SANGOMA: Stochastic Assimilation for the Next Generation Ocean Model Applications EU FP7 SPACE-2011-1 project 283580

Deliverable 4.3: How to use the probabilistic metrics on small & medium benchmarks ?

Due date: 31/03/2014

Delivery date: 18/04/2014

Delivery type: Report , public

Jean-Marie Beckers   Alexander Barth
University of Liège, BELGIUM

Peter Jan Van Leeuwen
University of Reading, UK

Lars Nerger
Alfred-Wegener-Institut, GERMANY

Arnold Heemink   Nils van Velzen
Martin Verlaan
Delft University of Technology, NETHERLANDS

Pierre Brasseur   Jean-Michel Brankart   Guillem Candille
Sammy Metref   Florent Garnier
CNRS-LGGE, FRANCE

Pierre de Mey
CNRS-LEGOS, FRANCE

Laurent Bertino
NERSC, NORWAY

April 17, 2014

Chapter 1

Introduction

The main purpose of SANGOMA is to advance the status of probabilistic assimilation methods in oceanography. The probabilistic approach means that the uncertainty associated with the estimated ocean flow is available. The challenge is then to qualify and quantify this uncertainty in order to get the most accurate information as possible about the state of the ocean flow.

In the previous WP4 report (4.1), the benchmarks methodolgy has been fully described. Also, probabibilistic metrics has been briefly introduced. The goal of this document is to provide a manual to perform probabilistic verification in practical cases: for the small and medium benchmarks. First, chapter 2 reminds the basic concepts of the probabilistic validation: the main probabilistic attributes and the theoretical description of the probabilistic metrics -commonly called scores- measuring these attributes. Then, chapter 3 shows how to specifically perform the probabilisitc diagnoses in the small and medium benchmarks situations.

Chapter 2

Probabilistic validation: theoretical considerations

This chapter first presents the main attributes on which the probabilisitc systems are evaluated. The following sections remind the theoretical properties of probabilisitc scores (see chaper 4 in report 4.1): the Rank Histogram & the RCRV, the CRPS and finally the Brier score & the Entropy.

2.1 Probabilistic attributes

How can one objectively evaluate the quality of a probabilistic system? Let us consider the following statement produced by the probabilistic system: ‘there is 30% probability that the Northern Sea Route is free of ice’. Assuming the event ‘free of ice’ is unambiguously defined, neither its observed occurrence nor its non-occurrence can be legitimately used to validate or unvalidate the produced ensemble. Contrary to a deterministic system, the validation of a probabilistic system cannot be performed over a single case (or realisation). One must use a statistical approach, based on a sufficiently large set of realisations, Meaning this validation requires an aggregation of a large set of independent realisations of the considered process. After accumulating independent realisations of the probabilistic system, two probabilistic properties (attributes) can be measured: the *reliability* and the *resolution* (*e.g.* Toth et al. 2003).

In the example cited above, one has to wait until the 30% probability is produced by the system a number of times. Then one can first check the proportion of actual observed occurrence of ‘free of ice’. If that proportion is equal or close to 30%, the probabilistic system can be considered as statistically consistent. If, on the contrary, that proportion is significantly different from 30%, the system is statistically inconsistent. One condition for the validy of a probabilistic system is therefore the statistical consistency between produced probabilities and observed frequencies of occurrence of the event under consideration. This property of statistical consistency is called *reliability*.

More generally, the reliability is the system ability in producing Probability Density Functions (pdf) in agreement with the observed pdf, *i.e.* distribution of

the observed variable when a given pdf is produced. Actually, one considers a system producing the pdf F_p with the frequency g_p. Then, one aggregates the observed variables when the system produces F_p in order to define the associated observed pdf F'_p. Thus, F_p can be directly compared to F'_p. The system is reliable, or statistically consistent, if and only if for each class p of produced pdf one has $F_p = F'_p$. Several scores are used to evaluate this property, and some of them are presented in the following sections: the Rank histogram & the RCRV, but also the reliability component of the CRPS & the Brier score.
The reliability attribute is a necessary condition to have a skillfull probabilistic system, but it is not a sufficient property. Actually, every system can be *a posteriori* calibrated. It can be transformed into a reliable system by replacing F_p by F'_p for each p over a given verification set, and by applying this correction to the pdfs produced by the system over the subsequent verification set (under stationnary assumption of the system). Also, if one knows the climatological distribution F_c of the observed variables over a given verification dataset, a system producing this climatological distribution for each realisation would be obviously reliable ... but it would provide no other usefull information than the climatology (no need to integrate a complex numerical model to get this result). For instance, one knows the climatological frequency when the Northern Sea Route is free of ice is 2 months a year (occurrence $\approx 16\%$). If a probabilistic system produce the 16% probability every day, it is reliable if one evaluates its performance over a year, but it cannot provide any information about the seasonal (for instance) variability of that probability of occurrence. In other words, a climatological system would be perfectly reliable without providing any additional useful information. To determine if one has a skillful probabilisitic system, another attribute is then needed.

