Ensemble Verification II

Martin Leutbecher

ECMWF

Training Course 2023

Ensemble Verification II

Martin Leutbecher

ECMWF

Training Course 2023

1. Proper scores
2. Continuous scalar variables
3. Observation uncertainty
4. Climatological distribution
5. Statistical significance
6. Outlook

Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.

Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.
- Let us assume that we predict the distribution $p_{\mathrm{fc}}(x)$ while the verification is distributed according to a distribution $p_y(x)$. Not all scores indicate maximum skill for $p_{\mathrm{fc}} = p_y$.

Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.
- Let us assume that we predict the distribution $p_{\mathrm{fc}}(x)$ while the verification is distributed according to a distribution $p_y(x)$. Not all scores indicate maximum skill for $p_{\mathrm{fc}} = p_y$.
- A score (or scoring rule) is *(strictly) proper* if the score reaches its optimal value if (and only if) the predicted distribution is equal to the distribution of the verification.

Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.
- Let us assume that we predict the distribution $p_{\mathrm{fc}}(x)$ while the verification is distributed according to a distribution $p_y(x)$. Not all scores indicate maximum skill for $p_{\mathrm{fc}} = p_y$.
- A score (or scoring rule) is *(strictly) proper* if the score reaches its optimal value if (and only if) the predicted distribution is equal to the distribution of the verification.
- If a forecaster is judged by a score that is not proper, (s)he is encouraged to issue forecasts which are not the best forecasts according to their true belief!

  In such a situation, one says that the forecast is *hedged* or that the forecaster *plays the score*.

Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.
- Let us assume that we predict the distribution $p_{\mathrm{fc}}(x)$ while the verification is distributed according to a distribution $p_y(x)$. Not all scores indicate maximum skill for $p_{\mathrm{fc}} = p_y$.
- A score (or scoring rule) is *(strictly) proper* if the score reaches its optimal value if (and only if) the predicted distribution is equal to the distribution of the verification.
- If a forecaster is judged by a score that is not proper, (s)he is encouraged to issue forecasts which are not the best forecasts according to their true belief!

  In such a situation, one says that the forecast is *hedged* or that the forecaster *plays the score*.
- Examples of proper scores are: Brier Score, continuous (and discrete) ranked probability score, quantile score, logarithmic score

Proper scores

- A score for a probabilistic forecast is a summary measure that evaluates the probability distribution. This condenses all the information into a single number and can be potentially misleading.
- Let us assume that we predict the distribution $p_{\mathrm{fc}}(x)$ while the verification is distributed according to a distribution $p_y(x)$. Not all scores indicate maximum skill for $p_{\mathrm{fc}} = p_y$.
- A score (or scoring rule) is *(strictly) proper* if the score reaches its optimal value if (and only if) the predicted distribution is equal to the distribution of the verification.
- If a forecaster is judged by a score that is not proper, (s)he is encouraged to issue forecasts which are not the best forecasts according to their true belief!
  In such a situation, one says that the forecast is *hedged* or that the forecaster *plays the score*.
- Examples of proper scores are: Brier Score, continuous (and discrete) ranked probability score, quantile score, logarithmic score
- see Gneiting and Raftery (2007) for more details

Example of a score that is not proper

- consider the linear score: $\mathrm{LinS} = |p - o|$
- dichotomous event e: e occured $(o = 1)$, e did not occur $(o = 0)$
- assume the event occurs with the true probability of 0.4
- If the prediction is 0.4, the *expected* linear score is

$$\mathrm{E}(\mathrm{LinS}) = 0.4|0.4 - 1| + (1 - 0.4)\,|0.4 - 0| = 0.48$$

- If the prediction is instead 0, the expected linear score is

$$\mathrm{E}(\mathrm{LinS}) = 0.4|0 - 1| + (1 - 0.4)|0 - 0| = 0.40$$

It is not hard to prove that the Brier score is strictly proper (e.g. Wilks 2011)

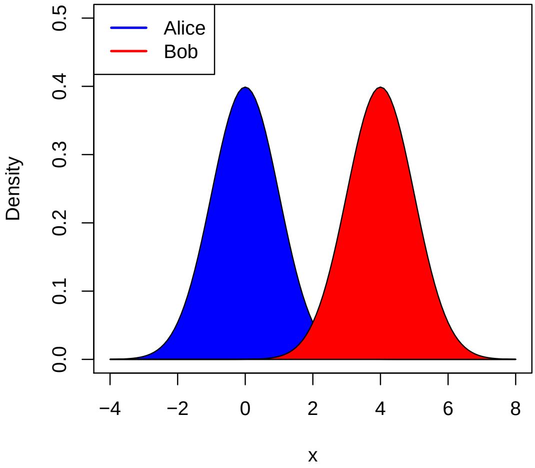
Sebastian Lerch's example with two proper scores

Simple idealised example

We compare Alice's and Bob's forecasts for $Y \sim \mathcal{N}(0,1)$,

$$F_{\text{Alice}} = \mathcal{N}(0,1)$$

$$F_{\text{Bob}} = \mathcal{N}(4,1)$$



Sebastian Lerch's example with two proper scores

Simple idealised example

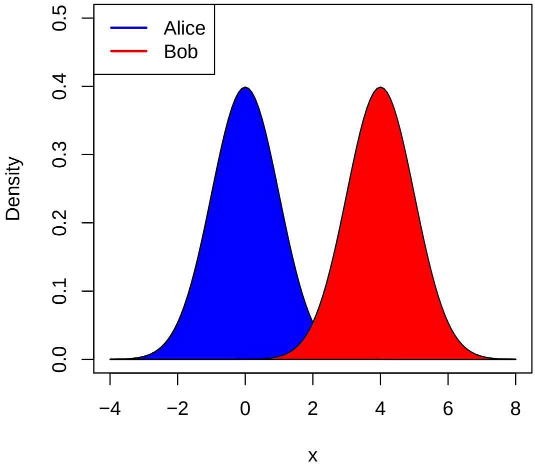
We compare Alice's and Bob's forecasts for $Y \sim \mathcal{N}(0, 1)$,

