TECHNICAL MEMORANDUM

754

The Forecast Skill Horizon

Roberto Buizza and Martin Leutbecher

Research Department

June 2015

Accepted for publication in the
Q. J. Roy. Meteorol. Soc., 29 June 2015.

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

© Copyright 2015

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

Numerical weather prediction has seen, in the past 25 years, a shift from a 'deterministic' approach, based on single numerical integrations, to a probabilistic one, with ensembles of numerical integrations used to estimate the probability distribution function of forecast states. This shift to a probabilistic approach enabled a better extraction of predictive signals at longer lead times and provided a meaningful framework for extending the forecast length beyond 10 days. In this work, the limit of predictive skill is assessed for ECMWF monthly ensemble forecasts at different spatial and temporal scales. The forecast skill horizon is defined as the lead time when the ensemble ceases to be more skilful than a climatological distribution, using a continuous ranked probability score as metric. Results based on 32-day ensemble forecasts indicate that the forecast skill horizon is sensitive to the spatial and temporal scale of the predicted phenomena, to the variable considered and the area analysed. On average over 1 year of forecast, the forecast skill horizon for instantaneous, grid-point fields is between 16–23 days, while it is considerably longer for time- and spatial-average fields. Forecast skill horizons longer than the 2 weeks that were thought to be the limit are now achievable thanks to major advances in numerical weather prediction. More specifically, they are possible because forecasts are now framed in probabilistic terms, with a probability distribution estimated using ensembles generated using forecast models that include more components (e.g. a dynamical ocean and ocean waves) and more faithfully represent processes. Moreover, the forecasts start from more accurate initial conditions constructed using better data-assimilation methods and more observational data.

1 Introduction

Two key aspects of operational weather prediction have changed substantially in the last 25 years, the adoption of a probabilistic approach and the extension of the forecast length beyond two weeks. These two aspects are linked: in fact, the shift from a 'deterministic' approach, based on single numerical integrations, to a probabilistic approach, with ensembles of numerical integrations used to estimate the probability distribution function of forecast states, contributed to the extension of the forecast length. More precisely, it has made it possible to extract predictive signals in the extended forecast range.

The forecast length extension to beyond two weeks may seem to contradict early results of the 1960s-1980s that suggested that the range of predictive skill is limited to about 2 weeks. It is thus relevant to ask whether it is still correct to think of two weeks as the predictive skill limit of numerical weather prediction, or whether advances in modelling and data assimilation, and the adoption of ensemble-based probabilistic approaches to numerical weather prediction have led to predictive skill beyond 2 weeks.

This work addresses this question by determining the forecast lead time at which the probabilistic skill of an ensemble forecast ceases to be higher than that of the climatological distribution. This forecast lead time will be referred to as the forecast skill horizon here. It will be quantified for ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF). In the forthcoming sub-sections of this Introduction, early estimates of the range of predictive skill and the reason why the forecast skill horizon should be estimated using ensembles are discussed. Then, some of the most recent estimates of the predictive skill limit are briefly reviewed, and the concept of the forecast skill horizon is introduced.

1.1 Early estimates of the range of predictive skill

Lorenz (1982) concisely summarizes early predictability studies that aimed to determine the range of predictive skill. The studies reported by Charney et al (1966) and subsequently Smagorinsky (1969) estimated error doubling times of 5 days and 3 days, respectively. With estimates of the magnitude of initial error at the time, this implied a limit of predictability of about two weeks. One might have concluded that further reduction of the initial error would allow extending the range of skilful forecasts further and further — at least in principle. However, the model estimates did not include all scales of motion and neglected the fact that errors grow the faster the smaller the spatial scale. This aspect was investigated quantitatively for the first time by Lorenz (1969a) using a two-dimensional vorticity equation. He showed that '… there may be some systems where a reduction of the initial error will not increase the range of predictability at all.' Whether this happens depends on the steepness of the spectrum of kinetic energy in the flow. For spectra shallower than K^{-3} with K being the total wavenumber, the predictability is limited to a finite time as long as there is some initial error regardless how small (Rotunno and Snyder 2008). For one of the configurations he studied, Lorenz estimated that this finite range of predictability was about two weeks, more precisely 16.8 days.

