JOURNAL OF CLIMATE

Ensemble averaging and mean squared error

JONATHAN ROUGIER*

School of Mathematics, University of Bristol, UK

ABSTRACT

In fields such as climate science, it is common to compile an ensemble of different simulators for the same underlying process. It is a striking observation that the ensemble mean often out-performs at least half of the ensemble members in mean squared error (measured with respect to observations). In fact, as demonstrated in the most recent IPCC report, the ensemble mean often out-performs all or almost all of the ensemble members across a range of climate variables. This paper shows that these could be mathematical results based on convexity and averaging, but with implications for the properties of the current generation of climate simulators.

1. Introduction

A striking feature of Fig. 9.7 of Ch. 9 of the report of Working Group 1 to the Fifth Assessment Report of the IPCC (Flato et al., 2013, p. 766) is that the ensemble mean consistently outperforms more than half of the individual simulators on all variables, as shown by the blue rectangles on the lefthand side of Fig. 9.7, in the column labelled 'MMM'. Even more strikingly, the ensemble mean often outperforms *all* of the individual simulators (deep blue rectangles). This paper provides a mathematical explanation of these features.

Section 2 shows that it is a mathematical certainty that the ensemble mean will have a Mean Squared Error (MSE) which is no larger than the arithmetic mean of the MSEs of the individual ensemble members. This result holds for any convex loss function, of which squared error is but one example. While this does not imply the same relation for Root-MSE (RMSE) and the median ensemble member (as represented in Fig. 9.7), it makes a similar result plausible.

Section 3 establishes a stronger result, concerning the rank of the ensemble mean MSE among the individual MSEs (with an identical result for RMSE). This is based on a simple model of simulator biases, and on an asymptotic treatment of the behaviour of MSE in the case where the number of pixels increases without limit. Section 4 argues that this is a plausible explanation for the stronger result that the ensemble mean outperforms all of the individual simulators. A crucial aspect of this explanation is that it does not rely on 'offsetting biases', which would be inappropriate for the current generation of climate simulators.

In this paper I exercise my strong preference for 'simulator' over 'model' when referring to the code that produces climate-like outputs (see Rougier et al., 2013). This allows me to use the word 'model' without ambiguity, to refer to *statistical* models.

2. Convex loss functions

Let X_{ij} be the output for simulator i in pixel j, where there are k simulators and n pixels. I will always use i to index simulators and j to index pixels, and suppress the limits on sums, to reduce clutter. Let $\overline{X}_j$ be the ensemble arithmetic mean for pixel j,

$$\overline{X}_j := \frac{1}{k} \sum_i X_{ij}.$$

Let Y_j be the observation for pixel j. Denote the mean squared error as M_i for simulator i, and $\overline{M}$ for the ensemble mean:

$$M_i := \frac{1}{n} \sum_j (X_{ij} - Y_j)^2, \quad \overline{M} := \frac{1}{n} \sum_j (\overline{X}_j - Y_j)^2.$$

Then we have the following result, that the MSE of the ensemble mean is never larger than the arithmetic mean of the MSEs of the individual simulators.

Result 1. $\overline{M} \le k^{-1} \sum_i M_i$.

Corresponding author address: Jonathan Rougier, School of Mathematics, University of Bristol, Bristol BS8 1TW, UK
E-mail: j.c.rougier@bristol.ac.uk

Proof.

$$\begin{aligned}
\overline{M} &= \frac{1}{n}\sum\nolimits_j \left[\overline{X}_j - Y_j\right]^2 \\
&= \frac{1}{n}\sum\nolimits_j \left[\frac{1}{k}\sum\nolimits_i X_{ij} - Y_j\right]^2 \\
&= \frac{1}{n}\sum\nolimits_j \left[\frac{1}{k}\sum\nolimits_i (X_{ij} - Y_j)\right]^2 \\
&\leq \frac{1}{n}\sum\nolimits_j \frac{1}{k}\sum\nolimits_i (X_{ij} - Y_j)^2 \qquad (*) \\
&= \frac{1}{k}\sum\nolimits_i \frac{1}{n}\sum\nolimits_j (X_{ij} - Y_j)^2 \\
&= \frac{1}{k}\sum\nolimits_i M_i,
\end{aligned}$$

where $(*)$ follows by Jensen's inequality. □

This result holds for any convex function; replacing x^2 with $|x|$ gives the same inequality for Mean Absolute Deviation (MAD). The result for x^2 has previously appeared in the climate literature in Stephenson and Doblas-Reyes (2000) and Annan and Hargreaves (2011), but these authors failed to discover that it is the *convexity* of the loss function which induces this result.