$$F_{\text{Alice}} = \mathcal{N}(0, 1)$$

$$F_{\text{Bob}} = \mathcal{N}(4, 1)$$

Based on $10,000$ forecast experiments,

Forecaster	CRPS	LogS
Alice	**0.56**	**1.42**
Bob	3.53	9.36



A conditional sample for evaluating Alice and Bob

Simple toy example

Based on the 10 largest observations,

Forecaster	CRPS	LogS
Alice	2.70	6.29
Bob	**0.46**	**1.21**

The forecaster's dilemma

More generally, for non-constant weight functions w, any scoring rule

$$S^*(F, y) = w(y)S(F, y)$$

is improper even if S is a proper scoring rule (Gneiting and Ranjan, 2011). Here, y and F denote the verifying observation and the predicted distribution, respectively.

Forecaster's dilemma

Forecast evaluation only based on a subset of extreme observations corresponds to *improper* verification methods and is bound to discredit skillful forecasters.

Acknowledgement: Forecaster's dilemma and Alice and Bob's forecast based on slides provided by Sebastian Lerch (Heidelberg Institute for Theoretical Studies), see also Lerch et al. (2017)

Scores for probabilistic/ensemble forecasts of continuous scalar variables

some (but not all) useful measures

- RMSE and other scores used for single forecasts applied to ensemble mean
- rank histograms (reliability again)

Scores for probabilistic/ensemble forecasts of continuous scalar variables

some (but not all) useful measures

- RMSE and other scores used for single forecasts applied to ensemble mean
- rank histograms (reliability again)
- continuous ranked probability score (reliability *and* resolution)
- quantile score (reliability *and* resolution)
- logarithmic score (for Gaussian) (reliability *and* resolution)
- reliability of the ensemble spread (domain-integrated and local)

Continuous ranked probability score

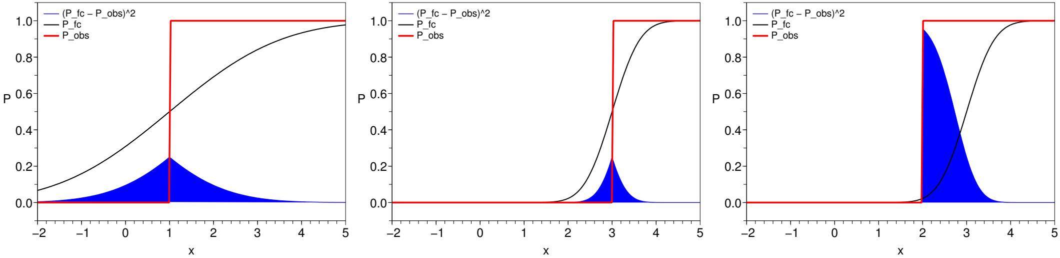
CRPS = Mean squared error of the cumulative distribution P_{fc}

cdf of observation $P_y(x) = P(y \leq x) = H(x - y) = \mathbb{1}\{y \leq x\}$

cdf of forecast $P_{\mathrm{fc}}(x) = P(x_{\mathrm{fc}} \leq x)$

Here, H and $\mathbb{1}$ denote the Heaviside step function and the indicator function, respectively.

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} (P_{\mathrm{fc}}(x) - P_y(x))^2 \, \mathrm{d}x = \int_{-\infty}^{+\infty} \mathrm{BS}_x \, \mathrm{d}x$$



Continuous ranked probability score

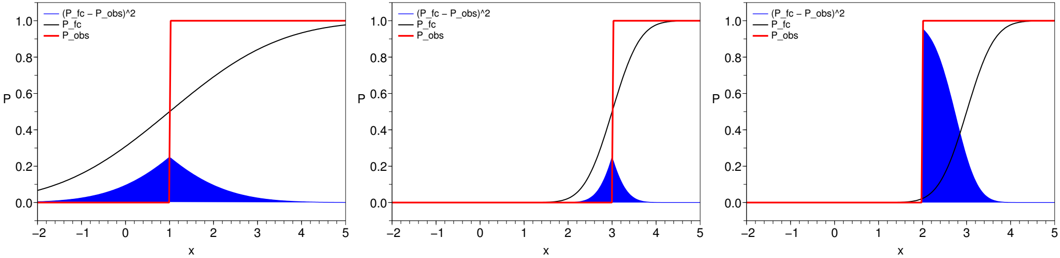
CRPS = Mean squared error of the cumulative distribution P_{fc}

cdf of observation $P_y(x) = P(y \leq x) = H(x - y) = \mathbb{1}\{y \leq x\}$

cdf of forecast $P_{\mathrm{fc}}(x) = P(x_{\mathrm{fc}} \leq x)$

Here, H and $\mathbb{1}$ denote the Heaviside step function and the indicator function, respectively.

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} \left(P_{\mathrm{fc}}(x) - P_y(x)\right)^2 \, \mathrm{d}x = \int_{-\infty}^{+\infty} \mathrm{BS}_x \, \mathrm{d}x$$



equal to mean absolute error for a single forecast

How to compute the CRPS

Ensemble

The integral $\int \ldots \mathrm{d}x$ can be evaluated exactly by using the intervals defined by the M ensemble forecasts and the verification rather than some fixed interval Δx:

HERSBACH (2000)

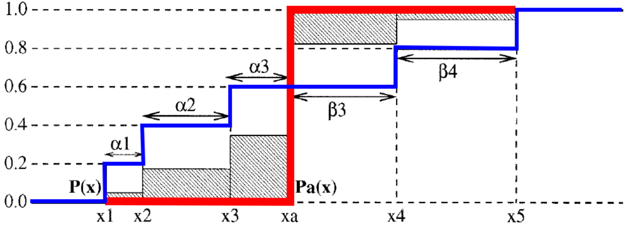


FIG. 2. Cumulative distribution for an ensemble $\{x_1, \ldots, x_5\}$ of five members (thick solid line) and for the verifying analysis x_a (thin solid line). The CRPS is represented by the shaded area. The α_i and β_i are defined in Eq. (26).