The *resolution* is the system ability in discriminating the disctinct observed situations; this property is closely related to the information content and the entropy (*e.g.* Roulston and Smith 2002). If the system is reliable, the resolution is also referred as the *sharpness* which measures the dispersion of *a priori* produced pdfs (or probabilities). Resuming the notations from the previous paragraph, the resolution can be seen as the dispersion of *a posteriori* observed pdf F'_p. The sharper the pdfs F'_p compared to F_c are, the better the resolution is. In other words, the resolution is the additional information, compared to the climatology, that can be potentially extracted from the probabilistic system.

A skillful probabilistic system must then satisfy two criteria:

- to be reliable, *i.e.* $F_p = F'_p$ for any p
- F_p are as sharp as possible compared to F_c

The following sections introduce some commonly used probabilistic scores (for univariate situation) and mostly remind the Appendix A from report 4.1.

2.2 Rank Histogram & RCRV

The probabilistic scores introduced here only verify the reliability of the probabilistic system by checking the indistinguishability between the observed variables and the produced pdfs. More specifically, an ensemble of N members and one observation are indistinguishable if one can consider they are independent draws from the same random variable.

Rank Histogram (RH) Let us consider the produced pdf is represented by an ensemble with size N. For each of M realisations of the system, the N ensemble members are ranked in increasing order, thereby defining $N+1$ intervals (or bins). If the verification is statistically indistinguishable from the ensemble, it must fall with equal probability in each of those $N+1$ intervals and then shows by accumulation a flat RH. The shape of the RH caracterizes the lack of reliability of the system. For instance, a strong U-shape with overpopulated outliers shows an underdispersive system. The deviation from the flatness, considering the finitness of M, is numerically measured by the following quantity

$$\delta = \frac{M(N+1)}{N} \sum_{i=1}^{N} \left(\frac{s_i}{M} - \frac{1}{N+1} \right)^2 \tag{2.1}$$

where s_i is the observed population of the i-th interval. For a reliable system, considering the finiteness of M, the expected value of δ must be equal to 1. Note that δ is strongly M-dependent and increases with increasing M once the system is (even slightly) non reliable ($\delta > 1$). In order to avoid this numerical effect, it is suitable to normalize this score by M, or to use the RCRV for the reliability characterization of a probabilisitc system.

Reduced Centered Random Variable (RCRV) For each realisation of the system, the following variable is defined

$$y = \frac{o - m}{\sigma} \tag{2.2}$$

where m and σ are the mean and the standard deviation of the produced pdf respectively and o the observed value. Note that the observation error σ_o can be simply introduced in y by considering $\sigma = \sqrt{\sigma_{ens}^2 + \sigma_o^2}$. The system is reliable, if the mean of y over all the M realisation of the probabilistic system is null and its standard deviation is equal to 1. Thus, the reliability is decomposed into (normalized) bias $b = E[y]$ and dispersion $d^2 = E[y^2] - b^2$. Here, the U-shape RH is equivalent to $d > 1$ for an underdispersive system.
Note that the RCRV bias/dispersion can be related to the RMSE at the first order approximation by

$$\text{RMSE}^2 \approx E[\sigma^2](b^2 + d^2) \tag{2.3}$$

2.3 Continuous Ranked Probability Score (CRPS)

The CRPS measures the global skill of a probabilistic system by evaluating both reliability and resolution. It is based on the the square difference between the produced Cumulative Distribution Functions (cdf) of a univariate variable x and the corresponding cdf of the observation:

$$\text{CRPS} = E\left[\int_{\mathbb{R}} (F_p(x) - F_o(x))^2 \, dx\right] \tag{2.4}$$

