On the predictability of extremes: does the butterfly effect ever decrease?

A. E. Sterk,[a]*D. B. Stephenson,[b] M. P. Holland,[b] K. R. Mylne[c]

[a] *Johann Bernoulli Institute, University of Groningen, The Netherlands*

[b] *College of Engineering, Mathematics and Physical Sciences, University of Exeter, United Kingdom*

[c] *Met Office, United Kingdom*

*Correspondence to: Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, PO Box 407, 9700 AK Groningen, The Netherlands. E-mail: a.e.sterk@rug.nl

This study investigates whether or not predictability always decreases for more extreme events. Predictability is measured by the Mean Squared Error (MSE), estimated here from the difference of pairs of ensemble forecasts, conditioned on one of the forecast variables (the "pseudo-observation") exceeding a threshold.

Using an exchangeable linear regression model for pairs of forecast variables, we show that the MSE can be decomposed into the sum of three terms: a threshold-independent constant, a mean term that always increases with threshold, and a variance term that can either increase, decrease, or stay constant with threshold. Using the Generalised Pareto Distribution to model wind speed excesses over a threshold, we show that MSE always increases with threshold at sufficiently high threshold. However, MSE can be a decreasing function of threshold at lower thresholds but only if the forecasts have finite upper bounds.

The methods are illustrated by application to daily wind speed forecasts for London made using the 24 member Met Office Global and Regional Ensemble Prediction System from 1 Jan 2009 to 31 May 2011. For this example, the mean term increases faster than the variance term decreases with increasing threshold, and so predictability decreases for more extreme events.

Key Words: extreme events; predictability; mean squared error; ensemble forecasts

Received . . .

1. Introduction

A classical issue in weather prediction is the inherent inaccuracy of forecasts (Tribbia and Anthes 1987). In one of the earliest studies, Thompson (1957) showed that the predictability of large-scale atmospheric flow patterns is sensitive to the uncertainty in the initial condition. Since the seminal work of Lorenz (1963a) it is well known that deterministic systems can be very unpredictable: small errors in the initial condition can grow exponentially in time. This phenomenon, now known colloquially as *the butterfly effect*[*], fundamentally limits the accuracy of long-term weather forecasts and stimulated mathematical research on nonlinear dynamics and chaos theory.

Predictability is often quantified in terms of the growth rate of errors in the initial condition. The earliest studies (Smagorinsky 1963; Mintz 1964; Leith 1965) computed the time needed for small errors in the initial condition to double in size. This so-called doubling time is inversely proportional to the largest Lyapunov exponent provided that the initial error is sufficiently small (Lorenz 2006). However, Lyapunov exponents are independent of the initial condition, which makes them unsuitable as measures for the predictability of specific events such as weather extremes. Finite-time Lyapunov exponents measure the exponential growth rate of nearby trajectories over a finite time and can strongly depend on the initial condition (Nese 1989; Abarbanel *et al.* 1991; Smith *et al.* 1999). Hence, they can be used as a measure of predictability which depends on the event being predicted such as extreme wind speeds at a specific geographical location (Sterk *et al.* 2012).

Exponential growth rates, as measured by Lyapunov exponents, only apply to small errors and short lead times. Harle *et al.* (2006) showed that for longer lead times the error growth follows a power law which systematically depends on the initial size of the error. For even longer lead times the error will saturate. Another way to quantify predictability is to compute the growth of Mean Squared Error (MSE) between pairs of forecasts starting at different initial conditions (Stephenson and Doblas-Reyes 2000). This squared distance approach is particularly useful in ensemble prediction systems which are not equipped with a tangent linear model needed to compute Lyapunov exponents or their finite-time counterparts. Stroe and Royer (1993) compared different growth formulas for the error between twin forecasts starting from nearby initial conditions. In particular, they showed that the growth rate for small errors and saturation rate for large errors have different characteristics.

Despite much research on forecast error growth, there is hardly any published research on how forecast errors behave for more extreme events. It is often assumed that more extreme events are harder to predict, but this belief needs to be rigorously tested. This study makes an initial attempt to address this problem by investigating how predictability depends on the choice of exceedance threshold used to define an event. We propose simple statistical models for representing ensemble forecasts and use them to find out how Mean Squared Error (MSE) of pairs of forecasts depends upon choice of threshold. The methods are applied to daily wind speed forecasts for London made using the 24 member Met Office Global and Regional Ensemble Prediction System from 1 Jan 2009 to 31 May 2011. Section 2 shows how the MSE depends on lead time and wind speed thresholds. Section 3 then uses extreme value theory (the Generalised Pareto Distribution) to quantify the behaviour in the limit of very high thresholds. Section 4 concludes the paper with a discussion of our results.

2. Mean Squared Error growth

2.1. *Ensemble forecast data: exchangeable variables*

We illustrate our methods using a data set of ensemble 10m wind speed forecasts produced by the Met Office Global and Regional Ensemble Prediction System (Bowler *et al.* 2008). In total, we have 866 forecasts at 5 locations in the United Kingdom (see Table 1) for the period 1 January 2009 until 31 May 2011. For each day, the $m = 24$ ensemble members are initialised at 0600 UTC, and predictions are made out to lead times $\tau = 3, 6 \ldots, 54$ hours ahead. For brevity, only results for London will be presented here but these are robust: similar results were found at the other 4 locations.