$0 < i < N$	α_i	β_i
$x_a > x_{i+1}$	$x_{i+1} - x_i$	0
$x_{i+1} > x_a > x_i$	$x_a - x_i$	$x_{i+1} - x_a$
$x_a < x_i$	0	$x_{i+1} - x_i$

$$\mathrm{CRPS} = \sum_{j=0}^{M} c_j$$

$$c_j = \alpha_j p_j^2 + \beta_j (1 - p_j)^2$$

$$p_j = j/M$$

How to compute the CRPS

Gaussian distribution

- For a Gaussian distribution an analytical formula for the CRPS is available.
- Assume that the predicted Gaussian has mean μ and variance σ^2 and that the verification is denoted by y.

$$\mathrm{CRPS} = \frac{\sigma}{\sqrt{\pi}}\left[-1 + \sqrt{\pi}\,\frac{y-\mu}{\sigma}\,\Phi\left(\frac{y-\mu}{\sqrt{2}\sigma}\right) + \sqrt{2}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)\right]$$

- Here, Φ denotes the error function $\Phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\,\mathrm{d}t$.
- This relationship is particularly useful for calibration purposes (Non-homogeneous Gaussian regression).
- Leutbecher and Haiden (2021) approximate the joint distribution of forecast and observations by a homogeneous Gaussian distribution. Then, the expected CRPS can be expressed as a function of the variance of the error of the ensemble mean, the spread-error ratio and the bias (mean error of the ensemble mean)

Kernel representation of CRPS

$$\mathrm{CRPS}\left(\{x_j\}_M, y\right) = \frac{1}{M}\sum_{j=1}^{M}|x_j - y| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|$$

where $x_j \quad (j = 1, \ldots, M)$ denote the ensemble members.

ECMWF

Kernel representation of CRPS

$$\mathrm{CRPS}\left(\{x_j\}_M, y\right) = \frac{1}{M}\sum_{j=1}^{M}|x_j - y| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|$$

where $x_j \quad (j = 1, \ldots, M)$ denote the ensemble members.

Using $1/\left[2M(M-1)\right]$ as normalisation factor for the second term yields the fair CRPS. Under certain assumptions (exchangeability of members) the expected value of the fair CRPS is independent of ensemble size. The fair CRPS estimates the CRPS one would obtain from an ensemble with infinitely many members that are sampled from the same distribution as the existing members.

Kernel representation of CRPS

$$\mathrm{CRPS}\left(\{x_j\}_M, y\right) = \frac{1}{M}\sum_{j=1}^{M}|x_j - y| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|$$

where $x_j \quad (j = 1, \ldots, M)$ denote the ensemble members.

Using $1/\left[2M(M-1)\right]$ as normalisation factor for the second term yields the fair CRPS. Under certain assumptions (exchangeability of members) the expected value of the fair CRPS is independent of ensemble size. The fair CRPS estimates the CRPS one would obtain from an ensemble with infinitely many members that are sampled from the same distribution as the existing members.

The kernel representation can be generalized to higher dimensions using the Euclidean norm of the vector differences of members and observation:

$$\text{energy score} = \frac{1}{M}\sum_{j}\|\mathbf{x}_j - \mathbf{y}\| - \frac{1}{2M^2}\sum_{j}\sum_{k}\|\mathbf{x}_j - \mathbf{x}_k\|$$

CRPS

Decomposition

- The CRPS can be decomposed into a reliability component and a resolution component.
- The CRPS is additive: The CRPS for the union of two samples is the weighted (arithmetic) average of the CRPS of the two samples with the weights proportional to the respective sample sizes.
- The components of the CRPS are not additive. The components can be computed from the sample averages of the α_j and β_j distances.
- This is similar to the decomposition of the Brier score. However, the reliability (resolution) component of the CRPS is not the integral of the reliability (resolution) component of the Brier scores.
- The reliability component of the CRPS is related to the rank histogram but not identical.
- see Hersbach (2000) for details (but note that there are alternative decompositions, e.g. Leutbecher and Haiden (2021))

CRPS with threshold-weighting

Can be used for instance to focus on the tails of the climatological distribution, e.g. strong wind, intense rainfall.

The threshold-weighted CRPS weights the integrand (= Brier score for threshold z)

$$\text{twCRPS}(F, y) = \int_{-\infty}^{\infty} (F(z) - \mathbb{1}\{y \leq z\})^2 w(z) dz$$

$w(z)$ is a weight function. The score twCRPS is proper and avoids the problem with looking only at a sample of extreme outcomes (Alice and Bob's example).

Gneiting, T. and Ranjan, R. (2011)

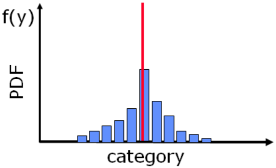
(adapted from a slide by Sebastian Lerch)

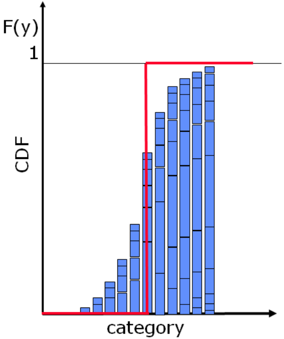
Ranked Probability Score (RPS)

- The RPS is a discrete analog to the integral over Brier scores $\mathrm{CRPS} = \int \mathrm{BS}_x \, \mathrm{d}x$

