PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2012) **370**, 1087–1099
doi:10.1098/rsta.2011.0307

Model complexity versus ensemble size: allocating resources for climate prediction

By Christopher A. T. Ferro[1,2,*], Tim E. Jupp[2], F. Hugo Lambert[2], Chris Huntingford[3] and Peter M. Cox[2]

[1]*National Centre for Atmospheric Science, and* [2]*Mathematics Research Institute, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK*
[3]*Centre for Ecology and Hydrology, Wallingford OX10 8BB, UK*

A perennial question in modern weather forecasting and climate prediction is whether to invest resources in more complex numerical models or in larger ensembles of simulations. If this question is to be addressed quantitatively, then information is needed about how changes in model complexity and ensemble size will affect predictive performance. Information about the effects of ensemble size is often available, but information about the effects of model complexity is much rarer. An illustration is provided of the sort of analysis that might be conducted for the simplified case in which model complexity is judged in terms of grid resolution and ensemble members are constructed only by perturbing their initial conditions. The effects of resolution and ensemble size on the performance of climate simulations are described with a simple mathematical model, which is then used to define an optimal allocation of computational resources for a range of hypothetical prediction problems. The optimal resolution and ensemble size both increase with available resources, but their respective rates of increase depend on the values of two parameters that can be determined from a small number of simulations. The potential for such analyses to guide future investment decisions in climate prediction is discussed.

Keywords: general circulation models; initial condition ensembles; mean-squared error; resolution; weather forecasting

1. Introduction

Two of the main drivers of improved weather forecasts and climate predictions in recent decades have been the increasing complexity of general circulation models and the increasing number of trajectories simulated with those models. Both of these advances place greater demands on computational resources and so a trade-off emerges: should future resources be invested in more complex models or in larger ensembles of simulations? This question pertains to forecasting on all time scales. We believe that answers to this question would benefit from quantitative analysis, but efforts are rarely made to collect the necessary data. In this paper,

*Author for correspondence (c.a.t.ferro@exeter.ac.uk).

One contribution of 13 to a Theme Issue ‘Climate predictions: the influence of nonlinearity and randomness’.

we illustrate how such analyses might be conducted when the relevant data are available. Our approach is highly simplified, but we hope that it will stimulate more detailed quantitative investigations of this important topic. We begin with a brief discussion of the roles of ensembles and model complexity.

Ensemble simulations are now routinely used to investigate the uncertainty that arises from limitations in our knowledge [1]. For example, uncertainty about external forcings of the climate system can be explored by running a model with each of several different boundary conditions. The impact of imperfections in models can be explored by running several models, differing either in the structure of their equations (to produce a multi-model ensemble) or just in the values of the parameters in their equations (to produce a perturbed-physics ensemble). An initial condition ensemble is formed by varying only the initial state of a model. These latter ensembles provide information about the frequency distribution of simulated weather that arises from a model's sensitivity to initial conditions. Increasing the number of members in an initial condition ensemble improves the precision with which this model climate (and properties of it such as its mean) is known, but the ensemble size will be limited by the available computational resources.

General circulation models are based on equations derived from physical laws, which are approximated and solved numerically on a discrete grid. The complexity of a model may be judged in terms of both the physical processes that are represented in the model and the resolution of the grid on which the equations are solved. Increasing model complexity may improve the accuracy with which a model simulates the climate system, but again the complexity will be limited by the available computational resources.

Weather forecasting and climate prediction, therefore, would benefit from both more complex models and larger ensembles. In order to simplify our illustrative analysis, we shall concentrate on the roles of model resolution and initial condition ensembles because their likely effects on forecast performance are more amenable to mathematical description. Extensions to other aspects of model complexity and other types of ensemble are left for future research. Larger initial condition ensembles provide more precise information about a model's climate, while increasing model resolution (reducing the distance between grid points) changes that climate, potentially making it closer to reality. Increasing ensemble size or model resolution is expensive, and resource constraints mean that a balance must be struck. Should future investment be targeted at higher resolution? If so, how high? Or should investment be targeted at larger ensembles? If so, how large? Is there an optimum trade-off between the two competing demands?

In §2, we survey existing evidence for how ensemble size and model resolution might be expected to affect the quality of predictions. In §§3 and 4, we outline a simple framework for describing these effects mathematically. In §5, we illustrate how such a framework can be used to identify an optimal trade-off between ensemble size and model resolution. We close with a discussion in §6.