[*] or *the sea gull effect* when first published: *"When the instability of a uniform flow with respect to infinitesimal perturbations was first suggested as an explanation for the presence of cyclones and anticyclones in the atmosphere, the idea was not universally accepted. One meteorologist remarked that if the theory was correct, one flap of a sea gull's wings would be enough to alter the course of the weather forever."* – Lorenz (1963b)

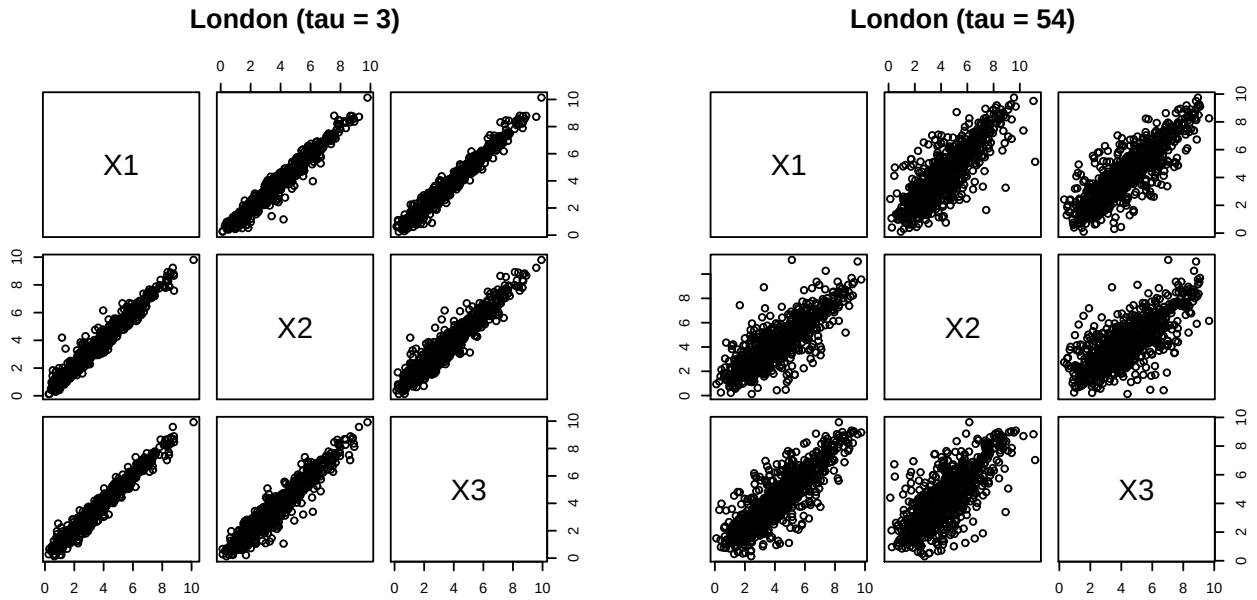


Figure 2. Scatter plots of the pairs (X_i, X_j) for $1 \leq i, j \leq 3$ for the lead times $\tau = 3$ and $\tau = 54$. Scatter plots for the remaining pairs (i, j) are qualitatively similar (not shown).

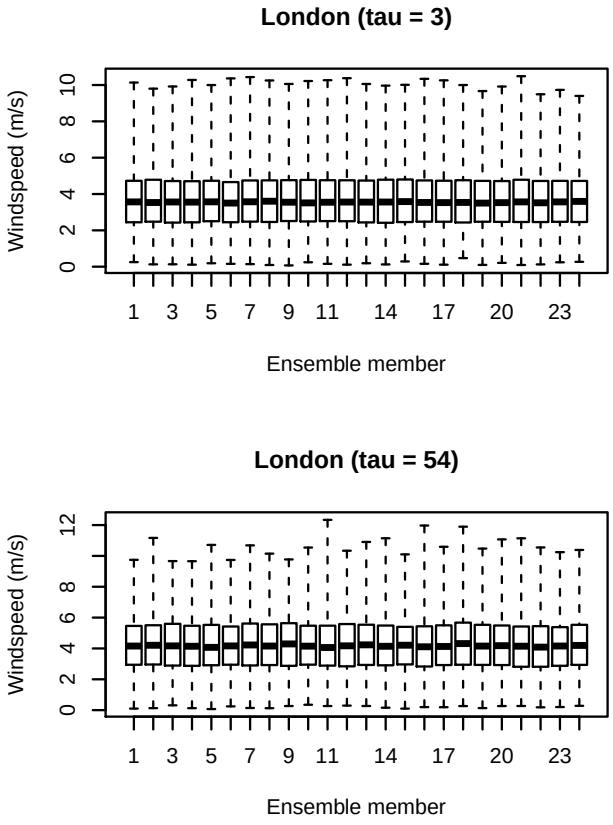


Figure 1. Box plots of the wind speed forecasts for lead times $\tau = 3$ and $\tau = 54$.

Let $X_i = X_i(\tau)$ denote the i-th ensemble member ($i = 1, \ldots, m = 24$) of the wind speed forecasts for a fixed lead time τ. Figure 1 shows box plots of the ensemble members X_i for lead times $\tau = 3$ and $\tau = 54$ hours. The box plots suggest that the X_i have similar distributions. It is therefore reasonable to make the assumption that the ensemble members X_i are exchangeable, which means that $(X_1, \ldots, X_m)$ and $(X_{\sigma(1)}, \ldots, X_{\sigma(m)})$ have the same joint distribution for any permutation σ of the indices $\{1, \ldots, m\}$ (Kallenberg 2005).

Table 1. Five locations in the United Kingdom for which we have wind speed forecasts.

ID	Location
00003134	Bishopton (Glasgow)
00003355	Church Fenton (Leeds)
00003535	Coleshill (Birmingham)
00003628	Filton (Bristol)
00003772	Heathrow Airport (London)

2.2. *Relationship between forecasts*

The scatter plots in Figure 2 suggest a linear association between X_i and X_j, which can be represented using a linear regression model:

$$X_i = \mu + \rho(X_j - \mu) + \varepsilon_{ij}. \quad (1)$$

where μ and ρ are respectively the mean and the covariance of the X_i's. The error ε_{ij} is considered to be zero mean and constant variance noise that is independent of X_j. For notational simplicity, the dependence of μ, ρ, and ε_{ij} on τ is suppressed. However, we assume that μ and ρ and the variance of the noise are identical for all pairs (i, j). Point estimates of the μ and ρ parameters can be

found easily by using an exchangeable extension of ordinary least squares estimation, see Appendix A.1.

Figure 3 shows graphs of μ and ρ versus the lead time τ. The mean μ follows a diurnal cycle and attains its minimum at lead times $\tau = 21$ and $\tau = 45$ hours which corresponds to 0300 UTC since all forecasts are initialised at 0600 UTC. This agrees with the fact that wind speeds are generally lower during the night. The correlation coefficient ρ (and predictability) decreases with lead time τ, related to the growth of forecast errors.