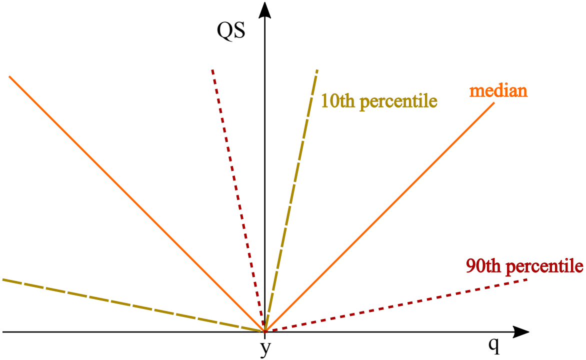
$$\mathrm{RPS} = \sum_{k=1}^{L} \mathrm{BS}_{x_k} = \sum_{k=1}^{L} \left(P_{\mathrm{fc}}(k) - P_y(k)\right)^2$$

- The thresholds x_k that separate the L categories can be chosen in various ways
  - equidistant (RPS $\rightarrow$ CRPS as $\Delta x \rightarrow 0$)
  - climatologically equally likely, e.g. tercile boundaries





Quantile score



$$\mathrm{QS}_\alpha(q,y) = 2\left(\mathbb{I}\{y < q\} - \alpha\right)(q - y)$$

where q, y and α denote the quantile, the observation and the probability level, respectively. The indicator function $\mathbb{I}$ returns 1 if its argument is true and 0 otherwise. For the median ($\alpha = 0.5$), the quantile score becomes symmetric with respect to $q - y$ and is equal to the mean absolute error.

$$\int_0^1 \mathrm{QS}_\alpha \, d\alpha = \mathrm{CRPS}$$

Logarithmic score

Ignorance score

- For a forecast consisting of a probability density $p_{\mathrm{fc}}(x)$, define

$$\mathrm{LS} = -\log(p_{\mathrm{fc}}(y))$$

  where y denotes the observation (or analysis).
- This score is proper and local.
- ensemble forecasts $\longrightarrow$ probability density

Logarithmic score

Ignorance score

- For a forecast consisting of a probability density $p_{\mathrm{fc}}(x)$, define

$$\mathrm{LS} = -\log(p_{\mathrm{fc}}(y))$$

  where y denotes the observation (or analysis).
- This score is proper and local.
- ensemble forecasts $\longrightarrow$ probability density
- A simple yet useful exercise is to use the Gaussian density given by the ensemble mean μ and the ensemble variance σ^2. Then, the logarithmic score is given by

$$\mathrm{LS} = \frac{(\mu - y)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$$

ECMWF

Logarithmic score

Ignorance score

- For a forecast consisting of a probability density $p_{\mathrm{fc}}(x)$, define

$$\mathrm{LS} = -\log(p_{\mathrm{fc}}(y))$$

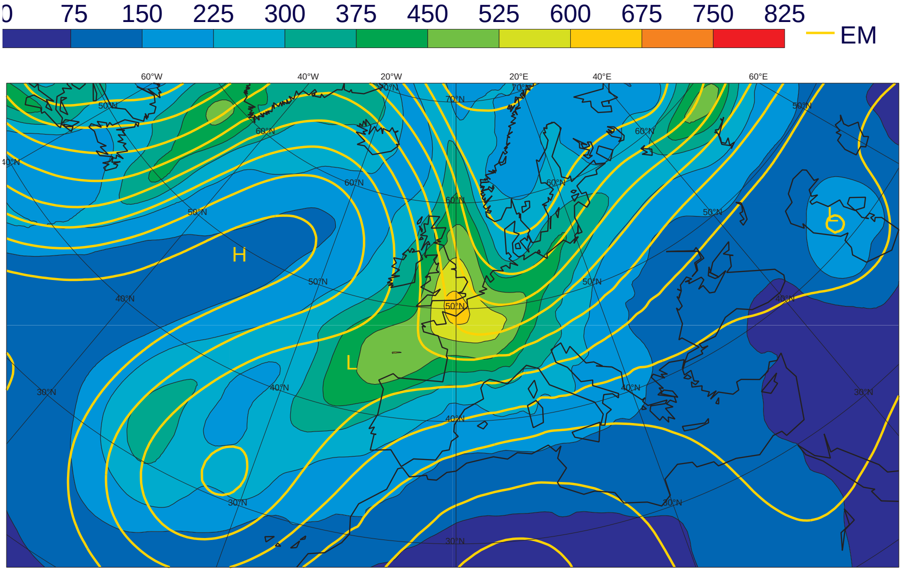
  where y denotes the observation (or analysis).
- This score is proper and local.
- ensemble forecasts $\longrightarrow$ probability density
- A simple yet useful exercise is to use the Gaussian density given by the ensemble mean μ and the ensemble variance σ^2. Then, the logarithmic score is given by

$$\mathrm{LS} = \frac{(\mu - y)^2}{2\sigma^2} + \tfrac{1}{2}\log(2\pi\sigma^2)$$

- The first term is a measure of the reliability and the second term is a measure of the sharpness of the forecast.
- This is a score in itself (Gneiting and Raftery, 2007, refer to this Dawid-Sebastiani two-moment score) and Siegert et al. (2019) have developed a fair version of this score.

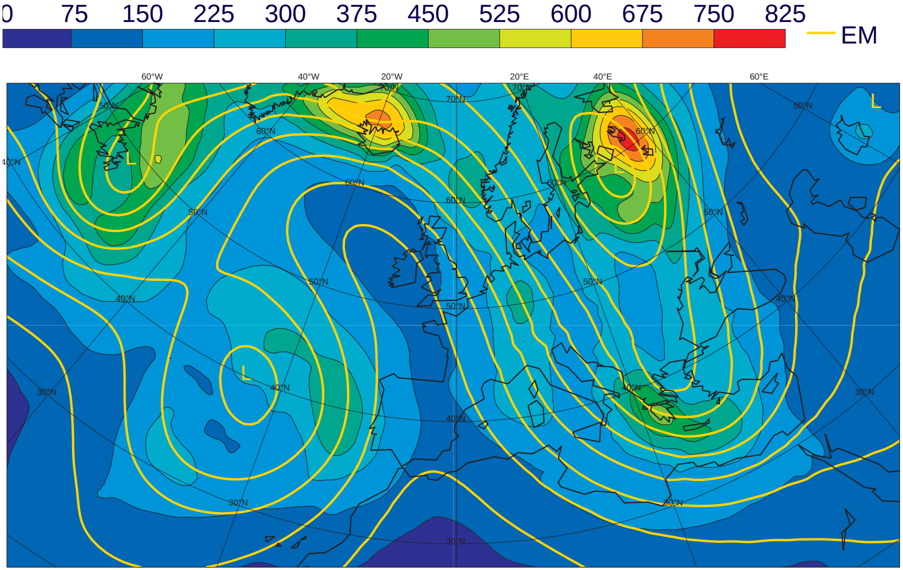
Daily EPS stdev (shaded) and ens. mean (cont.)