Lorenz (1969a) mentions what will later be known as the butterfly effect, although the paper talks about sea gulls: '.. if the theory of atmospheric instability were correct, one flap of a sea gull's wings would forever change the future course of the weather'. Rotunno and Snyder (2008) and Durran and Gingrich (2014) further pursued the error growth model that Lorenz (1969a) introduced (see discussion in Section 4). In another work published the same year as an MIT technical report, Lorenz (1969c) discussed the possible implications of these predictability studies for numerical weather prediction, and stated that '.. Certainly they offer little hope for those who would extend the two-week goal to one month or more. They are not especially reassuring even for two-week forecasting … '. Later, Lorenz (1982) examined the predictability using ECMWF forecasts in a 100 day period. He estimated error doubling times of about 2 days for small errors and concluded that 'better-than-guesswork forecasts of instantaneous weather patterns nearly two weeks in advance seem possible'.

Dalcher and Kalnay (1987) extended Lorenz (1982) model for error growth, and estimated the theoretical limit of dynamical predictability for different scales, identified by their total wave number. Looking at 10-day forecasts from the ECMWF model, they concluded that the limit is longer in winter than in summer, and that in winter for the longer waves it is longer than 10 days, while in summer it is about 10 days. Later on, Simmons and Hollingsworth (2002) also revisited the error growth model of Lorenz (1982) and estimated error doubling times as low as 1.5 d for 500 hPa geopotential in the Northern Hemisphere extra-tropics in boreal winter. Despite the even shorter doubling times they found evidence for predictive skill up to about 3 weeks after correcting for the forecast model bias. Thus, what was deemed impossible about 30 years earlier, now seemed achievable, and indeed the new century (2000s) saw the beginning of operational monthly weather prediction.

1.2 The range of predictive skill in the context of ensemble forecasts

As one approaches the limit of predictability, the error of single forecasts increases and reaches eventually twice the variance of the climatology — assuming that the model has no bias in the first and

second moments. The error of the single forecast will on average exceed the error of the climatological mean in magnitude before this happens. The ability to make decisions based on such a single forecast that are better than climatologically-based ones, is limited especially as the limit of predictability is approached. In contrast, an ensemble can still carry more useful information also close to the limit of predictability. Therefore, we propose to base the forecast skill horizon on the accuracy of ensemble forecasts. They provide a more complete estimate of the future forecast states, since this estimate includes not only the most likely scenario but also a confidence interval, expressed in terms of a range of possible scenarios, or probabilities of occurrence of events of interest (Palmer et al 2007, Buizza 2008).

To date, ensembles provide the only feasible way to provide a dynamical, and thus flow-dependent, estimate of the initial-time and forecast probability distribution of the atmospheric state. This has been possible since November and December 1992, when operational, global, medium-range ensemble prediction started at the National Centers for Environmental Prediction (NCEP, Toth & Kalnay 1993 and 1997) and the European Centre for Medium-Range Weather Forecast (ECMWF, Palmer et al 1993, Buizza & Palmer 1995, Molteni et al 1996). ECMWF and NCEP were followed in 1995 by the Meteorological Service of Canada (MSC, Houtekamer et al 1996). Following these three examples, six other centres started producing global ensemble forecasts in the following years. Today, eight meteorological centres issue daily medium-range global ensemble forecasts and share their data as part of the TIGGE project (see Buizza 2014 for a recent review of the TIGGE medium-range global ensembles).