2.3. *Conditional Mean Squared Error*

This predictability study adopts the perfect model scenario: instead of comparing forecasts with observations we will pairwise compare ensemble members with each other. Inclusion of observations would invalidate our assumption of exchangeability because of epistemic uncertainties in how the forecast model represents the real world (e.g. presence of model errors, biases, etc.). One member X_i is assumed to be the "forecast" and another member X_j (with $j \neq i$) is assumed to be a "pseudo-observation". As our measure of predictability, we consider the conditional Mean Squared Error averaged over all distinct pairs of ensemble members:

$$\mathrm{MSE}(\tau, u) = \frac{1}{m(m-1)} \sum_{i \neq j} \mathrm{E}\left([X_i - X_j]^2 \mid X_j > u\right) \qquad (2)$$

where the threshold u is taken here to be an empirical quantile of all the X_j's.

A key question is how the MSE depends on both the lead time τ and the threshold u. The top panel of Figure 4 shows a graph of the conditional MSE as a function of the lead time τ for fixed thresholds u. For $u = 0$ the MSE is an increasing function of τ. For $u > 0$, however, the MSE has local maxima near $\tau = 21$ and $\tau = 45$ (0300 UTC) and a local minimum near $\tau = 33$ (1500 UTC). Thus, errors are larger during the night than during the day. The bottom panel of Figure 4 shows a graph of the conditional MSE as a function of the threshold u for fixed lead times τ. The MSE increases with u, but the growth rates are larger for $\tau = 21$ and $\tau = 45$ than for $\tau = 33$.

But can we understand the reasons for this behaviour in MSE? From equation (1) it is shown in Appendix A.2 that the conditional MSE between a pair of forecasts, $\mathrm{E}\left([X_i - X_j]^2 \mid X_j > u\right)$, is the sum of three terms:

$$\begin{aligned}(1-\rho^2)\mathrm{Var}\left(X_j\right) + (1-\rho)^2 \mathrm{E}\left(X_j - \mu \mid X_j > u\right)^2 \\ + (1-\rho)^2 \mathrm{Var}\left(X_j \mid X_j > u\right)\end{aligned} \qquad (3)$$

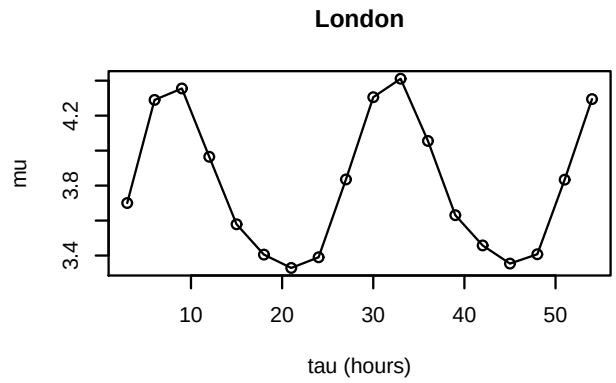
All terms contain the factor $1-\rho$ or $1-\rho^2$ and so vanish for perfectly dependent forecasts with $\rho = 1$ (e.g. at initial time $\tau = 0$). The first term involves the unconditional variance of the forecast and so is independent of threshold u. The second term is based on the squared mean of the difference of the forecasts from the mean and so increases monotonically for large threshold $u > \mu$. We will refer to this term as the squared mean excess. The third term is based on the conditional variance of the forecasts, which can either increase or decrease with increasing threshold. If X_j has a finite upper end point, then this term will tend to zero as threshold u approaches the end point.

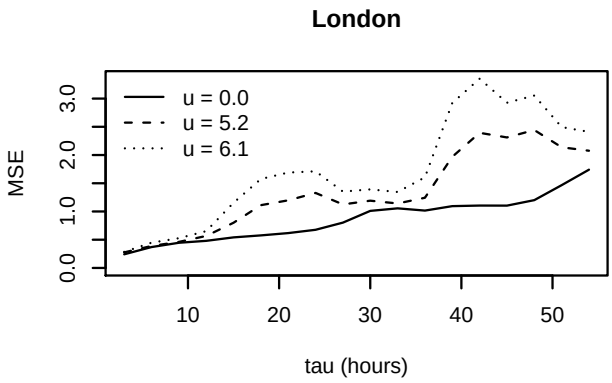
Ensemble summaries can be obtained by averaging each term over all distinct pairs of ensemble forecasts. For example, for the first term in equation (3) we obtain

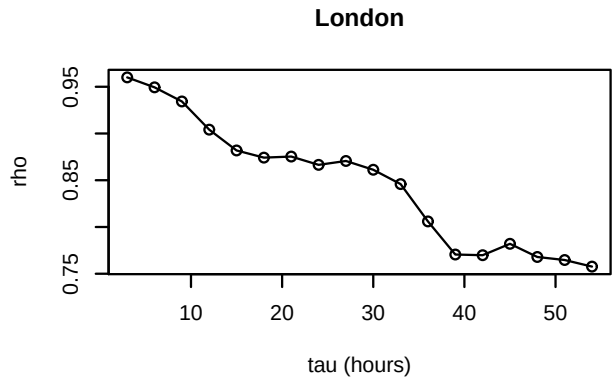
$$\frac{1}{m(m-1)} \sum_{i \neq j} (1-\rho^2)\mathrm{Var}\left(X_j\right) = \frac{1}{m} \sum_{j=1}^{m} (1-\rho^2)\mathrm{Var}\left(X_j\right).$$

Analogous expressions hold for the second and third term. To test the validity of the decomposition, Figure 5 shows the difference between the mean square error (2) and the sum of the ensemble averages of the three terms for threshold $u = 0$. Because of sampling variations in the estimates of the components, the difference is not identically zero, however, it is only a small fraction of the MSE and its variation with lead time and threshold shown in Figure 4.