This result falls short of being an explanation for the blue rectangles in the MMM column of Fig. 9.7 in two respects. First, Fig. 9.7 is drawn for Root Mean Squared Error (RMSE) not MSE, and second, it is drawn with respect to the median of the RMSEs of the ensemble, not the mean.

A weaker result is available for the RMSE; unfortunately Jensen's inequality does not help here, because the square root is a concave function (i.e., it bends the wrong way to extend the inequality). Let R_i be the RMSE of simulator i, and define

$$\tilde{R} := \frac{1}{k}\sum\nolimits_i R_i, \quad \sigma_R^2 := \frac{1}{k}\sum\nolimits_i (R_i - \tilde{R})^2,$$

the sample mean and sample variance of the RMSEs. Then, starting from Result 1,

$$\sqrt{\overline{M}} \leq \sqrt{\frac{1}{k}\sum\nolimits_i R_i^2} = \sqrt{\tilde{R}^2 + \sigma_R^2} = \tilde{R}\sqrt{1 + \frac{\sigma_R^2}{\tilde{R}^2}}.$$

If the variation in the simulators' RMSEs is small relative to their mean, then we would expect the RMSE of the ensemble mean to be no larger than the mean of the RMSEs of the individual simulators (although this outcome is not a mathematical certainty).

Extending the result from the mean to the median is trickier. A histogram of MSEs will typically be very positively skewed, and the histogram of RMSEs will remain positively skewed. Therefore the median and the mean of the RMSEs will not be similar. Typically the median will be lower, and therefore the mean being an upper bound does not imply that the median is an upper bound. But progress can be made using the result from the next section.

3. A simple systematic bias model

Flato et al. (2013, p. 767) and others have commented on the "notable feature" that $\overline{M}$ is typically smaller than any of the individual MSEs. This statement holds equally for MSE and RMSE, because rankings are preserved under increasing transformations, such as the square root.

A simple thought experiment suggests that this is indeed notable. If all of the simulators had MSE ε^2, and then we took pixel j in simulator i and steadily increased its X_{ij} value, then the MSE of simulator i and of the ensemble mean would increase. The rank of $\overline{M}$ among $M_1, \dots, M_k$ would be $k-1$, where the rank is defined as

$$\operatorname{rank}(\overline{M}, \boldsymbol{M}) := \sharp\left\{i : M_i \leq \overline{M}\right\} \qquad (1)$$

where $\boldsymbol{M} := (M_1, \dots, M_k)$, and $\sharp\{\cdot\}$ denotes the number of elements in the set. By Result 1, a rank of k is impossible if the M_i's are not identical. So ranks of between 0 and $k-1$ are attainable, in principle, and ranks that are consistently small invite an explanation.

A candidate explanation is found in weather forecast verification, in which it is sometimes found that a high resolution simulation has a larger MSE than a lower resolution simulation, when evaluated on high resolution observations (see, e.g. Mass et al., 2002). The explanation is that if the high resolution simulation puts a local feature such as a peak in slightly the wrong place (in space or time), then it suffers a 'double penalty', while a lower resolution simulation which does not contain the feature at all only suffers a single penalty. Following similar reasoning, we might argue that the ensemble mean is flatter than any individual member, and is thus penalized less if the individual members are putting local features in slightly wrong places. However, this argument is not compelling for the IPCC climate simulations, in which the observations have low resolution, and there is already substantial averaging in the individual simulator outputs.

I propose a different explanation, in terms of the simulators' 'biases'. Suppose each simulator has a systematic bias μ_i. Then over a large number of pixels the MSE of simulator i would be approximately μ_i^2 plus a constant. The ensemble mean at each pixel, though, would average the biases to the value $\bar{\mu} := k^{-1}\sum_i \mu_i$. Then over a large number of pixels the MSE of the ensemble mean would be approximately $\bar{\mu}^2$ plus a smaller constant (see the proof of Result 2). This looks to be a promising explanation, but there is work to be done, to establish the conditions under which the MSE of the ensemble mean is driven down towards or even below the smallest of the individual MSEs.