where F_p is the cdf associated with the produced pdf and F_o the cdf associated with the observation (a simple Heaviside distribution when no observation error is considered). Contrary to the reliability scores presented above, the CRPS has the dimension of the evaluated variables (for instance expressed in meters for the SSH, in Kelvin for the temperature, etc...). The CRPS can be decomposed into the reliability/resolution parts in the same way as the Brier score (see next section). But for practical and numerical considerations (see Candille and Talagrand 2005), the decomposition described by H. Hersbach (Hersbach 2000) is chosen. This decomposition is based on the same principle as the rank histogram construction. The reliability part ($Reli$) is null for a reliable system and the resolution part ($Resol$) goes from 0 for a perfect deterministic system to $Unc = \int_{\mathbb{R}} F_c(x)\,(1 - F_c(x))\,dx$ for a useless and non informative system (F_c is the climatological cdf associated with the verification data set).
Evaluated through the CRPS, a skillful probabilistic system must satisfy two criteria:

- $Reli = 0$
- $Resol \ll Unc$
- and $CRPS = Reli + Resol$

2.4 Brier score & Entropy

In this section, the probability of occurrence of a binary event is considered. The presented scores are then computed in probability space.

Brier score The Brier score is a restriction of the CRPS to the probability space. It measures the global skill of a probabililistic system by evaluating both reliability and resolution criteria. For simplifications, we denote p the probability $p(x)$ of occurrence of an event related to the state vector x. The Brier score is then written

$$\mathcal{B} = E[(p - o)^2] \tag{2.5}$$

where o is the probability associated with the observed occurrence of the considered event. By definning $p' = E_p[o]$, the probability of the observed occurrence of the event when p is produced, and $p_c = E[o] = E[p']$ the climatological probability of occurrence associated with the whole observation data set, the Brier score can be decomposed into reliability and resolution parts (Murphy 1973):

$$\mathcal{B} = \underbrace{E[(p - p')^2]}_{reliability} + \underbrace{p_c(1 - p_c) - E[p' - p_c)^2]}_{resolution} \tag{2.6}$$

The term $p_c(1-p_c)$ is also called *uncertainty* and only depends on the the observation data set. A skill score can be thus defined by

$$\mathcal{B}_s = 1 - \frac{\mathcal{B}}{p_c(1-p_c)} \tag{2.7}$$

The reliability part is null for a perfectly reliable system. The resolution part goes from 0 for a perfect deterministic system to 1 for a useless system, *i.e.* a system providing no more information than the climatology p_c.
Similarly to the CRPS, a system evaluated through the Brier score is skillful for a given event if:

- $p = p'$ for any p
- $resolution \ll uncertainty$ or $resolution_{skill} \ll 1$

Entropy The Entropy only measures the information content of the system which is closely related to the resolution as evaluated through the Brier score. Considering the notations above, the entropy is written

$$\mathcal{S} = -E[p' \ln p'] \tag{2.8}$$

This score goes to 0 for a perfect deterministic system to $p_c \ln p_c$ for a useless system, *i.e.* a system providing no more information than the climatology p_c. Note that this score does not provide any diagnosis on the reliability of the probabilistic system.

2.5 Some realistic examples

In this section, the behavior of some probabilistic scores is shown. The characterisation of the outputs ensemble systems are fully under control through synthetic data built from (log-)normal distributions:

- M=100000 number of realisations of the ensemble process
- N=50 ensemble size
- ensemble means: $m \in \mathcal{N}(0, S)$
  with $S = 1$ *climatological* spread
- ensemble standard deviations: $\sigma \in Log\mathcal{N}(s, 0.05)$
  with sharpness $s = 20\%.S$
- 1 observation $o \in \mathcal{N}(m, \sigma)$
- N=50 ensemble members $x \in \mathcal{N}(m - \alpha.s, \sigma/\beta)$
  (α and β characterise the bias and the dispersion respectively)

Figure 2.1 shows the rank histograms (RH and the δ score), the bias/dispersion from the RCRV and the $CRPS = Reli + Resol$ for different values α and β. 9 ensembles systems are simulated with 3 different biases α: no bias, 50% negative bias (the ensemble underestimates the observation) and 100% negative bias. Also 3 kinds of dispersive systems are simulated: twice overdispersive systems ($\beta = 1/2$), perfect dispersive systems ($\beta = 1$) and twice underdispersive systems ($\beta = 2$).