Figure 6 shows the averaged conditional variance and squared excess as functions of lead time τ and threshold u. The conditional variance term increases with τ for fixed u, but decreases with u for fixed τ. The squared mean excess term has local maxima near $\tau = 21$ and $\tau = 45$ and a local minimum near $\tau = 33$. For fixed lead time, it is an increasing function of u as expected. The mean excess term increases faster than the variance term decreases and so the conditional MSE increases for higher thresholds. In







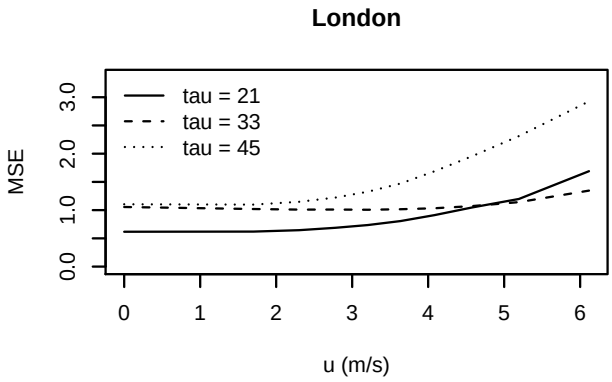


Figure 3. The coefficients μ and ρ in equation (1) estimated by the least squares method described in Appendix A.1 and plotted as a function of the lead time τ. Note that the mean μ follows a diurnal cycle and attains its minimum for $\tau = 21$ and $\tau = 45$ (0300 UTC).

Figure 4. Top panel: averaged conditional mean square error as a function of the lead time τ for three thresholds u. The thresholds $u > 0$ correspond to the 80th and 90th percentile of the distribution of wind speed forecasts. Bottom panel: averaged conditional mean square error as a function of the threshold u for different lead times τ.

other words, predictability decreases for extremes in this example because of extreme forecasts being further from the mean.

3. What happens at high thresholds?

The decomposition (3) shows that threshold dependence is related to how the 1st and 2nd moments of forecast excesses vary with threshold (note that for fixed τ the coefficient ρ is a threshold-independent constant). This section will model wind speed excesses over a threshold using extreme value theory to understand the possible behaviour at high thresholds.

3.1. Modelling excesses using the Generalised Pareto Distribution

Under widely applicable conditions (Balkema and de Haan 1974; Pickands 1975) the distribution of excesses $X - u$ is well-approximated by the Generalised Pareto Distribution (GPD)

$$G_{\xi,\beta(u)}(x) := \begin{cases} 1 - \left(1 + \frac{\xi x}{\beta(u)}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta(u)}\right) & \text{if } \xi = 0 \end{cases} \quad (4)$$

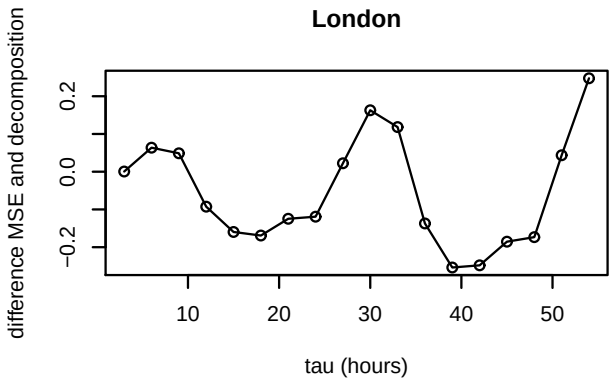


Figure 5. The difference between the mean squared error (2) and the decomposition terms in the right hand side of (3) as a function of the lead time τ.

for X exceeding sufficiently high thresholds. The scale parameter is a linear function of threshold, $\beta(u) = \sigma + \xi(u - \lambda)$, which either increases, stays constant, or decreases depending on whether the shape parameter ξ is positive, zero, or negative, respectively (Coles 2001). For $\xi < 0$, the distribution has a finite upper end point at $x_F = \lambda - \sigma/\xi$, but is unbounded when the shape parameter is non-negative. The distribution of excesses above any high threshold is completely determined by just three

