 2016)
  - Threshold exceedances (Scheuerer and Hamill, 2018)
  - Multivariate scoring rules (Knüppel et al., 2022)

- Studies recommend employing a variety of pre-rank functions (Wilks, 2017)

- Multivariate rank histograms are rarely employed in practice

Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$

Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$
- The function can extract information about the predicted…

- Average $\rho(\boldsymbol{x}) = \bar{x}$

Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$
- The function can extract information about the predicted…

- Average $\rho(\boldsymbol{x}) = \bar{x}$
- Variation $\rho(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^2$

Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$
- The function can extract information about the predicted…

- Average $\rho(\boldsymbol{x}) = \bar{x}$
- Variation $\rho(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^2$
- Dependence $\rho(\boldsymbol{x}) = \dfrac{\gamma_{\boldsymbol{x}}(h)}{\sigma_{\boldsymbol{x}}^2}$

Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$
- The function can extract information about the predicted…

- Average $\rho(\boldsymbol{x}) = \bar{x}$
- Variation $\rho(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^2$
- Dependence $\rho(\boldsymbol{x}) = \dfrac{\gamma_{\boldsymbol{x}}(h)}{\sigma_{\boldsymbol{x}}^2}$
- High-impact events $\rho(\boldsymbol{x}) = \sum \mathbb{1}\{x_j > t\}$

Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$
- The function can extract information about the predicted…

- Average $\rho(\boldsymbol{x}) = \bar{x}$
- Variation $\rho(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^2$
- Dependence $\rho(\boldsymbol{x}) = \dfrac{\gamma_{\boldsymbol{x}}(h)}{\sigma_{\boldsymbol{x}}^2}$
- High-impact events $\rho(\boldsymbol{x}) = \sum \mathbb{1}\{x_j > t\}$
- Weather patterns $\rho(\boldsymbol{x}) = \boldsymbol{e}^{(\boldsymbol{i})} \cdot \boldsymbol{x}$

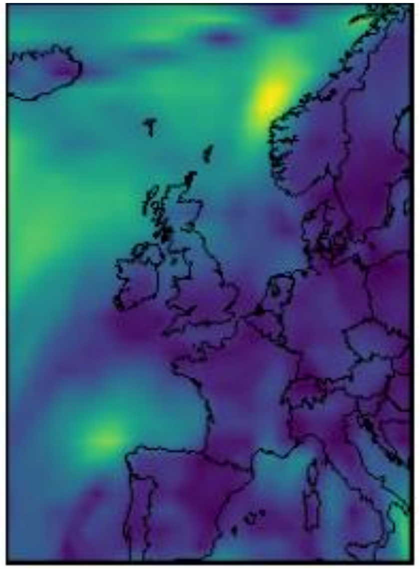
Pre-rank functions

- We can choose any function from $\mathbb{R}^d$ to $\mathbb{R}$
- The function can extract information about the predicted…

- Average $\rho(\boldsymbol{x}) = \bar{x}$
- Variation $\rho(\boldsymbol{x}) = \sigma_{\boldsymbol{x}}^2$
- Dependence $\rho(\boldsymbol{x}) = \frac{\gamma_{\boldsymbol{x}}(h)}{\sigma_{\boldsymbol{x}}^2}$
- High-impact events $\rho(\boldsymbol{x}) = \sum \mathbb{1}\{x_j > t\}$
- Weather patterns $\rho(\boldsymbol{x}) = \boldsymbol{e}^{(\boldsymbol{i})} \cdot \boldsymbol{x}$
- Isotropy $\rho(\boldsymbol{x}) = \left(\gamma_{\boldsymbol{x}}(h) - \gamma_{\boldsymbol{x}}(h')\right)^2$