Note that the uncertainty computed from the CRPS is equal to 0.57 for all the experiments. When the ensemble spreads are not degraded compared to the observations ($\beta = 1$), the ratio $Resol/Unc$ is equal to 19.3% (very close to the expected sharpness parameter 20%).

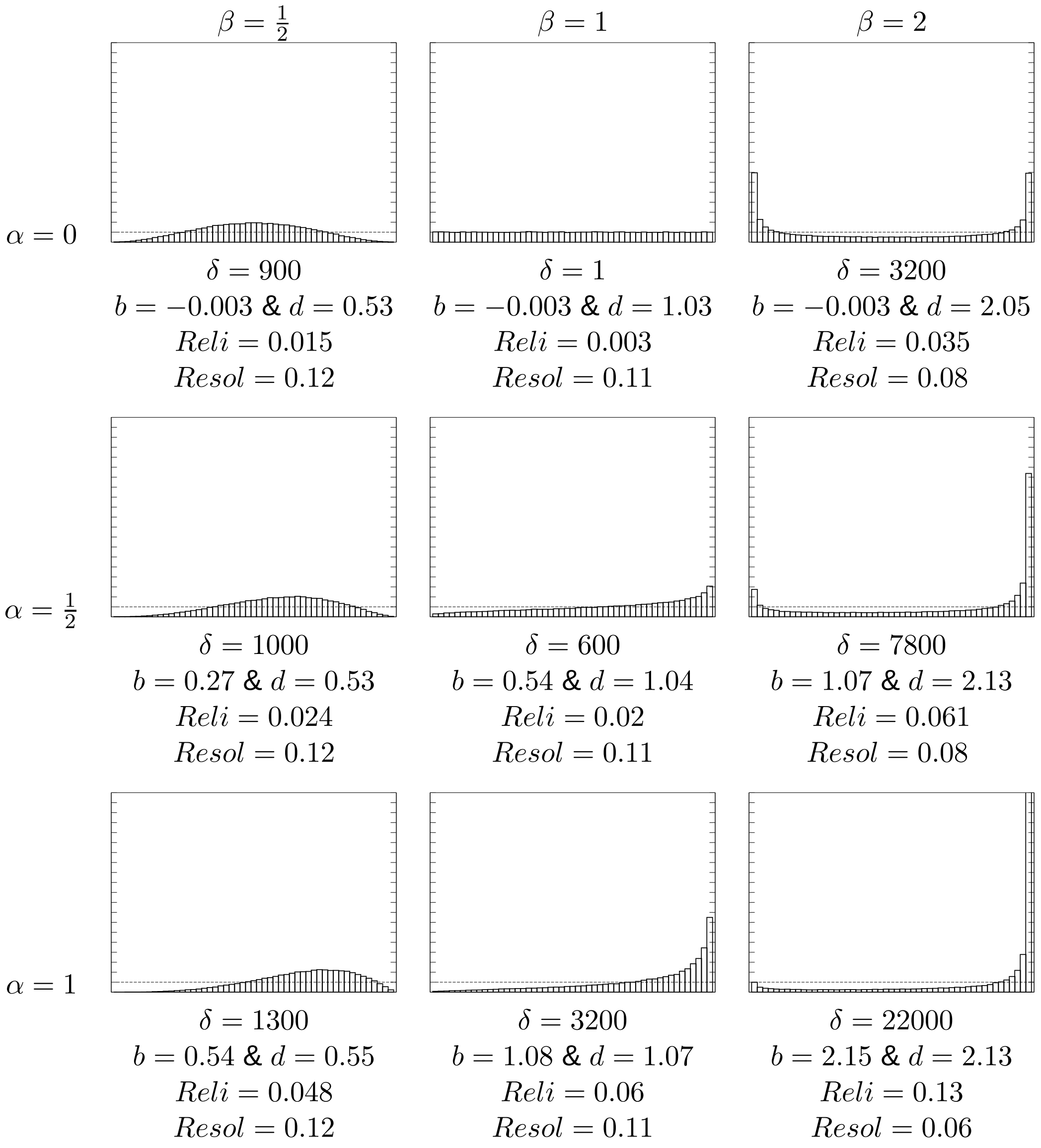

Figure 2.1: Probabilistic diagnoses for synthetic ensembles with different parameters α and β; uncertainty related to the CRPS computation: $Unc = 0.57$.