2. Evidence for the effects of ensemble size and resolution

In order to decide where to invest resources, we should like to be able to predict how changes in ensemble size and model resolution will affect the quality of

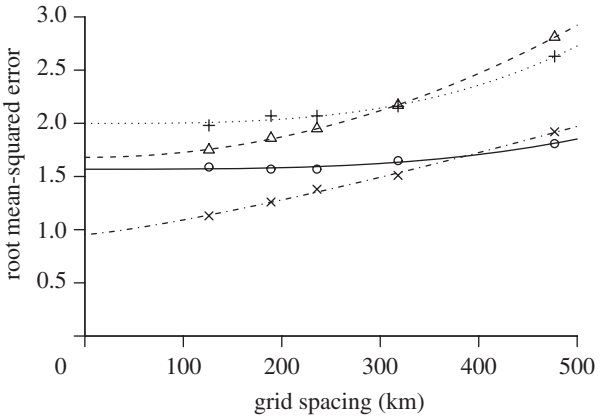

Figure 1. Root mean-squared errors against horizontal grid spacing for temperature at 850 hPa (T850 in K, circles), mean sea-level pressure (MSLP in hPa, triangles), geopotential height at 500 hPa (Z500 in dam, pluses) and zonal wind speed at 850 hPa (U850 in $\mathrm{m\,s^{-1}}$, crosses) simulated by the ECHAM5 model. Curves defined by expressions (4.1) and (4.2) and fitted by least squares are superimposed.

forecasts without actually having to generate extra ensemble members or create new models at higher resolutions. For this to be possible, we must first have some evidence that ensemble size and resolution have predictable effects on forecast performance.

For many measures of performance, increasing the size of initial condition ensembles does have a beneficial effect that is systematic and predictable. This is because the statistical properties of additional members of these ensembles are often similar to the properties of existing members. This has been known since the early days of ensemble weather forecasting. Leith [2], for example, showed that the expected mean-squared error (m.s.e.) of the ensemble mean changes at a rate that is inversely proportional to the ensemble size. Many other analytical and empirical demonstrations of such effects have been published for lead times ranging from a few days to months [3–12].

Evidence for predictable effects of increasing model resolution is rarer. Indeed, isolating an effect that is owing to model resolution alone is difficult in practice because, usually, at least some model parameters must be tuned to each resolution in order to obtain realistic simulations. Nonetheless, there are some studies in the literature that identify effects which can be attributed largely to changes in resolution [13–17]. These effects, however, are not always systematic. For example, the performance of simulations can jump by unpredictable amounts when the resolution crosses a threshold beyond which new physical processes are resolved by the model. Even when the effects are systematic, increasing resolution does not always improve performance. However, the studies cited above suggest that there are times when predictable effects exist, when smooth changes in performance can be expected over a limited range of resolutions, and it is on these situations that we shall focus. Roeckner *et al.* [16] provide the most accessible, quantitative examples in their comparison of multiple resolutions of the ECHAM5 general circulation model. Figure 1 presents part

of their data in graphical form, where, for four variables, the m.s.e. of global and seasonal means, averaged over seasons, is found to decrease smoothly across five different resolutions.

In figure 1 and throughout this paper, we consider the effects of model resolution on the prediction of scalar quantities, such as global means, rather than spatial fields. The former provides key tests of model quality, and understanding the effects of resolution in these terms is therefore important. By neglecting the case of spatial fields, however, we ignore a second, important benefit of increasing model resolution, which is to produce forecasts with greater spatial variability. Incorporating this second effect is left for future research.

3. Predicting the effects of ensemble size

We have seen that there are situations in which we may reasonably assume that the effects of increasing resolution and ensemble size can be predicted to some useful degree. In §§3 and 4, we present a simple example of how such effects might be described mathematically. This leads to a formula for how performance depends on resolution and ensemble size, which we then use to analyse the trade-off between the two effects.

Let Y represent the observed value of a scalar predictand, and let $X_1, \ldots, X_n$ represent the corresponding predictions made by n members of an initial condition ensemble. We derive the effect of ensemble size on the m.s.e.,

$$S = (\bar{X} - Y)^2,$$

of the ensemble mean, $\bar{X}$, where

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Our choice of the m.s.e. reflects its widespread use and the tractability of the results that follow, but other measures could be used instead.

We suppose that we are interested in the performance of the ensemble over some, possibly short, time interval during which the statistical properties of the ensemble members and observations are approximately stationary. We account for temporal variation in the ensemble members and observation over this interval by defining $X_1, \ldots, X_n$ and Y to be random variables and then seek the effect of ensemble size on the expected value, $E(S)$, of the m.s.e. To do so, we must make some assumptions about the statistical properties of the ensemble members, otherwise we would be unable to tell what effect adding or removing ensemble members would have on the expected m.s.e. The only assumption that we make is that the ensemble members are second-order exchangeable [18]. In other words, the ensemble members all have the same expectations and variances and all pairs of ensemble members are equally correlated. This is weaker than assuming independent and identically distributed ensemble members and assumes nothing about the relationship between the ensemble members and the observation. Our assumption is encapsulated in the following notation. For all i, the expectations of our random variables are

$$E(Y) = \mu_o \quad \text{and} \quad E(X_i) = \mu,$$

the variances are

$$\mathrm{var}(Y) = \sigma_o^2 \quad \text{and} \quad \mathrm{var}(X_i) = \sigma^2,$$

and the correlations are

$$\mathrm{corr}(X_i, Y) = \rho_o \quad \text{and} \quad \mathrm{corr}(X_i, X_j) = \rho,$$

for all $j \neq i$. With these assumptions, we may now derive the effect of ensemble size on the expected m.s.e.