Case study

- Consider wind speed forecast fields over western Europe
  - 1353 grid points
- Forecasts are obtained from NCEP's global ensemble forecasting system
  - Ensemble comprised of 10 exchangeable members
- Daily forecasts available over 10 extended cold seasons from 2001-2010
  - Lead time of 5 days
- Compare the NCEP ensemble forecasts to statistically post-processed forecasts (EMOS + ECC)

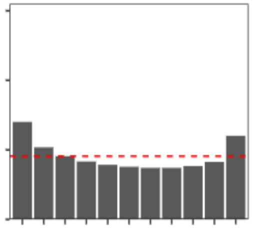


Case study

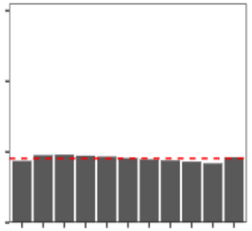
- Post-processing improves univariate calibration

Univariate

Raw ensemble

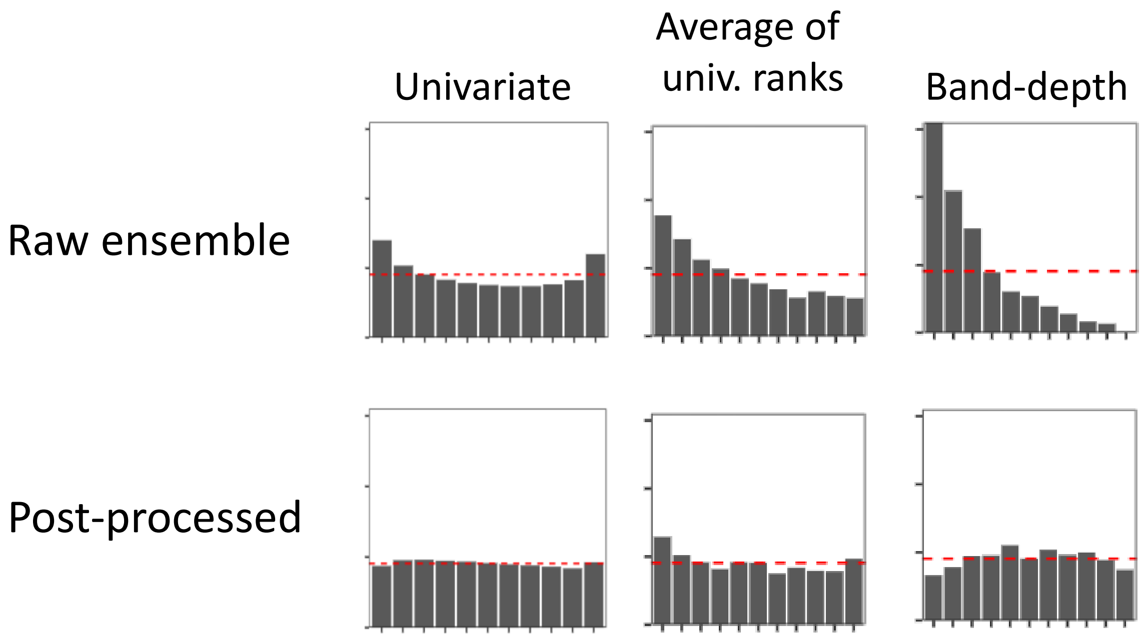


Post-processed