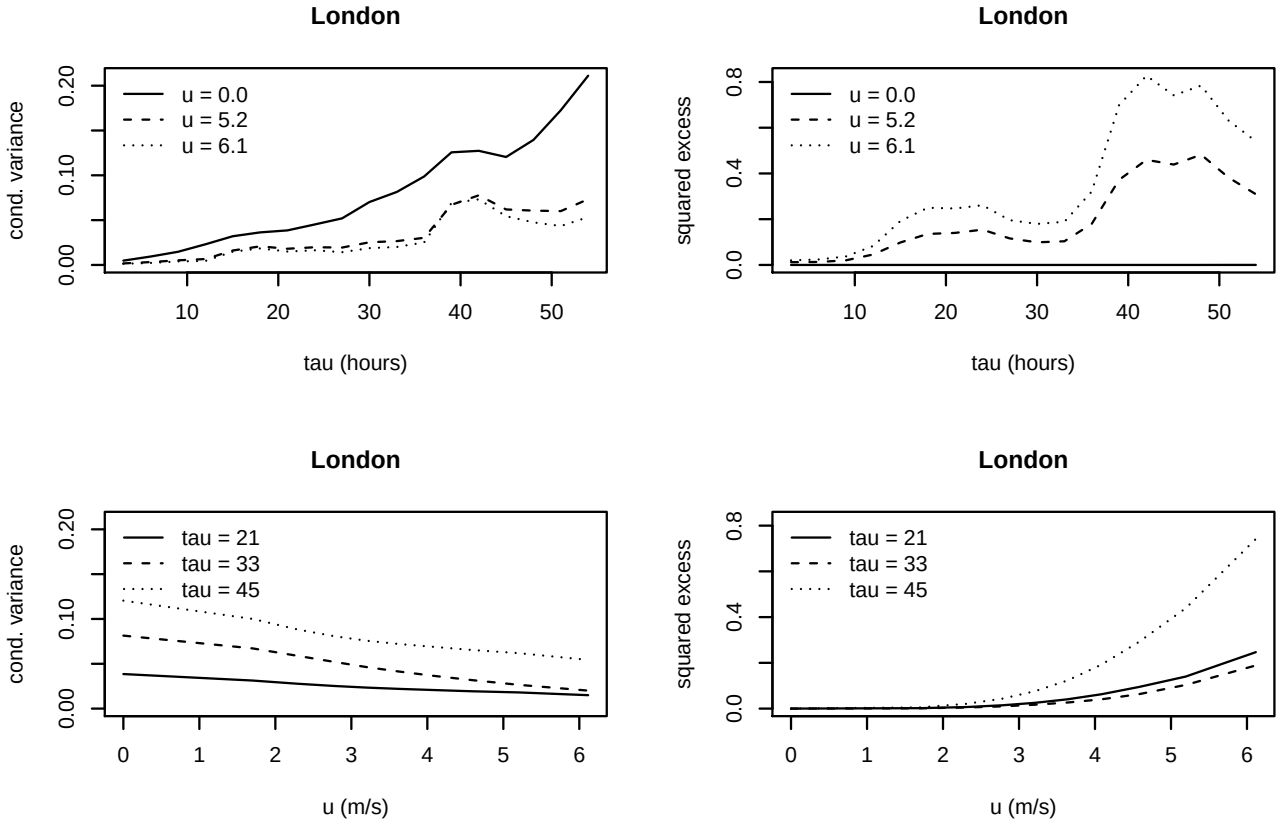


Figure 6. Top panels: the conditional variance $(\rho-1)^2\mathrm{Var}\,(X_j \mid X_j > u)$ and squared excess $(\rho-1)^2\mathrm{E}\,(X_j-\mu \mid X_j > u)$ (both averaged over all ensemble members) as a function of the lead time τ for fixed thresholds u. Bottom panels: the same quantities as function of the threshold u for fixed lead times.

parameters: (λ, σ, ξ). Note that the location parameter for the tail, λ, should not be confused with the central location parameter $\mu = \mathrm{E}\,(X)$. The GPD can now be used to find the components in the MSE decomposition.

3.2. *Conditional MSE components*

The 1st and 2nd moments of the GPD (see Appendix A.3) can be used to find expressions for the mean and variance components[†] of the MSE decomposition:

$$\begin{aligned} E(u) &= \mathrm{E}\,(X-\mu \mid X>u)^2 \\ &= \left(\mathrm{E}\,(X-u \mid X>u) + u - \mu\right)^2 \\ &= \left(\frac{\beta+(1-\xi)(u-\mu)}{1-\xi}\right)^2, \end{aligned}$$

and

$$\begin{aligned} V(u) &= \mathrm{Var}\,(X \mid X>u) \\ &= \mathrm{Var}\,(X-u \mid X>u) \\ &= \frac{\beta^2}{(1-\xi)^2(1-2\xi)}. \end{aligned}$$

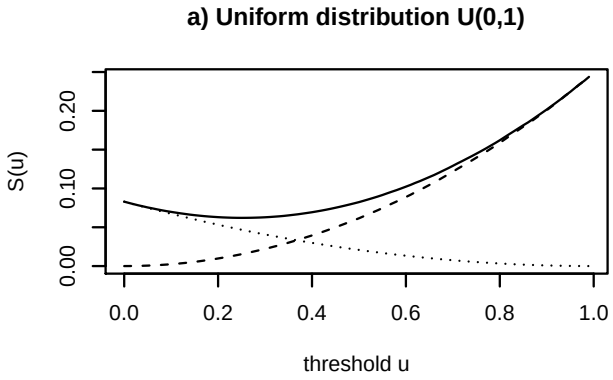
These expressions reveal that the mean component $E(u)$ increases monotonically for $u > \mu + \xi(\lambda-\mu) - \sigma$ regardless of the values of the other parameters. For large u, $E(u) \sim u^2$. In contrast, the variance component $V(u)$ either increases, stays constant, or decreases with u depending upon whether $\xi > 0$, $\xi = 0$, or $\xi < 0$, respectively. Hence, for $\xi \geq 0$, the sum of the two terms $S(u) = E(u) + V(u)$ is a strictly monotonically increasing function of threshold: the butterfly effect increases for higher thresholds. For $\xi < 0$, decreasing $V(u)$ compensates increasing $E(u)$ to create a minimum $S(u)$ at

$$u_{\min} = \frac{\mu(1-2\xi) + \xi\lambda - \sigma}{1-\xi}. \tag{5}$$

[†] without the ρ dependent multipliers

As u tends to x_F, $V(u)$ decreases at a slower rate than $E(u)$ increases and so $S(u)$ increases as the threshold approaches the end point. In summary, the MSE always increases with threshold at sufficiently high threshold, although it can be a decreasing function of threshold at lower thresholds but only if the forecasts have finite upper bounds ($\xi < 0$).

It is instructive to consider some special cases. Figure 7 shows the behaviour of the components when the forecasts are simulated from uniform and normal distributions. The results agree with what is expected from the above GPD theory. The $U(0,1)$ uniform distribution can easily be shown to have $\lambda = 0, \sigma = 1, \xi = -1$ and $\mu = 0.5$ and so one expects an increasing mean component, a decreasing variance component and a minimum at $u_{\min} = 0.25$, as can be seen in panel a. The Normal distribution can be shown to have $\xi = 0$ and so one expects to see an increasing mean component and a constant variance component, which is occurring for larger threshold in panel b.



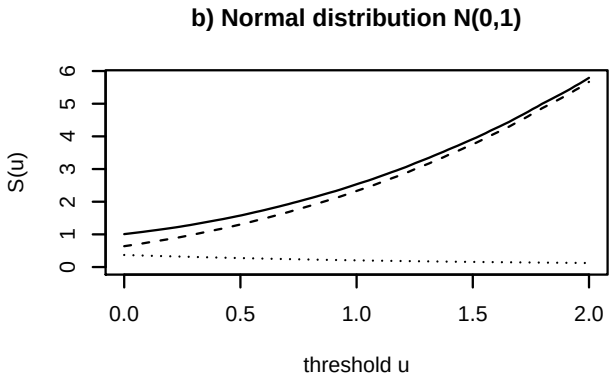
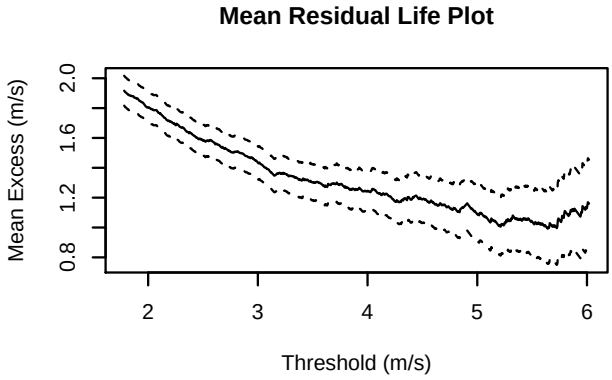
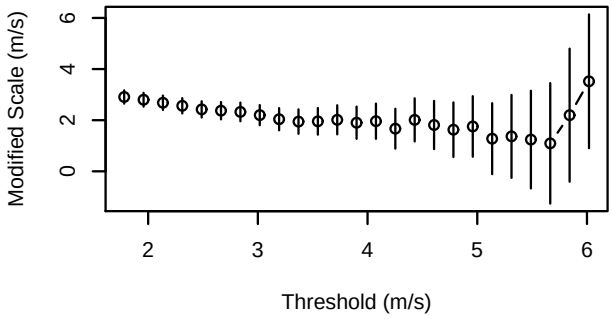