And also to establish that this is not just an 'offsetting biases' argument, which would be inappropriate for climate simulators (see section 4).

The mathematical challenge is that $(M_1, \dots, M_k, \overline{M})$ are not mutually independent: this difficulty was noted by Annan and Hargreaves (2011, p. 4532), who were unable to go beyond a heuristic explanation. One way to finesse this difficulty is with asymptotics, i.e. to consider the limit as the number of pixels increases without bound. This is not literally possible with climate simulators (which have a fixed domain), but, as is common practice in Statistics, asymptotic results serve to illuminate the situation when n is large (see, e.g., Cox and Hinkley, 1974, ch. 9). Results established asymptotically can be checked for finite n by simulation.

An asymptotic approach requires a statistical model of the joint relationship between the simulator outputs and the observations. Any results which are proved on the basis of the model are likely to hold for actual ensembles which might have been simulated from the model. Therefore we look to make the model as general as possible; the approach below is to start with a simple model, and then to check that the results generalize.

Define $Z_{ij} := X_{ij} - Y_j$ and take as the statistical model

$$Z_{ij} \mid \boldsymbol{\mu}, \sigma^2 \overset{\text{ind}}{\sim} \mathrm{N}(\mu_i, \sigma^2) \quad \text{for all } i \text{ and } j, \tag{2}$$

where μ_i is the 'bias' of simulator i, $\boldsymbol{\mu} := (\mu_1, \dots, \mu_k)$, and, for below, $\bar{\mu}$ is the arithmetic mean of $\boldsymbol{\mu}$. From now on, treat $\boldsymbol{\mu}$ and σ^2 as fixed constants, in order to avoid writing '$\mid \boldsymbol{\mu}, \sigma^2$' in every probability statement. The following result shows that, for this model, in the limit as $n \to \infty$ the rank of $\overline{M}$ among $M_1, \dots, M_k$ is completely and simply determined by $\boldsymbol{\mu}$ and σ^2.

Result 2. *For the model given in* (2),

$$\lim_{n \to \infty} \operatorname{rank}(\overline{M}, \boldsymbol{M}) = \sharp\{i : \mu_i^2 + \sigma^2 \le \bar{\mu}^2 + \sigma^2/k\},$$

where 'rank' *was defined in* (1).

The asymptotic theory in the following proof can be found in van der Vaart (1998, ch. 2); references to individual results are prefixed by 'vdV'.

Proof. In terms of Z_{ij},

$$M_i = \frac{1}{n} \sum\nolimits_j (X_{ij} - Y_j)^2 = \frac{1}{n} \sum\nolimits_j Z_{ij}^2.$$

Eq. (2) and the Weak Law of Large Numbers (WLLN, vdV 2.16) implies

$$M_i \xrightarrow{\mathrm{P}} \mathbb{E}(Z_{ij}^2) = \mu_i^2 + \sigma^2, \tag{3}$$

where '$\xrightarrow{\mathrm{P}}$' denotes convergence in probability (vdV, sec. 2.1). For $\overline{M}$,

$$\overline{M} = \frac{1}{n} \sum\nolimits_j \left(\frac{1}{k} \sum\nolimits_i (X_{ij} - Y_j) \right)^2 = \frac{1}{n} \sum\nolimits_j W_j^2$$

where $W_j := k^{-1} \sum_i Z_{ij}$. Eq. (2) and the WLLN implies

$$\overline{M} \xrightarrow{\mathrm{P}} \mathbb{E}(W_j^2) = \bar{\mu}^2 + \sigma^2/k.$$

Each M_i is converging in probability, and $\overline{M}$ is converging in probability, and hence $(M_1, \dots, M_k, \overline{M})$ is converging in probability (vdV 2.7). Then the Continuous Mapping Theorem (vdV 2.3) implies that

$$\frac{M_i}{\overline{M}} \xrightarrow{\mathrm{P}} \frac{\mathbb{E}(Z_{ij}^2)}{\mathbb{E}(W_j^2)} = \frac{\mu_i^2 + \sigma^2}{\bar{\mu}^2 + \sigma^2/k} \qquad i = 1, \dots, k.$$

This implies that in the limit as $n \to \infty$

$$M_i \le \overline{M} \iff \mu_i^2 + \sigma^2 \le \bar{\mu}^2 + \sigma^2/k \quad i = 1, \dots, k,$$

from which Result 2 follows directly. □

The result for ranking under RMSE is identical.