We show in appendix A that the expected m.s.e. is

$$E(S) = \sigma_o^2 - 2\rho_o\sigma\sigma_o + \rho\sigma^2 + (\mu - \mu_o)^2 + \frac{\sigma^2(1-\rho)}{n}. \tag{3.1}$$

This expression simplifies in situations in which the initial conditions confer no predictive skill on the ensemble members at the lead time of interest. This is typical of many climate predictions for which the lead time is longer than the time scales of any internal processes that influence significantly the variability of the predictand. In such cases, we may assume that the ensemble members are mutually uncorrelated and also uncorrelated with the observation so that $\rho = \rho_o = 0$. Under this assumption, the skill of the ensemble members comes from predicting the mean climate rather than the daily variations in weather.

We shall use the general expression (3.1) for the expected m.s.e. in order to analyse the benefit that would be gained from increasing the ensemble size. Although we would minimize the expected m.s.e. by choosing the ensemble size to be as large as possible, the benefit gained from increasing n depends on the magnitude of the ensemble variance, $\sigma^2(1-\rho)$, relative to the other terms in the expression. If the variance is relatively large, then a large percentage reduction in m.s.e. will result from increasing ensemble size; but, if the variance is relatively small, then the percentage reduction will be small. A law of diminishing returns also applies: one extra ensemble member has a greater impact on the m.s.e. if the current ensemble size is small than if the current ensemble size is large.

4. Predicting the effects of model resolution

In §3, we derived an expression for the effect of ensemble size on the expected m.s.e. under quite weak assumptions about the statistical properties of initial condition ensemble members. Finding an expression for the effect of resolution is more challenging. Rather than attempting to derive such an expression theoretically, we adopt an alternative strategy: assume a sufficiently flexible parametric form for the effect of resolution on the statistical properties of ensemble members. Such a form can be fitted to ensembles from a model run at a small number of different resolutions in order to assess empirically its adequacy and to estimate its parameters. We adopt one, simple form below for illustrative purposes, but alternatives may be used with equal ease.

Define the grid spacing, r, to be the horizontal distance between grid points. This is inversely proportional to spectral truncation and model resolution so that

Table 1. Parameter estimates. Parameter estimates and estimated standard errors in brackets for the data in figure 1 when the grid spacing, r, is rescaled to have units of 1000 km.

	ϵ	α	δ	σ
T850	1.57 (0.02)	2.7 (2.5)	3.2 (1.3)	0.17
MSLP	1.68 (0.01)	5.1 (0.2)	2.0 (0.1)	0.07
Z500	2.00 (0.03)	6.6 (3.2)	3.2 (0.7)	0.33
U850	0.95 (0.09)	2.4 (0.2)	1.2 (0.2)	0.13

higher resolution corresponds to smaller values of r. Let us write the effect of resolution on the expected value of an ensemble member as

$$\mu = \mu(r) = \mu_o + \epsilon + \beta(r), \tag{4.1}$$

where we assume that the resolution-dependent bias $\beta(r) \to 0$ as $r \to 0$. Then, ϵ can be interpreted as an irreducible error. We also assume for simplicity that the model standard deviation, σ, and the correlations ρ and ρ_o are constant in r, although resolution-dependent models for these parameters could also be incorporated.

We found earlier that the expected m.s.e. (3.1) decreases monotonically as the ensemble size, n, increases. In contrast, our model (4.1) admits non-monotonic dependence of the m.s.e. on the grid spacing, r. For example, if $\beta(r)$ and ϵ have different signs, then there may be a grid spacing $r > 0$ such that $\beta(r) = -\epsilon$ and, therefore, $\mu(r) = \mu_o$. If $\beta(r)$ and ϵ have the same sign, however, then our model (4.1) implies that the expected m.s.e. will decrease monotonically as $r \to 0$.

To complete our specification of the effect of resolution, we must specify a parametric form for $\beta(r)$. We consider the following example:

$$\beta(r) = \alpha r^{\delta}, \tag{4.2}$$

where $\delta > 0$. This choice is motivated by the finding that the energy spectra of atmospheric motions tend to follow power laws [19], such that the unresolved energy has a power-law dependence on model resolution. We assume that the bias, $\beta(r)$, associated with these unresolved motions also approximates a power law in r.