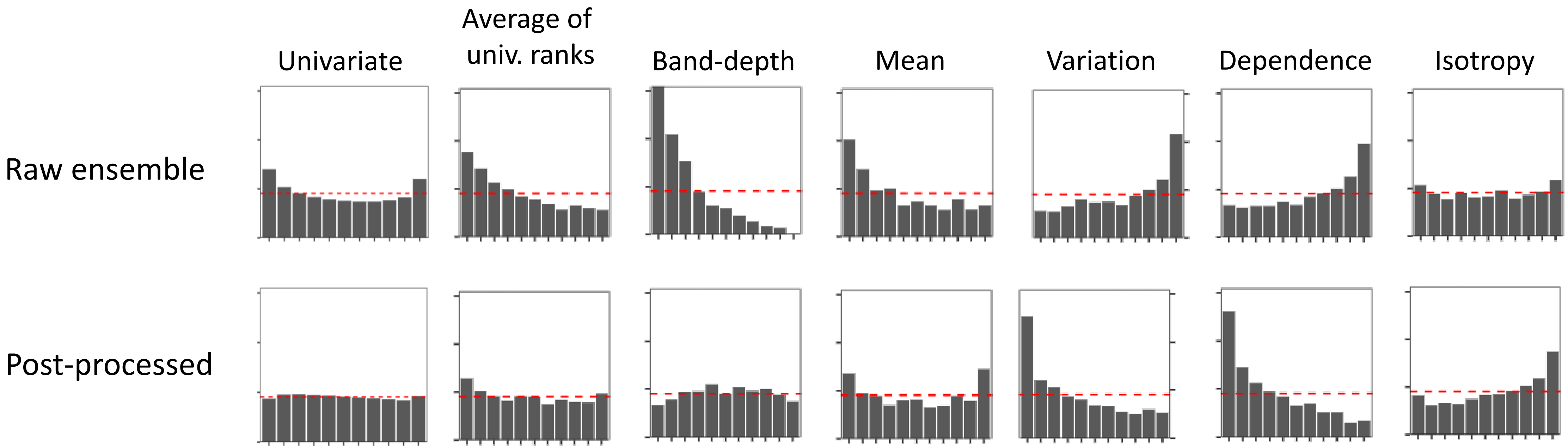
Case study

- Post-processing improves univariate calibration



Case study

- Post-processing improves univariate calibration
- But the forecasts do not reliably predict the dependence between neighbouring grid points

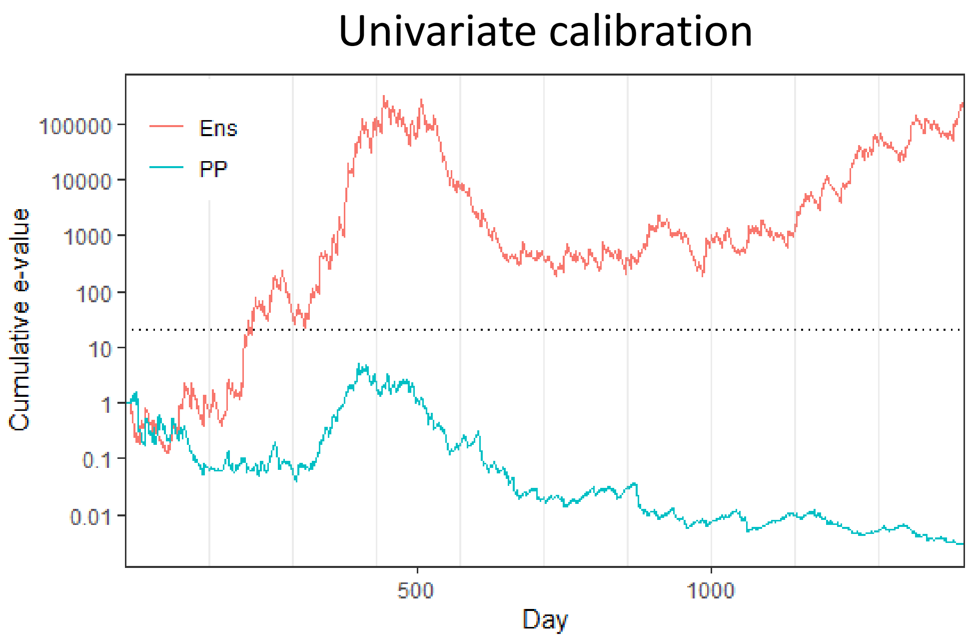


E-values

- We often want to formally test whether our forecasts are calibrated
- An appealing univariate approach is based on e-values (Arnold et al., 2021)
- E-values provide hypothesis tests that are valid under optimal stopping
  - We do not need to fix the test period apriori
- This accounts for the sequential nature of forecasting