500 hPa geopotential ($m^2\,s^{-2}$) at 72 h lead; init. time 6 December 2010

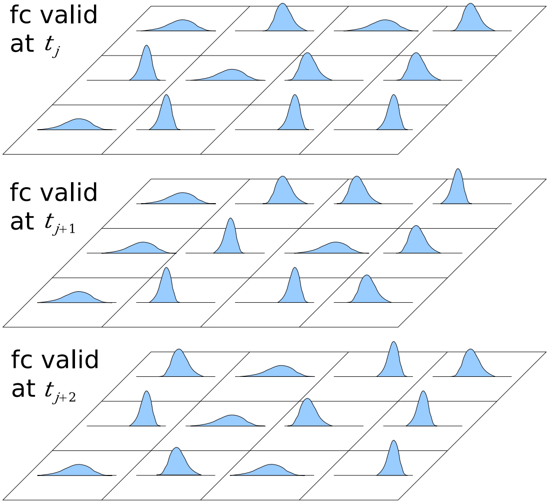


Daily EPS stdev (shaded) and ens. mean (cont.)

500 hPa geopotential ($m^2 s^{-2}$) at 72 h lead; init. time 8 December 2010



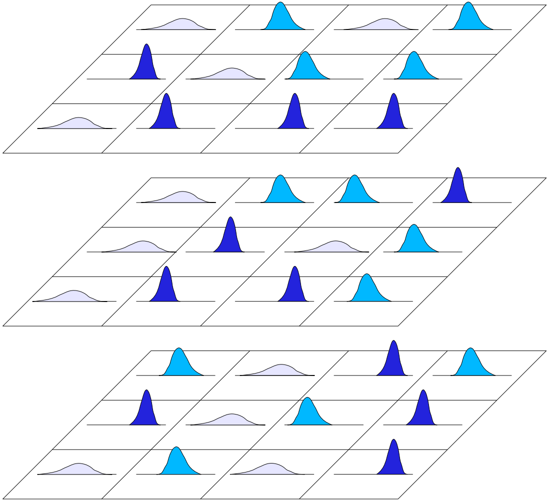
consider (local) pairs of ensemble variance and squared error of the ensemble mean



Spread-reliability

methodology

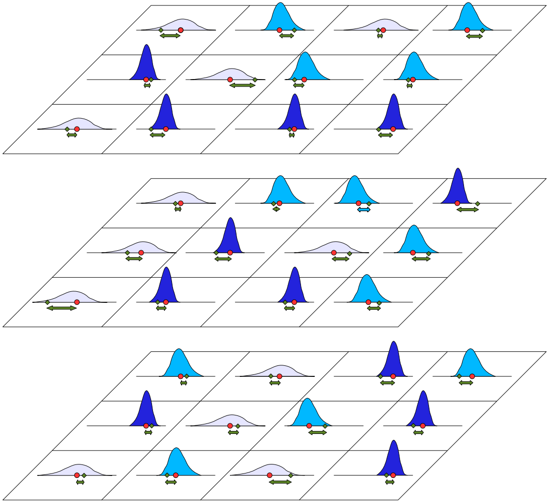
consider (local) pairs of ensemble variance and squared error of the ensemble mean



Spread-reliability

methodology

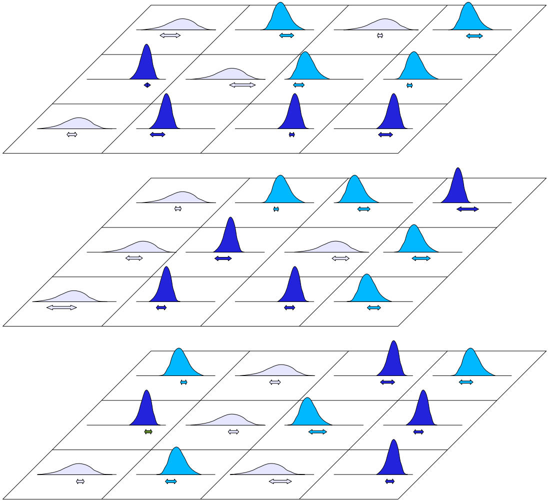
consider (local) pairs of ensemble variance and squared error of the ensemble mean



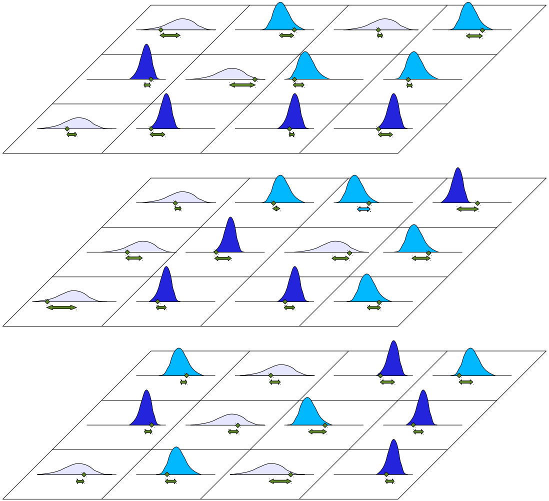
Spread-reliability

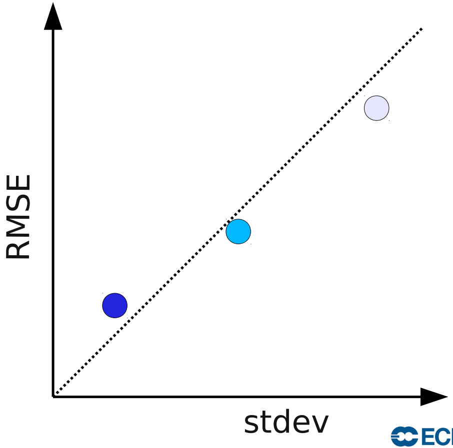
methodology

consider (local) pairs of ensemble variance and squared error of the ensemble mean — stratified by the ensemble variance