The parameters ϵ, α and δ in the model defined by expressions (4.1) and (4.2) may be estimated from a small number of model simulations run at different resolutions. For example, the bias, $\mu(r) - \mu_o$, may be estimated by the difference between the mean of an ensemble run at grid spacing r and a time mean of corresponding observations. If such bias estimates are available for a few different resolutions, then the curve $\epsilon + \alpha r^{\delta}$ can be fitted to them by least squares. We have been unable to obtain bias estimates with which to illustrate this procedure and so we fitted the curve to the data in figure 1 instead. This would be equivalent to fitting the curve to bias estimates were the biases of the simulations underlying figure 1 positive in all four seasons. The parameter estimates are presented in table 1 and the fitted curves are superimposed on the data in figure 1. Although some of the parameters are estimated with low precision, our simple parametric form is able to describe the effect of resolution quite closely.

5. Analysing the trade-off

(*a*) *Model cost*

In this section, we demonstrate how descriptions of the effects of ensemble size and model resolution on performance measures such as the m.s.e. can be used to identify an optimal allocation of finite resources. First, we need to know how the cost of running an ensemble member depends on the resolution of the model. The form of this dependence is determined by the details of how the model code would change with resolution, and so different forms of dependence will be required for different modelling systems. The change of cost with resolution or other aspects of model complexity will typically contain some discontinuities, as model parametrizations change their nature for example, but for our illustrative analysis we adopt a simple, smooth form based on the following generic ideas. Such forms may hold approximately over restricted resolution ranges, and alternative forms could be used with equal ease.

Halving the horizontal distance, r, between grid points quadruples the number of grid points and, therefore, quadruples the number of calculations required to run the model. So the number of calculations is approximately inversely proportional to r^2. In addition, however, the time step used to solve the model equations numerically tends to vary in direct proportion to r and so the number of calculations is approximately inversely proportional to r^3 [17]. If the vertical distance between grid points is also proportional to r [20], then the number of calculations is approximately inversely proportional to r^4. Rather than choosing a particular power now, we approximate the cost of running a single ensemble member at grid spacing r by c/r^γ for some positive constants c and γ. We shall specialize to the case $\gamma = 4$ later.

(*b*) *Resource allocation*

Suppose now that we have a total computational resource C for producing an initial condition ensemble of size n. Then, we are required to operate under the constraint $nc/r^\gamma = C$ so that $n = Cr^\gamma/c$. Given our expressions (4.1) and (4.2) for how resolution affects the model bias, the expected m.s.e. (3.1) becomes

$$E(S) = \sigma_o^2 - 2\rho_o\sigma\sigma_o + \rho\sigma^2 + (\epsilon + \alpha r^\delta)^2 + \frac{c\sigma^2(1-\rho)}{Cr^\gamma}.$$

This is minimized at $r = r_{\mathrm{opt}}$, where r_{opt} is defined implicitly by the equation

$$C = \frac{\gamma c\sigma^2(1-\rho)}{2\delta\alpha r_{\mathrm{opt}}^{\delta+\gamma}(\epsilon + \alpha r_{\mathrm{opt}}^\delta)} \tag{5.1}$$

when $\alpha \neq 0$, and the corresponding optimal ensemble size is

$$n_{\mathrm{opt}} = \frac{\gamma\sigma^2(1-\rho)}{2\delta\alpha r_{\mathrm{opt}}^\delta(\epsilon + \alpha r_{\mathrm{opt}}^\delta)}. \tag{5.2}$$

As C increases, more resources become available and so r_{opt} decreases and n_{opt} increases. As c increases, the model becomes more expensive and so r_{opt} increases and n_{opt} decreases. As σ increases or ρ decreases, the ensemble variance

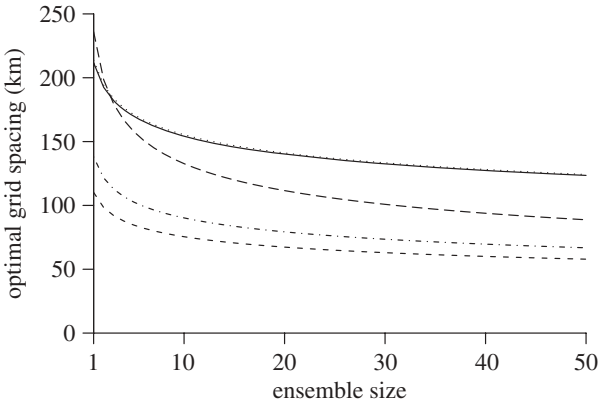

Figure 2. Optimal grid spacing against the ensemble size that can be afforded with a grid spacing of 276 km based on the data for the four variables in table 1: T850 (solid line), MSLP (dashed line), Z500 (dotted line) and U850 (dotted-dashed line). The smallest grid spacing (long-dashed line) that can be afforded with a one-member ensemble is also shown.