E-values

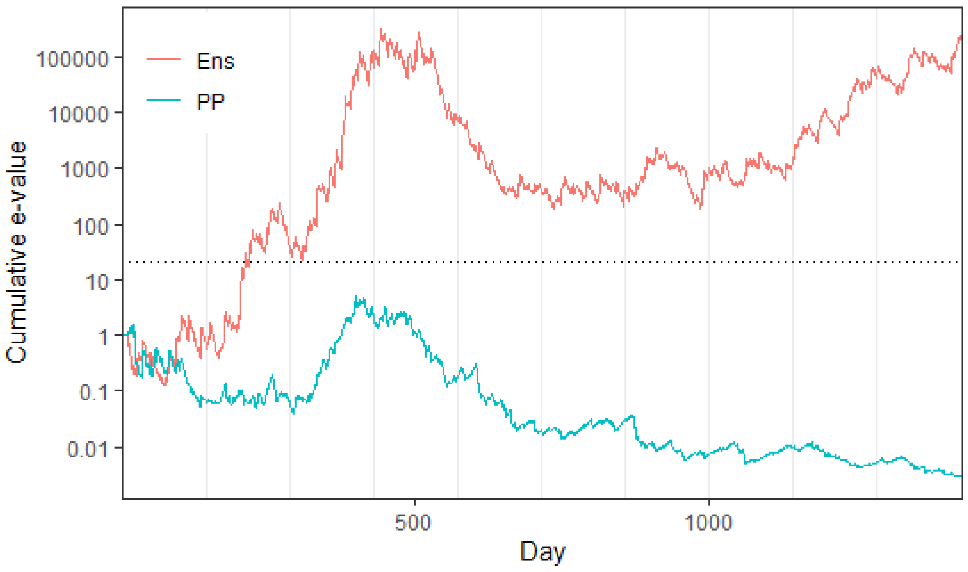
- E-values provide a measure of forecast miscalibration that can be monitored over time
- Inverted e-values are (conservative) p-values



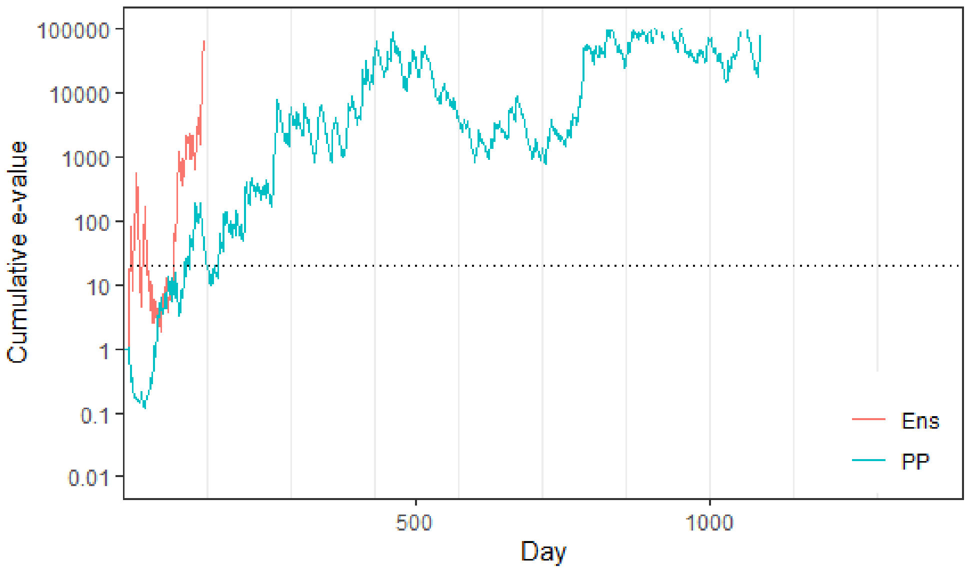
E-values

- E-values provide a measure of forecast miscalibration that can be monitored over time
- Inverted e-values are (conservative) p-values