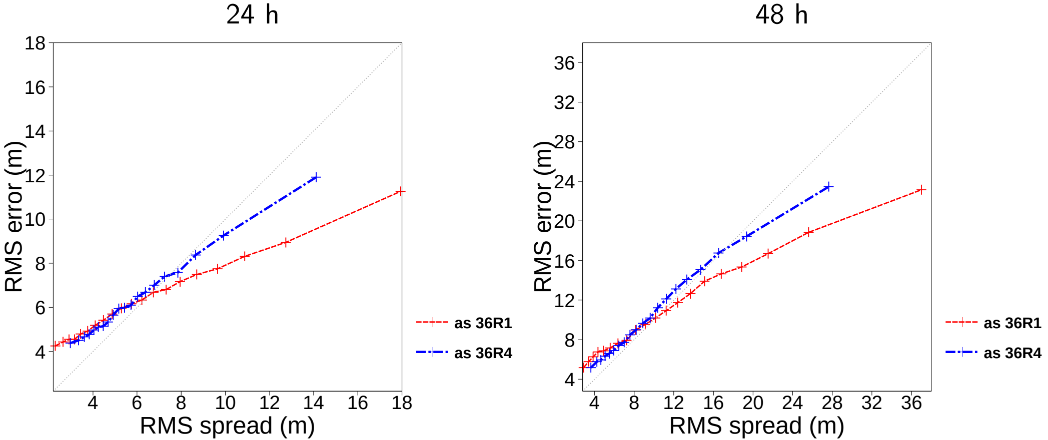
consider (local) pairs of ensemble variance and squared error of the ensemble mean — stratified by the ensemble variance





Spread-reliability: An example

500 hPa height — 20°–90° N



- 40 cases
- T639, 50 member
- Jan 2010 config. ("as 36r1")

- Nov 2010 config. ("as 36r4"): revised initial perturbations and revised tendency pertns.

Uncertainty of the verifying observations

or, more generally, the verifying data

- In real applications the true state x_t of the atmosphere is not know exactly. The observation y has an error

$$y = x_t + \epsilon$$

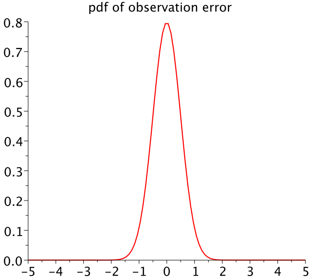
- Assume an ensemble is perfectly reliable, i.e. ensemble members $x_e \sim \rho_e$ and the true state $x_t \sim \rho_t$ are realisations of the same distribution $\rho_e = \rho_t$.
- Then, the observation y is a realisation of the distribution given by the convolution of the true distribution and the error distribution

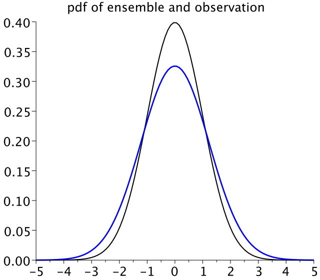
$$\rho_y = \rho_t * \rho_\epsilon$$

$$\rho_y(x) = \int_{-\infty}^{+\infty} \rho_t(z)\rho_\epsilon(x-z)\,dz$$

- Thus, a verification with respect to y will indicate a lack of reliability.

Verification in the presence of observation uncertainties





ρ_ϵ $\qquad$ $\rho_t = \rho_e, \qquad \rho_y = \rho_E$

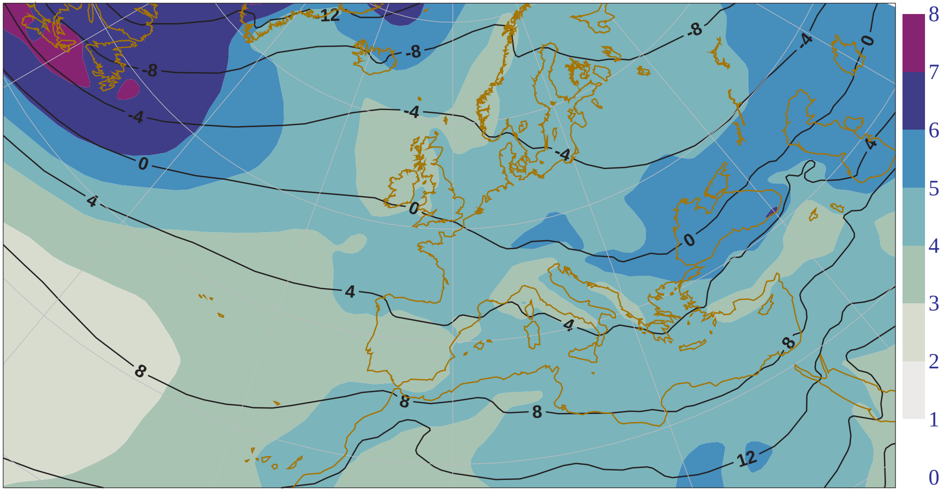
- A solution: postprocess ensemble members prior to verification
- Verify ensemble members to which noise has been added:
  $x_E = x_e + \epsilon$ with $\epsilon \sim \rho_\epsilon$
- Then, we have $\rho_E = \rho_y$

see Saetra et al. (2004) and see e.g. Ben Bouallègue (2020) Tech. Memo 865, www.ecmwf.int/node/19544