Generalizations. The Normal distribution for Z_{ij} is unnecessary; all that is required for the WLLN is that $\mathbb{E}(Z_{ij}^2) < \infty$. Different loss functions, such as Mean Absolute Deviation (MAD), can replace squared loss, providing that $\mathbb{E}\{|L(Z_{ij})|\} < \infty$, where L is the loss function. The common value of σ^2 for all simulators can be relaxed, at the expense of a slightly more complicated expression in Result 2. The independence can be relaxed somewhat, as long as the correlation length across pixels is small relative to the size of the domain. In particular, neighbouring pixels from the same simulator can have some positive correlation in their Z_{ij}'s. These generalizations affect the rate of convergence, and thus the accuracy of the result for finite n, but they do not affect the limit.

4. Interpretation

Result 2 shows that offsetting biases across the simulators in the ensemble, leading to $\bar{\mu} = 0$, would suffice to ensure that rank $= 0$ in the limit as $n \to \infty$. However, offsetting biases is an inplausible hypothesis for the current generation of climate simulators, as has been shown empirically in the 'genealogy' study of Knutti et al. (2013). Simulators have common biases, which cannot be expected to offset each other over the entire ensemble to give an overall mean of approximately zero. Additionally, as a reviewer has pointed out, the use of imperfect observations for Y_j introduces another common bias across all simulators.

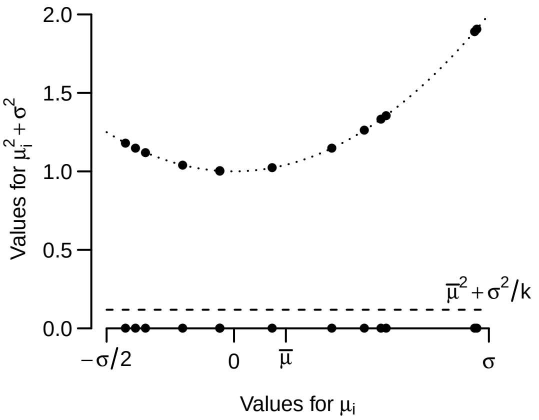

FIG. 1. A configuration of biases with $|\mu_i| < \sigma$, for which rank = 0 in the limit as $n \to \infty$. Here, $\sigma = 1$ and $k = 13$.

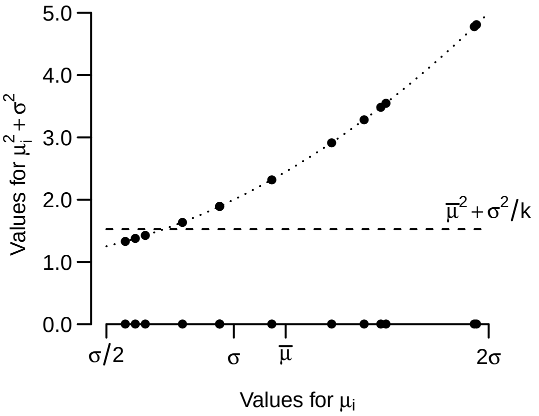

FIG. 2. Same as Figure 1, except with all of the μ_i's increased by σ; the asymptotic rank has increased from 0 to 3.

Therefore it is interesting that Result 2 can provide other sufficient conditions for which rank = 0, or very small.

Result 3. *If $|\mu_i| < \sigma$ for all i, then*

$$\lim_{k,n \to \infty} \text{rank}(\overline{M}, \boldsymbol{M}) = 0.$$

Proof. If $|\mu_i| < \sigma$ for all i then $|\bar{\mu}| < \sigma$, and $\bar{\mu}^2 < \sigma^2$. Thus a necessary (but not sufficient) condition for $M_i \le \overline{M}$ in the limit as $n \to \infty$ is $\mu_i^2 < \sigma^2/k$, where the righthand side goes to zero as $k \to \infty$. □

The condition in Result 3 can be summarized as *the simulators' biases are smaller in absolute size than the large pixel errors.* If individual simulators are tuned more on their overall bias than their large pixel errors, then we might expect something similar to this condition to hold.