Univariate calibration



Multivariate dependence calibration



Summary

- Calibration is a minimum requirement for probabilistic forecasts to be used optimally
- Multivariate calibration can be assessed using multivariate rank histograms
  - These are rarely employed in practice
- We can choose any function to transform multivariate observations to univariate objects
  - This leads to interpretable, user-specific checks for calibration
- These transformations could also be employed within scoring rules
- E-values provide an appealing framework with which to monitor and test forecast calibration

References

- Arnold, S., Henzi, A., & Ziegel, J. F. (2021). Sequentially valid tests for forecast calibration. *arXiv preprint arXiv:2109.11761*.

- Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B, 69*, 243-268.

- Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., & Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test*, *17*, 211-235.

- Knüppel, M., Krüger, F., & Pohle, M. O. (2022). Score-based calibration testing for multivariate forecast distributions. *arXiv preprint arXiv:2211.16362*.

- Scheuerer, M., & Hamill, T. M. (2018). Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *Journal of Hydrometeorology*, *19*, 1651-1670.

References

- Smith, L. A., & Hansen, J. A. (2004). Extending the limits of ensemble forecast verification with the minimum spanning tree. *Monthly Weather Review*, *132*, 1522-1528.

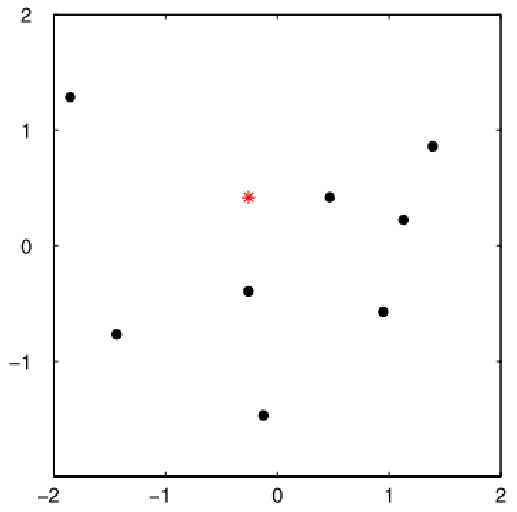
- Thorarinsdottir, T. L., Scheuerer, M., & Heinz, C. (2016). Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *Journal of computational and graphical statistics*, *25*, 105-122.

- Wilks, D. S. (2004). The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review*, *132*, 1329-1340.

- Wilks, D. S. (2017). On assessing calibration of multivariate ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *143*, 164-172.

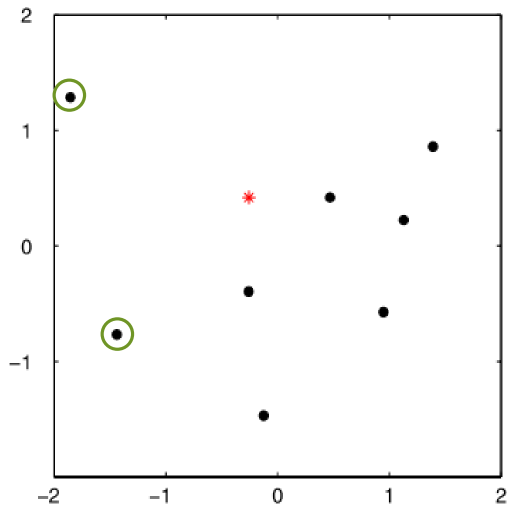
Pre-rank functions

- Example: Average rank (Thorarinsdottir et al,. 2016)
- The average rank considers the average of the ranks along each dimension



Pre-rank functions

- Example: Average rank (Thorarinsdottir et al,. 2016)
- The average rank considers the average of the ranks along each dimension