Figure 7. Examples of threshold dependence in $S(u) = E(u) + V(u)$ (solid line), $E(u)$ (dashed line), and $V(u)$ (dotted line) for forecasts that are a) uniformly distributed $X \sim U(0,1)$, and b) Normally distributed $X \sim N(0,1)$. Curves were obtained by simulating 10^5 values from the respective distributions.





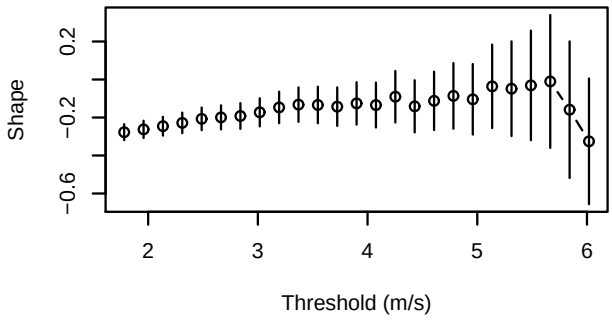


Figure 8. Threshold selection plots for the variable X_1 and lead time $\tau = 21$.

3.3. *GPD analysis of the wind speed data*

Can we use GPD fits to understand the behaviour seen earlier in our wind speed forecast example?

The parameter μ can be estimated from the linear model (1) using the least squares procedure described in Section A.1. The parameters (λ, σ, ξ) can be estimated by first selecting a fixed threshold u_0 and then fit the exceedances $X - u_0 > 0$ to a Poisson process (Coles 2001). This procedure is implemented in the `fpot` command, which is part of the R package `evd` (Stephenson 2002).

The selection of the threshold u_0 depends on two contradicting requirements. On the one hand u_0 should be as small as possible in

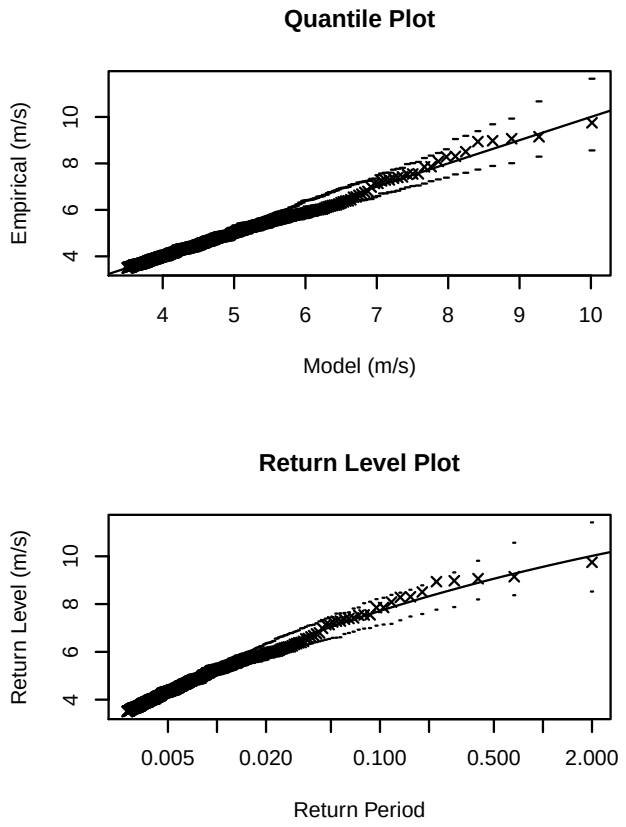


Figure 9. Diagnostic plots obtained from estimating the parameters (λ, σ, ξ) for the lead time $\tau = 21$. The parameters were estimated using 866 daily data values and so the return period, which is the reciprocal of the exceedance probability, is in units of 866 days.

order to have many exceedances $X - u_0 > 0$ thereby decreasing the variance of the estimates for (λ, σ, ξ). On the other hand u_0 should be large in order to ensure that the asymptotic GPD is a valid model for the exceedances $X - u_0 > 0$. If $G_{\beta,\xi}$ is a valid model for $X \mid X > u_0$ then the mean excess function $\mathrm{E}(X - u \mid X > u)$ for $u > u_0$ is linear in u. Moreover, the shape parameter ξ and the modified scale parameter $\beta - \xi u$ should be approximately linear in u. We adopt these criteria for selecting u_0.

For example, Figure 8 shows threshold selection plots for the variable X_1 of the London wind speed forecasts for the lead time $\tau = 21$. Based on these plots we set $u_0 = 3.5$. Subsequently, the parameters (λ, σ, ξ) can be estimated. The diagnostic plots in Figure 9 indicate that the fit is fairly good: the quantile plot is nearly diagonal and the empirical return levels agree well with the fitted return levels. This procedure has been repeated for the lead times $\tau = 33$ and $\tau = 45$, see Table 2 for the results.

The wind speed forecasts have a negative shape parameter which implies that the function $S(u)$ might have a minimum. However, the value $u_{\min}$ given in Table 2, computed from the GPD parameter estimates using equation (5), is far below the average of the wind speed forecasts (see Figure 1). At such low values, the GPD is no longer a good approximation for excesses, which may help to explain why no such minimum can be seen in Figure 4.

Table 2. Point process estimates for the parameters (λ, σ, ξ) for different lead times τ. The standard errors are listed between parentheses.