increases and both r_{opt} and n_{opt} increase. As the absolute value of α increases, the resolution-dependent bias increases and both r_{opt} and n_{opt} decrease. If $\epsilon\alpha > 0$, then, as the absolute value of ϵ increases, the impact of resolution on the squared bias increases and both r_{opt} and n_{opt} decrease. The opposite holds if $\epsilon\alpha < 0$ since then a large resolution (small grid spacing) is needed to balance a large irreducible error. If $\epsilon\alpha > 0$, then both r_{opt} and n_{opt} are minimized at an intermediate value of δ; both increase as $\delta \to 0$ since then the effect of resolution on the bias reduces; and both increase to the asymptotes $r_{\text{opt}} = 1$ and $n_{\text{opt}} = C/c$ as $\delta \to \infty$ since then any grid spacing below $r = 1$ yields a small bias. If $\epsilon\alpha < 0$, then r_{opt} and n_{opt} decrease monotonically as δ increases.

We can identify the optimal resolution and ensemble size to use for a given resource C if we can determine the values of the parameters in our equations (5.1) and (5.2). We have already described how ϵ, α and δ may be estimated. The ensemble variance, $\sigma^2(1-\rho)$, may be estimated from small initial condition ensembles. For illustration, we use the parameter estimates presented in table 1 where, for want of relevant model data, σ has been estimated by the standard deviation of global, seasonal means from the ERA-40 reanalysis [21] for the same time period, 1979–1993, used by Roeckner *et al.* [16]. We can assume $\rho = \rho_o = 0$ for these ensembles and we set $\gamma = 4$. Figure 2 shows how r_{opt} changes with the available resource, where resource is measured in terms of the number of ensemble members that could be afforded when the model is run at grid spacing 276 km (corresponding to spectral truncation T85). For example, an ensemble size equal to one should be interpreted as the resource sufficient to run one member at resolution T85. For all variables and all levels of resource, we find that the optimal grid spacing is lower than 276 km. The optimal resolutions for T850 and Z500 are similar to one another, and simulations of U850 and mean sea-level pressure (MSLP) would benefit from higher resolutions. The optimal resolutions for U850 and MSLP correspond to optimal ensemble sizes that are less than one, however, and so the optimal resolution for these two variables is actually the highest resolution that can be afforded with a one-member ensemble.

(*c*) *Large and small irreducible errors*

In two special cases, equations (5.1) and (5.2) yield closed-form expressions for the optimal grid spacing and ensemble size. When the irreducible error, ϵ, is small compared with the resolution-dependent bias, $\alpha r_{\mathrm{opt}}^{\delta}$, equations (5.1) and (5.2) yield

$$r_{\mathrm{opt}} \approx \left[\frac{\gamma c\sigma^2(1-\rho)}{2\delta\alpha^2}\right]^{1/(2\delta+\gamma)} C^{-1/(2\delta+\gamma)}$$

and

$$n_{\mathrm{opt}} \approx \left[\frac{\gamma\sigma^2(1-\rho)}{2\delta\alpha^2 c^{2\delta/\gamma}}\right]^{\gamma/(2\delta+\gamma)} C^{2\delta/(2\delta+\gamma)},$$

while the cost of each member for this optimal model is

$$cr_{\mathrm{opt}}^{-\gamma} \approx \left[\frac{2\delta\alpha^2 c^{2\delta/\gamma}}{\gamma\sigma^2(1-\rho)}\right]^{\gamma/(2\delta+\gamma)} C^{\gamma/(2\delta+\gamma)}.$$

The logarithms of the optimal grid spacing, ensemble size and model cost, therefore, scale approximately linearly with the logarithm of the total resource, C, and this scaling depends on the values of δ and γ only. Moreover, as resource increases, the optimal model cost increases faster than the ensemble size if and only if $\delta < \gamma/2$.

When the irreducible error is large compared with the resolution-dependent bias, something that can happen only when $\epsilon\alpha > 0$, we obtain the approximations

$$r_{\mathrm{opt}} \approx \left[\frac{\gamma c\sigma^2(1-\rho)}{2\delta\alpha\epsilon}\right]^{1/(\delta+\gamma)} C^{-1/(\delta+\gamma)}$$

and

$$n_{\mathrm{opt}} \approx \left[\frac{\gamma\sigma^2(1-\rho)}{2\delta\alpha\epsilon c^{\delta/\gamma}}\right]^{\gamma/(\delta+\gamma)} C^{\delta/(\delta+\gamma)},$$

while the model cost is

$$cr_{\mathrm{opt}}^{-\gamma} \approx \left[\frac{2\delta\alpha\epsilon c^{\delta/\gamma}}{\gamma\sigma^2(1-\rho)}\right]^{\gamma/(\delta+\gamma)} C^{\gamma/(\delta+\gamma)}.$$

A different log-linear relationship holds in this case, with the optimal model cost increasing faster than the ensemble size if and only if $\delta < \gamma$.