Chapter 3

Practical verification

This chapter provides methodologies for evaluating -in practice- the probabilistic assimilation systems developped on both small and medium benchmarks through the scores described in the previous chapter. Note that the presented methods are not unique but can be seen as recommandations for practical verification processes.

As mentioned in the previous chapter, the probabilisitc verification is based on statistical accumulation of independent realisations of the evaluated system. The challenge is then -in practical situations- to aggregate a large enough verification dataset in order to get statistically significant diagnoses. For instance, to validate an ensemble with size around $N = 50 - 100$, a reasonable verification dataset should be at least around $M = 5000$ (actually, the larger the better).

3.1 Small case benchmark

The small benchmark is based on the Lorenz-96 model with 40 variables (see report 4.1 for details). The experiments based on this model are highly reproducible. It is then feasible to perform several totally independent experiments for each data assimilation tested method. Of course, to perform $M = 5000$ independent experiments is out of bound, but produced ensembles and their associated observations can be aggregated over different lead times from the same experiment (for instance 1 for each assimilation cycle), or several variables of the model can also be aggragated/gathered. Nevertheless, remind that the diagnoses are univariate. Whatever the way to obtain the verification dataset, it is composed with M couples ensembles/observations, denoted by (x, o) (for the SANGOMA-standard format, see report 3.1). Note that different relevant datasets (different periods or different sets of variables) could be necessary to get a complete evaluation of a probabilistic system. This generally depends on the users' needs.
Remarks: in this chapter, the term ‘observation’ means the value used to perform the verification; this could be either observed or non-observed data as named in the common assimilation terminology.

All the routines mentioned below are listed and documented in report 2.4.

- How to compute the Rank Histogram for (x, o) ?
  - use sangoma_ComputeHistogram.F90 with direct input (x, o).
  - outputs: distribution of the aggregated observations over the ensemble bins (Rank Histogram) and the δ score.

  – the RH provides a graphical diagnosis of the reliability: flat histogram (perfectly reliable), U-shape (underdispersive), bell-shape (overdispersive), asymetry (bias).

  – δ provides the numerical distance between the RH and the perfectly flat histogram; this shows the lack of reliability, but with no information about the characteristic of the non-reliability.

- How to compute the RCRV for (x, o) ?
  - use sangoma_ComputeRCRV.F90 with direct input (x, o).
  - outputs: numerical values for the normalized bias and the dispersion of the ensembles compared to the observations.

  – bias b: expressed in percentage of the 'uncertainty' related to the evaluated system (combination of ensemble spread and observation error); for instance $b = 0$ indicates a no-bias system and $b = 1$ shows a negative bias ($o - m$) with the same order of magnitude of the system 'uncertainty'.

  – dispersion d: expressed in percentage of the system 'uncertainty'; for instance $d = 1$ means the ensembles are perfectly reliable, $d = 2$ means the ensembles are twice underdispersive and $d = 0.5$ means the ensembles are twice overdispersive.

- How to compute the CRPS for (x, o) ?
  - use sangoma_ComputeCRPS.F90 with direct input (x, o).
  - outputs: numerical values for CRPS, $Reli$, $Resol$ and Unc.

  – *uncertainty* Unc: CRPS value if only computed with the observations from the verification dataset; independent from the produced ensembles, Unc then only depends on the observations; a large value of Unc shows the large variability of the evaluated quantity over the chosen verification dataset (long verification period, very different aggregated variables); since the ability of an ensemble system is mostly based on the comparison with the *uncertainty* Unc, the definition of the verification dataset is then crucial in interpreting the skill of the system (but this is also true for all statistical diagnoses).

  – reliability $Reli$: this value is null for a perfectly reliable system; and $Reli > 0$ shows lack of reliability of the system.

  – resolution $Resol$: this value is null for a perfectly accurate deterministic system; $Resol > 0$ must be compared to Unc and the system is considered as usefull when $Resol \ll Unc$; as already mentioned, the resolution is closely related to the sharpness, so that $Resol \ll Unc$ implies $\overline{\sigma} \ll S$, where $\overline{\sigma}$ is the mean of the produced ensemble spreads and S is the spread of the observations from the verification dataset (or climatological spread).