τ	u_0	$\widehat{\lambda}$	$\widehat{\sigma}$	$\widehat{\xi}$	$u_{\min}$
21	3.5	9.56 (0.50)	0.69 (0.16)	-0.13 (0.04)	2.00
33	4.8	9.74 (0.25)	0.29 (0.06)	-0.32 (0.04)	2.90
45	3.5	10.17 (0.64)	0.91 (0.22)	-0.07 (0.05)	2.04

4. Conclusion

This study has investigated whether or not predictability can increase for more extreme events. Predictability is measured here by the Mean Squared Error (MSE) between pairs of ensemble forecasts, conditioned on one of the forecast variables (the "pseudo-observation") exceeding a threshold. Assuming an exchangeable linear regression model for pairs of forecast variables, we have found that the MSE can be decomposed into the sum of three terms: a threshold-independent constant, a mean term that always increases with threshold, and a variance term that can either increase, decrease, or stay constant with threshold. Extreme value theory has revealed that MSE always increases with threshold at sufficiently high threshold. Furthermore, MSE can be a decreasing function of threshold at lower thresholds but only if the forecasts have finite upper bounds.

In summary, our modelling approach suggests that the butterfly effect generally increases for more extreme events. This is predominantly due to the average distance between pairs of forecasts increasing if one of the forecasts is constrained to be in a more extreme edge of state space. For bounded forecasts, a reduction in variance can help to compensate this but is insufficient to do so as one approaches the edge of the attractor.

The methods have been applied to daily wind speed forecasts for London made using the 24 member Met Office Global and Regional Ensemble Prediction System from 1 Jan 2009 to 31 May 2011. The MSE has a local minimum during the nights and a local maximum during the afternoons. The MSE is found to increase

monotonically with threshold, especially so during the night. The mean component increases and the variance component decreases with threshold as to be expected for variables that are estimated to be bounded from above (negative GPD shape parameter).

Our approach relies on several assumptions, which appear reasonable but should be investigated in more detail in future work. For example, we have modelled dependency between forecasts using a simple linear model with constant variance noise. This appears to work well for our data example but more complex models may be appropriate for other data sets. It is possible that there could be more predictability at higher thresholds if one were to allow for extremal dependency between forecasts using bivariate copula approaches (Stephenson *et al.* 2008).

This study has addressed predictability by focussing on mean squared error—a measure of skill for deterministic forecasts. In future studies, it would be of interest to extend the approach to address probability forecasts e.g. by using modified decompositions to decompose the Brier score, which for frequency of exceedance above a threshold may be expressed as a mean squared error of binary forecasts. One other caveat is that this study has measured predictability using an absolute score but one may find alternative conclusions if one were to assess predictability by measuring skill against no-skill forecasts (e.g. using the MSE skill score against either persistence or climatological forecasts).

As with all previous perfect model approaches, we have made the assumption that observations are exchangeable with ensemble forecasts. When this is not the case and the forecasts contain more unpredictable noise than the observations (e.g. Eade *et al.* 2014), then real world observations will give smaller MSE than the MSE of pairs of forecasts. However, we suspect that the broad conclusions about increase of MSE with threshold may still generally apply to unreliable forecasts that are co-exchangeable with the observations.

Although we have illustrated our approach using weather forecasts of wind speeds, the methodology is generic and so could be applied to other predicted quantities of interest such as temperatures or river flows.

Acknowledgements

This research has been carried out within the project "PREDEX: PREdictability of EXtreme weather events", funded by Complexity-NET (www.complexitynet.eu). The authors gratefully acknowledge support by the UK and Dutch funding agencies involved in Complexity-NET: the Engineering and Physical Sciences Research Council (EPSRC) and the Netherlands Organisation for Scientific Research (NWO).

The authors wish to thank Dr Renato Vitolo for his contributions to the project, Dr Lisa Murray (Met Office) for preparing the data and Dr Jan Barkmeijer (Royal Netherlands Meteorological Institute) for useful discussions. The authors also wish to thank the reviewers for their insightful comments that have helped improve the manuscript.

A. Appendix

A.1. Ordinary least squares estimation for exchangeable variables

The coefficients μ and ρ in equation (1) are estimated by a least-squares method as follows. First, rewrite equation (1) as

$$X_i = a + bX_j + \varepsilon_{ij}$$

where $a = (1-\rho)\mu$ and $b = \rho$. The coefficients a and b are obtained by minimizing the expression

$$\mathrm{SSE} = \frac{1}{N}\sum \left(\mu + \rho(X_j - \mu) - X_i\right)^2,$$

where the summation ranges over all 866 days and all pairs $i \neq j$ so that $N = 866 \cdot m(m-1)$. Setting the partial derivatives of the SSE with respect to μ and ρ to zero gives the equations

$$Na + b\sum X_j = \sum X_i$$
$$a\sum X_j + b\sum X_j^2 = \sum X_i X_j$$

from which a and b are readily solved. Finally, $\rho = b$ and $\mu = a/(1-b)$. This computation is repeated for each lead time τ.

A.2. *Decomposition of the MSE*

Recall the formula for conditional variance (Ross 1997):

$$\begin{aligned}\mathrm{Var}\,(X \mid Y) &= \mathrm{E}\,([X - \mathrm{E}\,(X \mid Y)]^2 \mid Y) \\ &= \mathrm{E}\,(X^2 \mid Y) - \mathrm{E}\,(X \mid Y)^2.\end{aligned}$$

Using equation (1), the MSE of $X_i - X_j$ conditional on $X_j > u$ can be written as

$$\begin{aligned}&\mathrm{E}\,([X_i - X_j]^2 \mid X_j > u) \\ &\quad = \mathrm{Var}\,(X_i - X_j \mid X_j > u) + \mathrm{E}\,(X_i - X_j \mid X_j > u)^2 \\ &\quad = \mathrm{Var}\,((1-\rho)\mu + (\rho-1)X_j + \varepsilon_{ij} \mid X_j > u) \\ &\qquad + \mathrm{E}\,((1-\rho)\mu + (\rho-1)X_j + \varepsilon_{ij} \mid X_j > u)^2 \\ &\quad = \mathrm{Var}\,((\rho-1)X_j + \varepsilon_{ij} \mid X_j > u) \\ &\qquad + \left\{(\rho-1)\mathrm{E}\,(X_j - \mu \mid X_j > u) + \mathrm{E}\,(\varepsilon_{ij} \mid X_j > u)\right\}^2.\end{aligned}$$