1.3 The start of monthly prediction and evidence of long-range predictive skill

ECMWF has been at the forefront of dynamical extended range forecasting since the mid 1980's, when experimentation on ensemble forecasting for the monthly time scale started (Molteni et al 1986, Brankovic et al 1990). In 2002, ECMWF was one of the first operational centres to run an experimental monthly ensemble, based on coupled ocean-atmosphere integrations (Vitart 2004), a system that became operational in October 2004. Today, several meteorological centres produce operational sub-seasonal forecasts. In some cases (e.g. at the Japanese Meteorological Agency, JMA), a stand-alone forecasting system is used to produce monthly forecasts. Other centres use their seasonal forecasting system to produce sub-seasonal forecasts by increasing the frequency of seasonal forecast integrations (e.g. at the Centre for Australian Weather and Climate research, CAWCR) or the number of ensemble members (e.g. at the UK Meteorological Office). At ECMWF, following Tracton and Kalnay (1993), since March 2008 monthly forecasts are produced using a variable resolution approach (Buizza et al 2007) an extension of the medium-range ensemble, thus producing 'seamless' probabilistic forecasts from day 1 to week 4 using the same coupled ocean-atmosphere model.

Looking at the ECMWF monthly forecasts, the skill has improved significantly since operational production started in 2004, in particular for the prediction of large-scale, low-frequency events (Vitart et al 2014). Considering for example the Madden-Julian Oscillation (MJO, Madden and Julian 1971), an important source of predictability on the sub-seasonal time scale, Vitart et al (2014) show that the 2013 version of the ECMWF monthly ensemble predicted it skilfully up to about 27 days (see their Fig. 4). This is a skill gain of an impressive 1-day per year since 2004. Looking at another large-scale pattern

that influences the weather over the Euro-Atlantic sector, the North Atlantic Oscillation (NAO, Hurrell et al 2003), Vitart et al (2014) report that the ECMWF ensemble also showed clear improvements, with skill in 2013 up to about forecast day 13, compared to about day 9 ten years earlier (see their Fig. 7). Similar, although slightly smaller, improvements were reported for the prediction of sudden-stratospheric warming events (see their Fig. 10). These results provide statistically significant evidence that some phenomena can be predicted up to a few weeks ahead.

Although this paper focuses on the sub-seasonal time scale, it is also worth to briefly review evidence of seasonal forecast skill for large-scale, low-frequency phenomena. Research on predictability on seasonal time scale in the early 1990's (e.g. Palmer and Anderson 1994) led to the implementation of the first ECMWF seasonal forecast system based on a global ocean-atmosphere coupled model in 1997 (referred to as System-1). This seasonal System-1 gave a successful 6-month forecast of the 1997-98 El Niño (Stockdale et al 1998). This first coupled system was followed by System-2 in 2001, System-3 in March 2007, and the currently operational System-4 in 2011 (Molteni et al 2011). Considering this most recent ECMWF System-4, Molteni et al (2011) documented that large-scale phenomena such as monthly-average sea-surface-temperature anomalies linked to El Nino, or monthly-average ocean-basin tropical storm activities can be skilfully predicted months ahead.

These achievements in the monthly (sub-seasonal) and seasonal forecast range provide evidence of the range of predictability of some phenomena with different spatial and temporal scales. Therefore, it seems timely to examine systematically the range of predictive skill of forecasts of the atmosphere and refine the earlier estimates of the 2 week range of predictive skill.

1.4 The rationale for this work and the concept of the forecast skill horizon

What has been missing so far has been a unified, consistent approach to the estimation of the range of predictive skill for different spatial and temporal scales. To date, the skill of these different phenomena (such as the MJO, NAO, El Nino related patterns) has been assessed in different ways, using different fields, averages, and accuracy metrics. This hinders to some extent the quantitative comparison of the skill of numerical weather prediction for local, instantaneous values with the skill of spatial and temporal averages used in extended-range numerical weather prediction.