(*d*) *Non-dimensionalization*

A useful graphical representation of how the optimal grid spacing and ensemble size change with resource can be obtained by non-dimensionalizing our equations. Define the following dimensionless quantities:

$$\tilde{r} = \frac{r}{r_{\mathrm{c}}}, \quad \tilde{n} = \left[\frac{2\delta\epsilon^2}{\gamma\sigma^2(1-\rho)}\right] n \quad \text{and} \quad \tilde{C} = \tilde{n}\tilde{r}^{-\gamma},$$

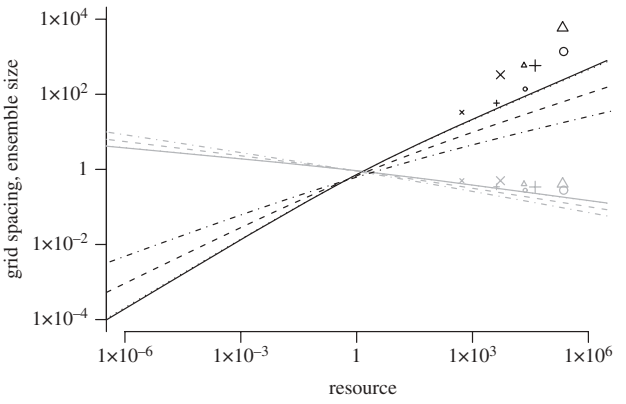

Figure 3. Dimensionless grid spacing ($\tilde{r}$, grey) and dimensionless ensemble size ($\tilde{n}$, black) against dimensionless resource ($\tilde{C}$) on logarithmic axes for the parameter estimates in table 1: temperature at 850 hPa (T850: solid line, circles), mean sea-level pressure (MSLP: dashed line, triangles), geopotential height at 500 hPa (Z500: dotted line, pluses) and zonal wind speed at 850 hPa (U850: dotted-dashed line, crosses). The curves show how the optimal grid spacing decreases and the optimal ensemble size increases as resource increases. The symbols correspond to a grid spacing of 276 km and ensemble sizes 10 (large symbols) and one (small symbols).

where $r_c = |\epsilon/\alpha|^{1/\delta}$ is a 'critical' grid spacing. As before, the optimal, dimensionless grid spacing and ensemble size are defined implicitly by two equations. When $\epsilon\alpha > 0$, these equations are

$$\tilde{C} = \frac{1}{\tilde{r}_{\mathrm{opt}}^{\delta+\gamma}(1+\tilde{r}_{\mathrm{opt}}^{\delta})}$$

and

$$\tilde{n}_{\mathrm{opt}} = \frac{1}{\tilde{r}_{\mathrm{opt}}^{\delta}(1+\tilde{r}_{\mathrm{opt}}^{\delta})}.$$

The relationships between $\tilde{r}_{\mathrm{opt}}$, $\tilde{n}_{\mathrm{opt}}$ and $\tilde{C}$, therefore, depend only on δ and γ. Moreover, relative changes in these dimensionless quantities equal the relative changes in r, n and C since $\tilde{r} \propto r$, $\tilde{n} \propto n$ and $\tilde{C} \propto C$. The curves in figure 3 show how $\tilde{r}_{\mathrm{opt}}$ and $\tilde{n}_{\mathrm{opt}}$ change with $\tilde{C}$ for the four variables listed in table 1 when $\gamma = 4$. At small and large values of $\tilde{C}$, the curves approach the log-linear relationships identified in §5*c* for the cases of small and large irreducible errors, respectively. The balance of grid spacing and ensemble size for existing ensembles can also be judged by plotting $\tilde{r}$ and $\tilde{n}$ against $\tilde{C}$ and seeing how far the points fall above or below the curves of optimal resource allocation. We illustrate this idea for the variables in table 1 by plotting points corresponding to the situations in which the available resource is sufficient to run either 10 or one ensemble members at grid spacing 276 km. As discovered earlier, in all cases, the resolution is too low (grid spacing too large) and the ensemble size too high.