  - CRPS: this is the global score $CRPS = Reli + Resol$ and summarize the global skill of the system taking into account both reliability and resolution; CRPS is null for a perfectly reliable deterministic system.

- How to compute the Brier score & the Entropy for (x, o) ?
  - define the vector threshold x_{th} for the binary event: for each evaluated variable x, the easiest choice is to apply the same single value th -based on the users' knowledge and needs- to all the realisations of the system; but one can also apply different values th depending on relevant subsets from the verification dataset.
  - use sangoma_ComputeBRIER.F90 with direct input (x, o) and x_{th}.
  - outputs: Brier skill score + reliability/resolution decomposition, *uncertainty*, Entropy.
  Remark: since the Brier score is a reduction of the CRPS to the probability space, all the general comments about the CRPS remain valid for the Brier score.

  - *uncertainty* Unc: $Unc = p_c(1 - p_c)$ where p_c is the climatological probabililty of occurrence of the event (over the verification dataset); this is the reference used to normalize the Brier score and defined the skill scores; the definition of the verification dataset is then crucial in interpreting the skill of the system.
  - Brier skill score $\mathcal{B}_s$: this score is equal to 1 for a perfectly reliable deterministic system and is null for a climatological system, *i.e.* a system always producing p_c; negative values of this scores indicates poorer informative systems than the climatology.
  - reliability: this value is null for a perfectly reliable system.
  - resolution: this value is null for a perfectly accurate deterministic system and is equal to 1 for a climatological system.
  - Entropy: an informative system gives $0 < \mathcal{S} \ll -p_c \log p_c$.

3.2 Medium case benchmark

The medium case benchmark is based on an idealized configuration of the NEMO primitive equation ocean model: a square and 5000-meter deep flat bottom ocean at mid latitudes (the so called square-box SQB configuration, see details in report 4.1). In this case, the practical computation of the diagnoses does not really differ from the computation presented in the previous section (section 3.1). Nevertheless, there is one crucial difference: the experiments are no longer so easily reproducible so that the realisations of the probabibilistic system must be aggregated over the grid points (3D) of the model (or over the simulated observation locations). The balance must then be done between the size of the verification dataset and the relevance of temporal/spatial distributions of the aggregated grid points. While thousands grid points are available at each time step of the model, time series of each numerical score can be computed (for instance 1 diagnosis per assimilation cycle).

For the reliability scores (RH and RCRV), their practical computations are similar to the ones presented in previous section 3.1. For the other scores (CRPS, Brier and Entropy), the raw comparisons over the whole SQB-domain with the *uncertainty* and the climatology could be meaningless and the skill of the probabilistic system overestimated. Actually, when the the climatology is computed over all the grid points of the SQB-domain and over a long period, its spread is very large and the *uncertainty* is too. This reflects a global *uncertainty*. On the other hand, the probabilistic system provides local ensembles for a specific geographical location and time. It then seems to be skillful, but it is essentially due to the ability of the model to simulate the climatological diversity. To avoid this skill overestimation, the anomalies with respect to the local climatology can be considered. The local climatology is the temporal mean of the variable at each local verification points. Then, these values are subtracted from both the ensembles members and the verifications. After this translation, the CRPS, the Brier score and the Entropy can be assessed as in section 3.1. The main effect is the reduction of the *uncertainty* which then reflects the local variabilities.
In the same way, the definition of the thresholds x_{th} for the probability space diagnoses can be changed compared to section 3.1. The binary events are then locally defined by the temporal mean $\overline{x}$ and the temporal spread s_x at each verification point: for instance, $x_{th} = \overline{x} + s_x$ at each local verification point.

Chapter 4

References

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

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weatther and Forecasting*, **15**, pp 559–570.

Murphy A. H. 1973. A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, pp 595–600.

Roulston M. S. and L. Smith, 2002. Evaluating probabilistic forecast using information theory. *Mon. Wea. Review*, **130**, pp 1653–1660.

Toth Z., O. Talagrand, G. Candille, and Y. Zhu. 2003. 'Probability and ensemble forecasts' in Forecast Verification: A Practitioner's Guide in Atmospheric Science, Jolliffe I., Stephenson D.B. (eds), Wiley: UK. pp 137–163.