When the forecast model is perfect, the predictability is limited due to the uncertainties in the initial conditions only (e.g. due to observation errors). In this context, it is appropriate to define the predictability limit as the time when the forecast distribution is no longer different from the climatological distribution. It is possible to define general metrics based on information theory, such as relative entropy, to quantify the difference between the two distributions. DelSole and Tippet (2007) reviewed the range of approaches in this area. For instance, Kleeman (2008) estimated that the forecast distribution converges to the climatological distribution after about 45 days in the mid-latitudes using relative entropy. The estimates exclude the possibility that predictable signals originate in the ocean as the SSTs were prescribed. In addition, a fairly low-resolution T42 atmospheric model was employed. However, this configuration enabled integrations of large ensembles with about 103–104 members, which are required to compute the entropy-based estimates.

Here, we are interested in the actual predictability with an imperfect yet state-of-the art coupled atmosphere-ocean model. Due to the higher computational cost of the integrations, we are limited to a moderate ensemble size and the general information theoretical tools are deemed to be impractical to use. We select the continuous ranked probability score (CRPS, Brown 1974, Hersbach 2000) as metric to decide when the forecast distribution ceases to be more skilful than the climatological distribution. Thus, we quantify actual predictability rather than a theoretical value based on a perfect model assumption. Figure 1 illustrates the concept of a skilful PDF forecast. The climatological (reference) PDF (top panel, black dashed line) has a small overlap with the observation (top panel, solid black line), and its cumulative distribution function (CDF; bottom panel, black dashed line) is far from the observed CDF (bottom panel, solid black line). By contrast, the forecast PDFs are closer to the observation, and they get closer as the forecast time shortens (top panel). Consistently, the forecast CDFs gets closer to the observed CDF (bottom panel). The CRPS is equal to the mean squared error of the CDF (see Section 2.3 for more details). The forecast skill horizon is then defined as the forecast lead time when the forecast ceases to be more skilful than the climatological distribution. The decision is based on a statistical significance test applied to the differences of the CRPS with the 99th-percentile level as threshold.

If one considers the case of a single scalar variable and a perfect model, the predictability limit defined with the CRPS should agree with information theoretical estimates as the CRPS is a strictly proper score (i.e. the skill cannot be increased by modifying the forecast probability) for distributions with finite first moments (Gneiting and Raftery 2007). However, the actual predictability in the full state space may be larger than that obtained by studying single scalar variables. In Kleeman (2008), it was found that predictability successively increases from uni-variate, then bi-variate to tri-variate marginal entropies. However, estimation of these relative entropies in higher-dimensional spaces necessitates a coarse partitioning of the state space. It is an advantage of using the CRPS that no partitioning of the state space is required. Furthermore, there is no need to use an a priori reduction of the state space with EOFs to reduce dimensionality.

We are interested in the actual predictability with an imperfect forecast model. The predictability estimated from direct model output will underestimate the true predictability that users could exploit by correcting for known errors in the forecasts that can be diagnosed from the joint distribution of forecasts and verification data. DelSole (2005) introduces the general concept of a "regression forecast" to examine the predictability in the context of imperfect models. This correction can be seen as a calibration step, which should yield a distribution of calibrated forecasts that is more consistent with the verification data than the distribution obtained from direct model output. Depending on the volume of training data that is available to inform the calibration step, different levels of sophistication are possible. Given the moderate sample size of the training data, we decided to opt for a basic correction only. It consists of a bias correction that depends on the seasonal cycle, location and forecast lead time. Details are described in the next section. Here, an ensemble forecast will be considered skilful if the bias-corrected forecast distribution has a statistically significantly (say at the 99th-percentile level) smaller CRPS than the climatological distribution, on average over an area and for a large number of cases.