The latter expression simplifies under the additional assumption that the error ε_{ij} is independent of X_j. This implies that

$$\mathrm{E}\,(\varepsilon_{ij} \mid X_j > u) = \mathrm{E}\,(\varepsilon_{ij}) = 0$$

and

$$\begin{aligned}&\mathrm{Var}\,((\rho-1)X_j + \varepsilon_{ij} \mid X_j > u) \\ &\qquad = (\rho-1)^2\mathrm{Var}\,(X_j \mid X_j > u) + \mathrm{Var}\,(\varepsilon_{ij} \mid X_j > u) \\ &\qquad = (\rho-1)^2\mathrm{Var}\,(X_j \mid X_j > u) + \mathrm{Var}\,(\varepsilon_{ij}).\end{aligned}$$

Note that equation (1) implies that

$$\mathrm{Var}\,(X_i) = \rho^2\mathrm{Var}\,(X_j) + \mathrm{Var}\,(\varepsilon_{ij}).$$

Assuming that $\mathrm{Var}\,(X_i) = \mathrm{Var}\,(X_j)$ gives

$$\mathrm{Var}\,(\varepsilon_{ij}) = (1-\rho^2)\mathrm{Var}\,(X_i) = (1-\rho^2)\mathrm{Var}\,(X_j).$$

Hence, the final decomposition of the conditional mean square error reads as

$$\begin{aligned}&\mathrm{E}\,([X_i - X_j]^2 \mid X_j > u) \\ &\quad = (1-\rho^2)\mathrm{Var}\,(X_j) + (\rho-1)^2\mathrm{Var}\,(X_j \mid X_j > u) \\ &\qquad + (\rho-1)^2\mathrm{E}\,(X_j - \mu \mid X_j > u)^2.\end{aligned}$$

A.3. *Moments of the generalised Pareto distribution*

Let X have a generalised Pareto distribution with scale parameter β and tail index ξ. For any natural number r satisfying $\xi < 1/r$ we have

$$\mathrm{E}\,(X^r) = \frac{\beta^r}{\xi^{r+1}} \cdot \frac{\Gamma(\xi^{-1} - r)}{\Gamma(\xi^{-1} + 1)} r!,$$

where Γ is the Gamma function (Embrechts *et al.* 1997). Note that the expectation of X^r does not exist when $1/r \leq \xi$. For $\xi < \frac{1}{2}$ we have

$$\mathrm{E}\,(X) = \frac{\beta}{1-\xi} \quad \text{and} \quad \mathrm{E}\,(X^2) = \frac{2\beta^2}{(1-\xi)(1-2\xi)}.$$

Hence, the variance is given by

$$\mathrm{Var}\,(X) = \frac{\beta^2}{(1-\xi)^2(1-2\xi)}.$$

Furthermore, the mean excess function is given by

$$\mathrm{E}\,(X - u \mid X > u) = \frac{\beta + \xi u}{1-\xi}.$$

References

Abarbanel H, Brown R, Kennel M. 1991. Variation of Lyapunov exponents on a strange attractor. *Journal of Nonlinear Science* **1**: 175–199.

Balkema A, de Haan L. 1974. Residual life time at great age. *The Annals of Probability* **2**: 792–804.

Bowler N, Arribas A, Mylne K, Robertson K, Beare S. 2008. The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **34**: 703–722.

Coles S. 2001. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics, Springer: New York.

Eade R, Smith D, Scaife A, Wallace E, Dunstone N, Hermanson L, Robinson N. 2014. Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters* **41**: 5620–5628.

Embrechts P, Klüppelberg C, Mikosch T. 1997. *Modelling extremal events*, *Applications of Mathematics: Stochastic Modelling and Applied Probability*, vol. 33. Springer–Verlag.

Harle M, Kwasniok F, Feudel U. 2006. Growth of finite errors in ensemble prediction. *Nonlinear Processes in Geophysics* **13**: 167–176.

Kallenberg O. 2005. *Probabilistic symmetries and invariance principles*. Probability and Its Applications, Springer–Verlag.

Leith C. 1965. Numerical simulation of the earth's atmosphere. In: *Methods in Computational Physics*, Alder B, Fernbach S, Rotenberg M (eds), Academic Press, pp. 1–28.

Lorenz E. 1963a. Deterministic nonperiodic flow. *Journal of the Atmospheric Science* **20**: 130–141.

Lorenz E. 1963b. The predictability of hydrodynamic flow. *Transactions of the New York Academy of Sciences* **25**: 409–432.

Lorenz EN. 2006. Predictability—A problem partly solved. In: *Predictability of Weather and Climate*, Palmer TN, Hagedorn R (eds), Cambridge University Press, pp. 40–58.

Mintz Y. 1964. Very long term global integration of the primitive equations of atmospheric motion. Technical Report Tech. note No. 66.

Nese J. 1989. Quantifying local predictability in phase space. *Physica D* **35**: 237–250.

Pickands J. 1975. Statistical inference using extreme order statistics. *The Annals of Statistics* **3**: 119–131.

Ross S. 1997. *A first course in probability*. Prentice Hall: New Jersey, 5th edn.

Smagorinsky J. 1963. General circulation experiments with the primitive equations. I. the basic experiment. *Monthly Weather Review* **91**: 99–164.

Smith L, Ziehmann C, Fraedrich K. 1999. Uncertainty dynamics and predictability in chaotic systems. *Quarterly Journal of the Royal Meteorological Society* **125**: 2855–2886.

Stephenson A. 2002. evd: Extreme value distributions. *R News* **2**(2): 0, URL http://CRAN.R-project.org/doc/Rnews/.

Stephenson D, Casati B, Ferro C, Wilson C. 2008. The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications* **15**: 41–50.

Stephenson D, Doblas-Reyes F. 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus A* **52**: 300–322.

Sterk A, Holland M, Rabassa P, Broer H, Vitolo R. 2012. Predictability of extreme values in geophysical models. *Nonlinear Processes in Geophysics* **19**: 529–539.

Stroe R, Royer J. 1993. Comparison of different error growth formulas and predictability estimation in numerical extended-range forecasts. *Annales Geophysicae* **11**: 296–316.

Thompson P. 1957. Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus* **9**: 275–295.

Tribbia J, Anthes R. 1987. Scientific basis of modern weather prediction. *Science* **31**: 493–499.