The climatological distribution

temperature at 850 hPa

15 March (based on ERA-Interim 1989–2008)

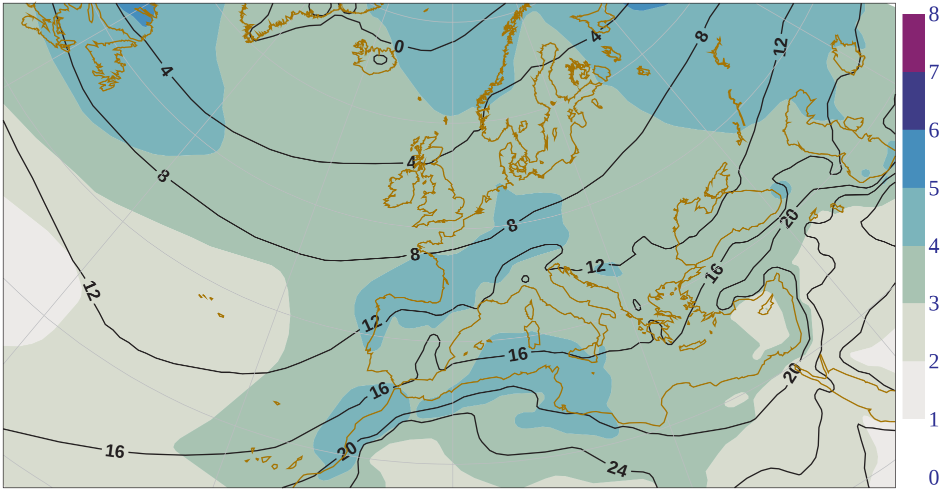


contours: mean — shading: stdev

The climatological distribution

temperature at 850 hPa

15 June (based on ERA-Interim 1989–2008)



contours: mean — shading: stdev

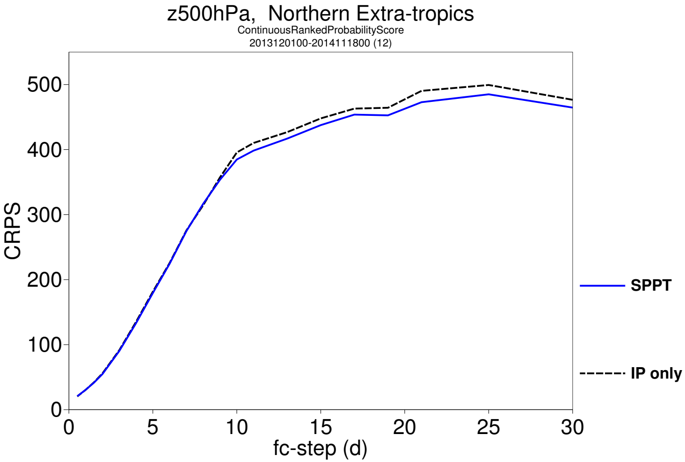
Ficticious skill due to a poor climatological distribution

- If one uses the same climatological distribution for a domain with different climatological characteristics (mean, stdev, . . .), the skill with respect to that distribution is not real skill. It reflects the poor quality of the climatological distribution.
- Same applies if seasonal variations of the climatological distribution are not represented.
- This criticism applies for instance if the climatological distribution is derived from the verification sample itself by aggregating different start times and different locations.

Ficticious skill due to a poor climatological distribution

- If one uses the same climatological distribution for a domain with different climatological characteristics (mean, stdev, ...), the skill with respect to that distribution is not real skill. It reflects the poor quality of the climatological distribution.
- Same applies if seasonal variations of the climatological distribution are not represented.
- This criticism applies for instance if the climatological distribution is derived from the verification sample itself by aggregating different start times and different locations.
- It can also be misleading to compare skill scores from different prediction centres when the skill scores have been computed against own analyses.
- If the same climatological distribution (say ERA-Interim) is used as reference, this climatological distribution has the lowest skill when verified against the analysis that deviates most from the analyses used for computing the climatological distribution.

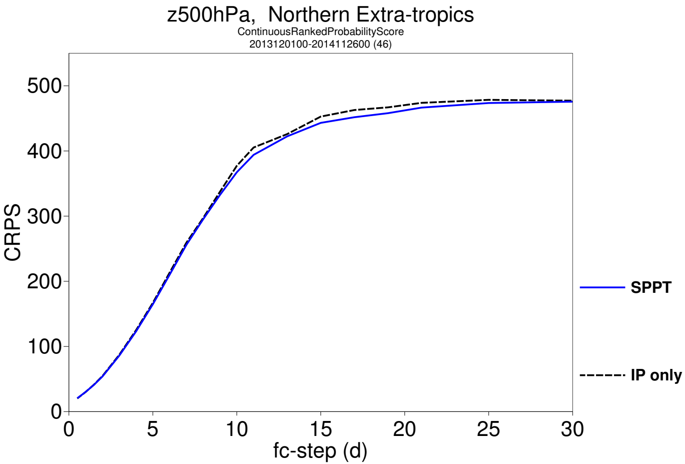
Comparing model versions/ numerical experiments



- 12 cases (1 year, every 32 days)
- Could difference in score be a result of chance?
- How large does a difference have to be to be trusted?