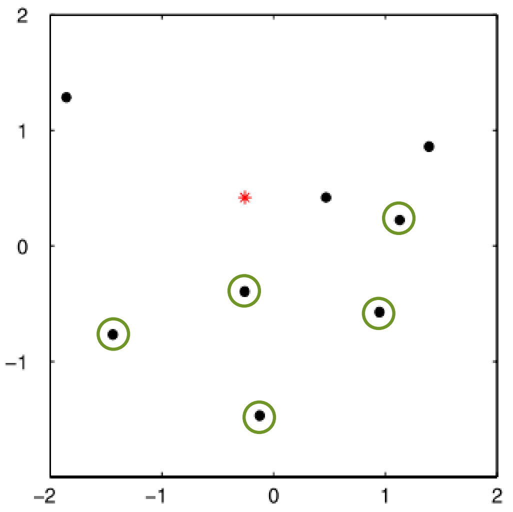


Rank of observation in first dimension: 3

Pre-rank functions

- Example: Average rank (Thorarinsdottir et al,. 2016)
- The average rank considers the average of the ranks along each dimension

Rank of observation in second dimension: 6

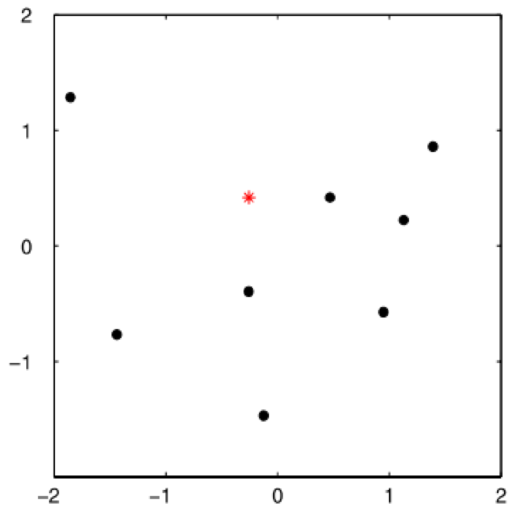


Rank of observation in first dimension: 3

Pre-rank functions

- Example: Average rank (Thorarinsdottir et al,. 2016)
- The average rank considers the average of the ranks along each dimension

Rank of observation in second dimension: 6



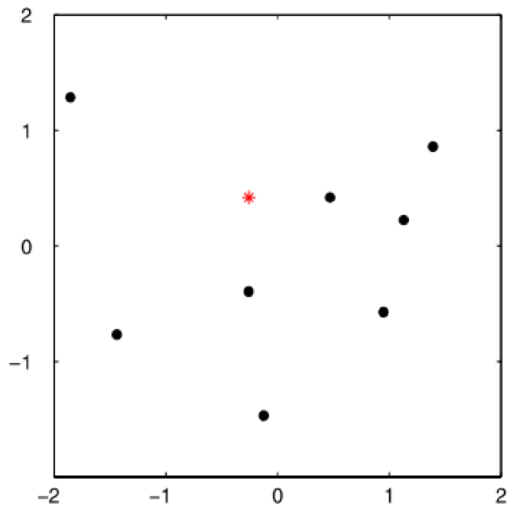
Rank of observation in first dimension: 3

$$\text{Average rank} = \frac{3+6}{2} = 4.5$$

Pre-rank functions

- Example: Average rank (Thorarinsdottir et al,. 2016)
- The average rank considers the average of the ranks along each dimension

Rank of observation in second dimension: 6



Rank of observation in first dimension: 3

$$\text{Average rank} = \frac{3+6}{2} = 4.5$$

Average ranks of the 8 ensemble members: $(2, 3, 4, 5, 5, 6.5, 6.5, 8{,}5)$

The overall rank of the observation is: 4

Simulation study

- We can simulate multivariate forecasts with particular errors ($d = 30 \times 30 = 900$)

Simulation study