6. Discussion

The question of resource allocation for weather forecasting and climate prediction is ubiquitous. Should the priority be to create more models, more complex models, or to learn more about existing models by creating larger ensembles? In this paper, we have argued that attempts should be made to answer these questions quantitatively, and we have provided a simple illustration of the sort of approach that might be taken. We showed how expressions for the effects of ensemble size and model resolution on the m.s.e. of an initial condition ensemble mean can be used to analyse the trade-off between these two effects and to identify the optimal ensemble size and resolution for a given level of computational resource. We derived our expression for the effect of ensemble size analytically under the weak assumption of second-order exchangeable ensemble members and obtained our expression for the effect of model resolution empirically from a small number of model simulations. The simplifying assumptions that we adopted may need to be relaxed in order to develop more realistic frameworks. We illustrated our approach with data from simulations of the ECHAM5 climate model published previously by Roeckner *et al.* [16]. Our analysis suggested that, at current resource levels, higher resolution should perhaps take priority over larger ensembles, but we emphasize that our numerical results should be viewed as an illustration of an approach rather than as a reliable guide for resource allocation for ECHAM5 or any other climate model. Although our framework has various limitations that we discuss below, we believe that extensions of our approach could be highly informative and we hope that these ideas will be pursued by other researchers.

We considered only one measure of performance: the m.s.e. of the ensemble mean. This is a widely used measure for which the ensemble-size effect is tractable, but it fails to assess other properties of the ensemble such as its spread. Extensions to other measures such as ranked probability scores are possible [12], but incorporating the effects of model complexity in those cases may be more challenging. Another improvement may be to measure performance relative to the magnitude of the predictand. For example, if the ensemble mean represents the proportion of ensemble members predicting the occurrence of a rare event, then the absolute m.s.e. typically will be small, but achieving a small relative m.s.e. is more important. Although the same allocation of resources will optimize both the absolute and relative m.s.e., the level of resource needed to achieve a small relative m.s.e. for rare events can be very large: high resolution (small grid spacing) may be needed to control the relative bias and large ensembles may be needed to control the relative precision of the ensemble mean. A third improvement would be to incorporate a measure of the benefit of the greater spatial detail provided by higher resolution models.

An important extension of our work would accommodate perturbed-physics and multi-model ensembles. These ensembles are most relevant when forecasting at seasonal and longer lead times, over which the impacts of model imperfections become more significant. However, the effects of increasing the size of such ensembles are harder to predict than for initial condition ensembles because the statistical properties of ensemble members are affected by model structures and parameter values in complex ways. Multi-model ensembles are especially challenging. Determining how to distribute resources across a given set of existing models may be the best approach in such circumstances. More is possible for

perturbed-physics ensembles if it can be assumed that ensemble members respond smoothly to changes in parameter values. In that case, statistical theory on the design of experiments [22] provides a way forward.

We relied on empirical information to predict the effects of changing model resolution on the statistical characteristics of ensemble members. Supplementing this information with theoretical results from numerical analysis would be beneficial, particularly, as the empirical evidence available at present is sparse. We also interpreted model complexity in the narrow sense of model resolution. Extending our approach to accommodate the expected effects of changing models in other ways would be desirable.

Finally, decisions about resource allocation can rarely be formalized entirely in the sort of framework that we have outlined. For example, optimal allocations are typically unique to each predictand, and so compromises must be made to find an allocation that yields acceptable performance across all predictands of interest. Nonetheless, such compromises should be based on relevant evidence. We believe that valuable additions to such evidence could be provided by the type of analysis proposed in this paper.

The authors are grateful to Dr Renate Brokopf for providing access to data from the ECHAM5 simulations, and to the staff of the Isaac Newton Institute for Mathematical Sciences and the organizers of its programme on Mathematical and Statistical Approaches to Climate Modelling and Prediction (11 August–22 December 2010), during which this work was initiated.

Appendix A

We derive the expression (3.1) for the expected m.s.e. We have

$$S = [(\bar{X} - \mu) + (\mu - \mu_o) - (Y - \mu_o)]^2$$

and, therefore,

$$\begin{aligned}
E(S) &= E[(Y - \mu_o)^2] + E[(\mu - \mu_o)^2] + E[(\bar{X} - \mu)^2] \\
&\quad - 2E[(\bar{X} - \mu)(Y - \mu_o)] + 2(\mu - \mu_o)E[(\bar{X} - \mu)] - 2(\mu - \mu_o)E[(Y - \mu_o)] \\
&= \mathrm{var}(Y) + (\mu - \mu_o)^2 + \mathrm{var}(\bar{X}) - 2\mathrm{cov}(\bar{X}, Y) \\
&= \sigma_o^2 + (\mu - \mu_o)^2 + \frac{1}{n^2}\left[\sum_{i=1}^{n} \mathrm{var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j)\right] - \frac{2}{n}\sum_{i=1}^{n} \mathrm{cov}(X_i, Y) \\
&= \sigma_o^2 + (\mu - \mu_o)^2 + \frac{1}{n^2}\left[n\sigma^2 + n(n-1)\rho\sigma^2\right] - 2\rho_o\sigma\sigma_o \\
&= \sigma_o^2 - 2\rho_o\sigma\sigma_o + \rho\sigma^2 + (\mu - \mu_o)^2 + \frac{\sigma^2(1-\rho)}{n}.
\end{aligned}$$