Result 2 also illustrates when the ensemble mean performs badly. The two situations, good (for the ensemble mean, according to Result 3) and bad, are shown in Figures 1 and 2, in the limit as $n \to \infty$. In the first case, $|\mu_i| < \sigma$ and rank = 0. In the second case, $\mu_i \ge 0$ and $\bar{\mu} > \sigma$. The μ_i's larger than σ pull the value of $\bar{\mu}^2$ above σ^2, and this allows the small μ_i's to pass under the threshold in Result 2, and raise the rank.

There is a reason to distrust the asymptotic result when n is small. The distribution of M_i is very positively skewed, so that, for small n, the value of M_i will typically be less than its expectation, possibly much less. The value of $\overline{M}$ will typically be closer to its expectation. Therefore, the asymptotic rank is likely to be similar to a lower bound on the finite-n rank.

This can be tested in a stochastic simulation study. In each simulation, the configuration $\boldsymbol{\mu}$ is generated using

$$\mu_i \overset{\text{iid}}{\sim} \text{Unif}(-\sigma/2, \sigma) \qquad i = 1, \ldots, k. \tag{4}$$

For each configuration, the distribution of rank$(\overline{M}, \boldsymbol{M})$ is computed using the model in (2). I then repeat with all of the μ_i's increased by σ. I set $k = 42$, the same as Fig. 9.7 in Flato et al. (2013), and $n = 22$, which is the number of land regions in Giorgi and Mearns (2002), and much lower than the typical number of pixels. This small n is used to challenge the asymptotic nature of Result 2.

The simulation study reveals that the asymptotic approximation is accurate in the 'good' configurations of Figure 1. For all 30 configurations, the asymptotic value for the rank of $\overline{M}$ in $M_1, \ldots, M_k$ is 0. In every configuration, the probability of rank = 0 is at least 0.94, and the median probability across the configurations is 0.99. Across the configurations, the maximum value for the largest rank is 3, and the median value for the largest rank is 1. In summary, under the good configuration of the biases it is highly probable that the ensemble mean will outperform all of the individual simulators.

The outcome for the 'bad' configurations (see Figure 2) is shown in Figure 3. As anticipated, the distribution of the rank has shifted upwards away from 0 for each configuration, and it is clearer that the asymptotic result provides an approximate lower bound on the rank. The median rank for this simulation study is 21, since $k = 42$. Under the bias model, the distribution of the rank is located entirely below the median, for each configuration. Were n to be increased, the distribution of the rank in each configuration would collapse towards its asymptotic value. The median asymptotic value across the configurations is rank = 6. In summary, under the bad configuration of the biases it is highly probable that the ensemble mean will perform far better than the median simulator.

Thus the mathematics and the stochastic simulations show that the simulator biases model provides an explanation for Flato et al.'s "notable feature" of Fig. 9.7: perhaps it is because for many of the variables the simulators'