- case-to-case variability of predictability implies distribution of score for given lead time is fairly wide
- ⇒ not easy to get enough cases to distinguish score distributions of two numerical experiments

Comparing model versions/ numerical experiments

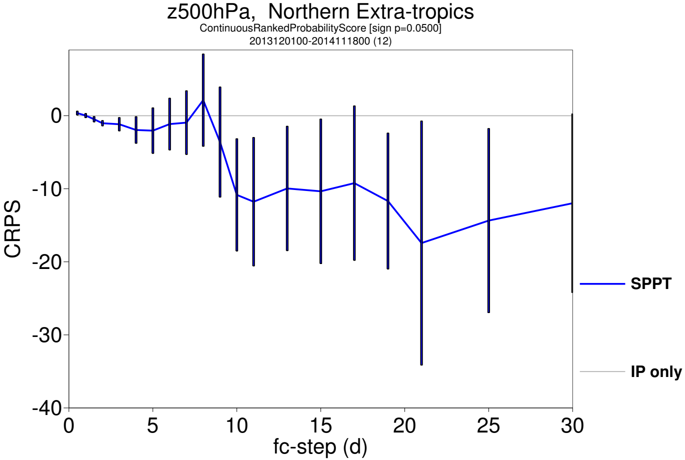


- 46 cases (1 year, every 8 days)
- Could difference in score be a result of chance?
- How large does a difference have to be to be trusted?

- case-to-case variability of predictability implies distribution of score for given lead time is fairly wide
- ⇒ not easy to get enough cases to distinguish score distributions of two numerical experiments

95% confidence intervals

Paired sample of cases: t test applied to score differences



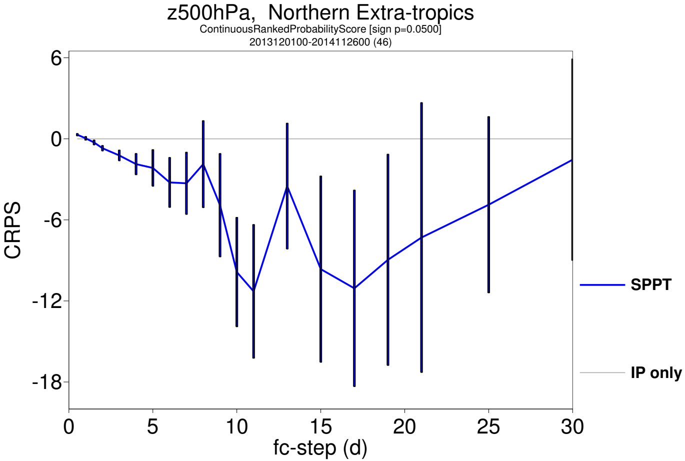
- 12 cases (1 year, every 32 days)
- Variability of score differences is much smaller!
- $\Rightarrow$ Paired sample of cases (start dates)
- For each forecast lead time, consider sample of score *differences*

- Temporal auto-correlation taken into account using AR(1) model when estimating variance of mean difference
- see also Diebold and Mariano (1995) and Geer (2016)

ECMWF

95% confidence intervals

Paired sample of cases: t test applied to score differences



- 46 cases (1 year, every 8 days)
- Variability of score differences is much smaller!
- ⇒ Paired sample of cases (start dates)
- For each forecast lead time, consider sample of score *differences*

- Temporal auto-correlation taken into account using AR(1) model when estimating variance of mean difference
- see also Diebold and Mariano (1995) and Geer (2016)

More verification topics

- sensitivity to ensemble size M and estimation of verification statistics in the limit $M \to \infty$: fair scores, see e.g. Ferro et al. (2008); Leutbecher (2019); Siegert et al. (2019) and the Lorenz 1996 practical
- skill on different spatial scales, see Jung and Leutbecher (2008)
- multivariate aspects (e.g. energy score, slide 11)
- forecast user perspective, decision making and verification
  - yes/no decisions and the cost-loss model, see Richardson (2000)
  - weather roulette, see Hagedorn and Smith (2009)
  - elementary scores are building blocks for many proper scores; with different weighting functions, one obtains many different scores such as the CRPS or the diagonal score, see Ben Bouallègue et al. (2018)

References I

Ben Bouallègue, Z., T. Haiden, and D. S. Richardson, 2018: The diagonal score: Definition, properties, and interpretations. *Q. J. R. Meteor. Soc.*, **144**(714), 1463–1473.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteor. Soc.*, **131**, 2131–2150.

Diebold, F. X. and R. S. Mariano, 1995: Comparing predictive accuracy. *Journal of Business & Economic Statistics*, **13**(3), 253–263.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteor. Appl.*, **15**, 19–24.

Geer, A. J., 2016: Significance of changes in medium-range forecast scores. *Tellus A*, **68**(1), 30229.

Gneiting, T. and A. E. Raftery, 2007: Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

Gneiting, T. and R. Ranjan, 2011: Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, **29**(3), 411–422.

Hagedorn, R. and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.*, **16**, 143–155.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

References II

Hamill, T. M. and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Q. J. R. Meteor. Soc.*, **132**, 2905–2923.

Hersbach, H., 2000: Decomposition of the continous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559–570.

Jung, T. and M. Leutbecher, 2008: Scale-dependent verification of ensemble forecasts. *Q. J. R. Meteor. Soc.*, **134**, 973–984.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Statistical Science*, **32**(1), 106–127.

Leutbecher, M., 2009: Diagnosis of ensemble forecasting systems. In *Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, ECMWF, Reading, UK, 235–266.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Q. J. R. Meteor. Soc.*, **145**(S1), 107–128.

Leutbecher, M. and T. Haiden, 2021: Understanding changes of the continuous ranked probability score using a homogeneous gaussian approximation. *Quarterly Journal of the Royal Meteorological Society*, **147**(734), 425–442.

Murphy, A. and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

References III

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteor. Soc.*, **126**, 649–667.

Roulston, M. S. and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

Saetra, Ø., H. Hersbach, J.-R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501.

Siegert, S., C. A. T. Ferro, D. B. Stephenson, and M. Leutbecher, 2019: The ensemble-adjusted ignorance score for forecasts issued as normal distributions. *Q. J. R. Meteor. Soc.*, **145**(S1), 129–139.

Weigel, A. P., 2011: Ensemble verification. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe, I. T. and Stephenson, D. B., editors. Wiley, 2nd edition.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 3rd edition.