References

1 Collins, M. 2007 Ensembles and probabilities: a new era in the prediction of climate change. *Phil. Trans. R. Soc. A* **365**, 1957–1970. (doi:10.1098/rsta.2007.2068)

2 Leith, C. E. 1974 Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418. (doi:10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2)

3 Murphy, J. M. 1988 The impact of ensemble forecasts on predictability. *Q. J. R. Meteorol. Soc.* **114**, 463–493. (doi:10.1002/qj.49711448010)

4 Déqué, M. 1997 Ensemble size for numerical seasonal forecasts. *Tellus* **49A**, 74–86. (doi:10.1034/j.1600-0870.1997.00005.x)

5 Buizza, R. & Palmer, T. N. 1998 Impact of ensemble size on ensemble prediction. *Mon. Weather Rev.* **126**, 2503–2518. (doi:10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2)

6 Kumar, A. & Hoerling, M. P. 2000 Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bull. Am. Meteorol. Soc.* **81**, 255–264. (doi:10.1175/1520-0477(2000)081<0255:AOACMO>2.3.CO;2)

7 Kumar, A., Barnston, A. G. & Hoerling, M. P. 2001 Seasonal predictions, probabilistic verifications, and ensemble size. *J. Clim.* **14**, 1671–1676. (doi:10.1175/1520-0442(2001)014<1671:SPPVAE>2.0.CO;2)

8 Richardson, D. S. 2001 Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.* **127**, 2473–2489. (doi:10.1002/qj.49712757715)

9 Müller, W. A., Appenzeller, C., Doblas-Reyes, F. J. & Liniger, M. A. 2005 A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Clim.* **18**, 1513–1523. (doi:10.1175/JCLI3361.1)

10 Weigel, A. P., Liniger, M. A. & Appenzeller, C. 2007 The discrete Brier and ranked probability skill scores. *Mon. Weather Rev.* **135**, 118–124. (doi:10.1175/MWR3280.1)

11 Ferro, C. A. T. 2007 Comparing probabilistic forecasting systems with the Brier score. *Weather Forecast.* **22**, 1076–1088. (doi:10.1175/WAF1034.1)

12 Ferro, C. A. T., Richardson, D. S. & Weigel, A. P. 2008 On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.* **15**, 19–24. (doi:10.1002/met.45)

13 Tibaldi, S., Palmer, T. N., Branković, Č. & Cusbach, U. 1990 Extended-range predictions with ECMWF models: influence of horizontal resolution on systematic error and forecast skill. *Q. J. R. Meteorol. Soc.* **116**, 835–866. (doi:10.1002/qj.49711649404)

14 Williamson D. L., Kiehl, J. T. & Hack, J. J. 1995 Climate sensitivity of the NCAR Community Climate Model (CCM2) to horizontal resolution. *Clim. Dyn.* **11**, 377–397. (doi:10.1007/BF00209513)

15 Branković, Č. & Gregory, D. 2001 Impact of horizontal resolution on seasonal integrations. *Clim. Dyn.* **18**, 123–143. (doi:10.1007/s003820100165)

16 Roeckner, E., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kornblueh, L., Manzini, E., Schlese, U. & Schulzweida, U. 2006 Sensitivity of simulated climate to horizontal and vertical resolution in the ECHAM5 atmosphere model. *J. Clim.* **19**, 3771–3791. (doi:10.1175/JCLI3824.1)

17 Buizza, R. 2010 Horizontal resolution impact on short- and long-range forecast error. *Q. J. R. Meteorol. Soc.* **136**, 1020–1035. (doi:10.1002/qj.613)

18 Goldstein, M. & Wooff, D. 2007 *Bayes linear statistics*. Chichester, UK: Wiley.

19 Tung, K. K. & Orlando, W. W. 2003 The k^{-3} and $k^{-5/3}$ energy spectrum of atmospheric turbulence: quasigeostrophic two-level model simulation. *J. Atmos. Sci.* **60**, 824–835. (doi:10.1175/1520-0469(2003)060<0824:TKAKES>2.0.CO;2)

20 Lindzen, R. S. & Fox-Rabinovitz, M. S. 1989 Consistent vertical and horizontal resolution. *Mon. Weather Rev.* **117**, 2575–2583. (doi:10.1175/1520-0493(1989)117<2575:CVAHR>2.0.CO;2)

21 European Centre for Medium-Range Weather Forecasts. 2011 ECMWF ERA-40 re-analysis data. NCAS British Atmospheric Data Centre, Harwell Oxford, UK, 28 February 2011. See http://badc.nerc.ac.uk/view/badc.nerc.ac.uk_ATOM_dataent_ECMWF-E40.

22 Chaloner, K. & Verdinelli, I. 1995 Bayesian experimental design: a review. *Stat.Sci.* **10**, 273–304. (doi:10.1214/ss/1177009939)