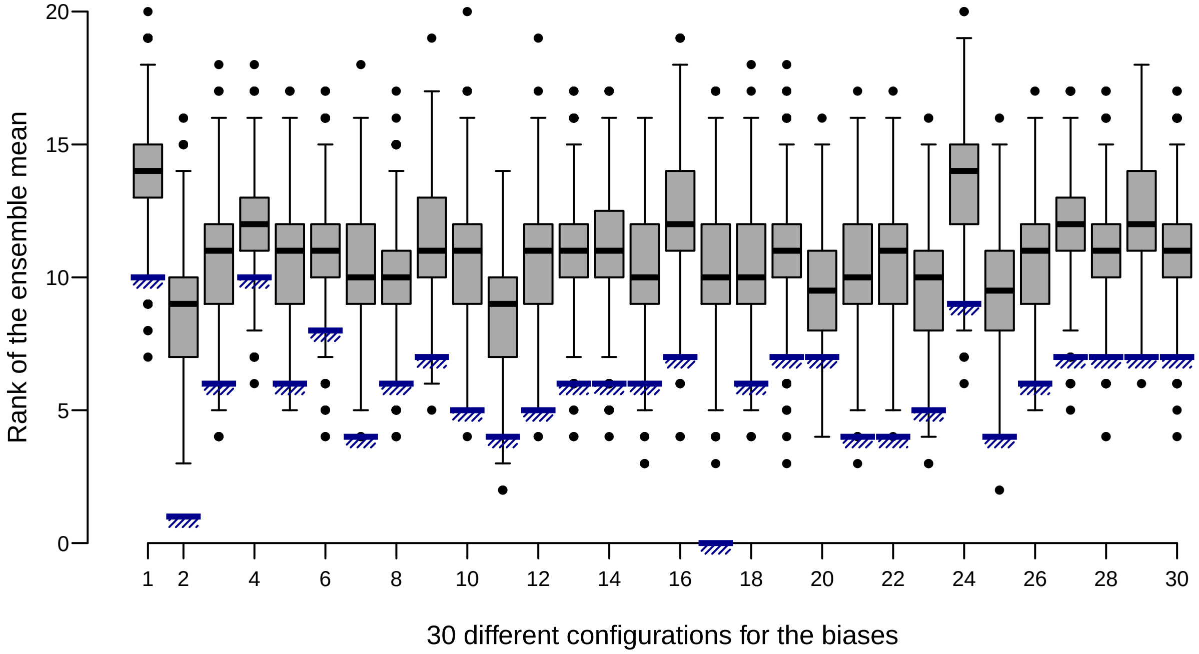

FIG. 3. Simulation study, for when the configuration of $\boldsymbol{\mu}$ does not satisfy $|\mu_i| < \sigma$, as shown in Figure 2. 30 different configurations of $\boldsymbol{\mu}$ are used, with the boxplots showing the the distribution of the rank of $\overline{M}$ in $M_1, \dots, M_k$. The 'ledge' for each configuration shows the asymptotic rank, using Result 2.

biases are smaller in absolute size than the large pixel errors. In this case, the notable feature of Fig. 9.7 is not just a mathematical artefact, but is telling us something interesting about the current generation of climate simulators.

Finally I would like to end with a caution about how to report and summarize ensemble model experiments. During the process of tuning the parameters of a climate simulator, a research group creates an ensemble of simulator versions with slightly different parameterisations. Result 2 suggests that they may get a lower MSE from the ensemble mean, than from their best-tuned simulator—again, we cannot assume offsetting biases in this case. If simulators are judged by the wider community on their MSEs, with more kudos and funding going to those research groups with lower MSEs, then the temptation will be to publicize the output from the ensemble mean rather than the best-tuned simulator. And yet the ensemble mean is 'less physical' at the pixel scale, since the space of climate states is not convex: linear combinations of valid climate states are not necessarily valid climate states. This makes the ensemble mean less suitable for providing boundary conditions, e.g. for regional downscaling and risk assessment. So research groups might consider how to certify the outputs they publicize, if they do not want to put their simulators in the public domain.

Acknowledgments. I would like to thank Yi Yu for her very helpful comments on my mathematics; Reto Knutti and Ben Sanderson for a discussion which prompted me to investigate the result in section 3; and Ken Mylne for an illuminating conversation about weather forecast verification. Two reviewers made perceptive comments on all aspects of this paper, and substantial improvements in its clarity. This research was supported by the EPSRC grant SuSTaIn, reference EP/D063485/1.

References

Annan, J., and J. Hargreaves, 2011: Understanding the CMIP3 multi-model ensemble. *Journal of Climate*, **24**, 4529–4538, doi:10.1175/2011JCLI3873.1.

Cox, D., and D. Hinkley, 1974: *Theoretical Statistics*. Chapman and Hall, London, UK.

Flato, G., and Coauthors, 2013: Evaluation of climate models. Stocker et al. (2013), chap. 9, 741–866.

Giorgi, F., and L. Mearns, 2002: Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via 'reliability ensemble averaging' (REA) method. *Journal of Climate*, **15**, 1141–1158.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, **40**, 1194–1199.

Mass, C., D. Owens, K. Westrick, and B. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, **83 (3)**, 407–430.

Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multi-model ensembles. *Journal of the American Statistical Association*, **108**, 852–863.

Stephenson, D., and F. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus*, **50**, 300–322.

Stocker, T., and Coauthors, 2013: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge UK and New York NY, USA.

van der Vaart, A., 1998: *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.