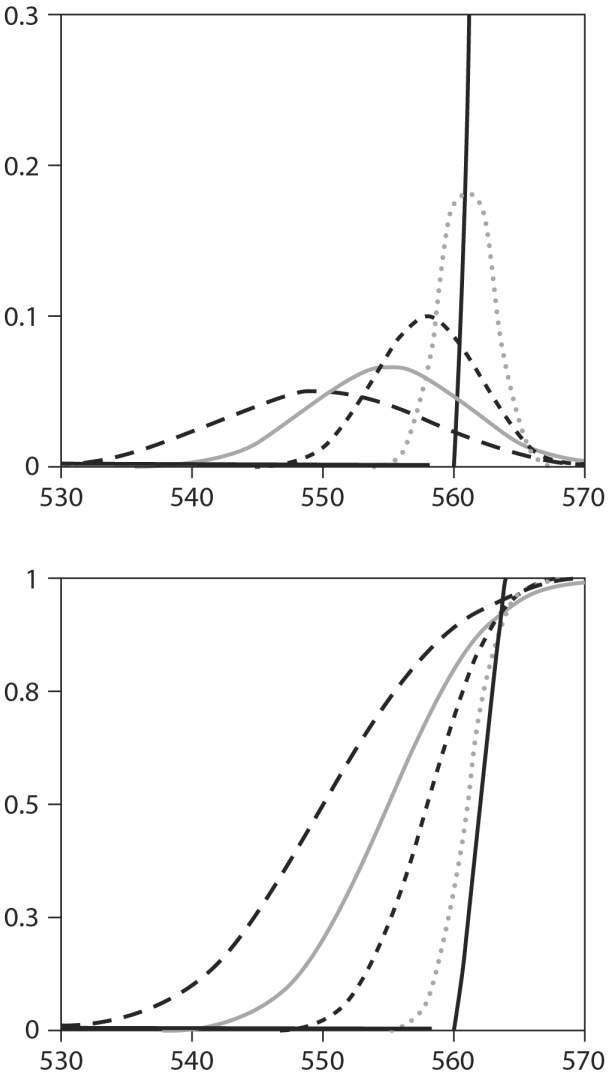

Figure 1. Schematic of a forecast of a probability density function (PDF) for the 500 hPa geopotential height at a point over the NH extra-tropics. Top panel: observed PDF [solid line, defined by a Gaussian distribution with mean 560 and standard deviation 0.5, N(560, 0.5)], forecast PDFs at three different lead times [grey dashed lines, defined by N(561,2), N(558,4) and N(555,6)] and climatological PDF [dashed black line, defined by N(550,8)]. Bottom panel: as top panel but for the corresponding cumulative distribution functions (CDFs). The CRPS is the mean squared error of the CDF. For the three forecast distributions, the CRPSs are, respectively, 0.3, 1.9 and 3.8, and the CRPS of the climatological distribution is 7.4.

One should expect that the range of predictive skill is a function of what one is trying to predict. The idea of considering weekly and even longer time averages for lead times beyond 2 weeks was mentioned by Lorenz (1982). Previous studies have attempted to quantify how probabilistic forecast skill depends on the scales. Dalcher and Kalnay (1987) have investigated the sensitivity of error growth to wave number using 10-day single ECMWF forecasts, and concluded that the longer waves are more predictable, with predictability limits beyond 10 days in winter and close to 10 days in summer, while the shorter waves showed limits below 10 days. Shukla (1981) looked at monthly means in a perfect

model context. concluded that ‘the evolution of long waves remains sufficiently predictable at least up to one month’, and suggested that improvements in model resolution and physical parameterisations could extend the predictability of time averages beyond one month. Jung and Leutbecher (2008) studied how the probabilistic skill of the ECMWF ensemble depends on spatial scales. They used waveband filters to extract the signals in planetary, synoptic and sub-synoptic scales. The predictability clearly increased with spatial scale exceeding the 15-day forecast range for the planetary scales. The sensitivity to temporal scale was determined for instance by Weigel et al (2008) for ECMWF monthly forecasts of 2-metre temperature.

The outline of the paper is as follows. In Section 2 the methodology and data used in this study are described. Section 3 presents estimates of the forecast skill horizon obtained using 1-year of ECMWF ensemble forecasts and 20 years of reforecasts. Skill horizons of different spatial and temporal scales and variables and areas will be compared. Section 4 discusses these estimates, and links our results with other published results. Finally, conclusions follow in section 5.

2 Data sets, accuracy metrics and methodology

Two ensembles are going to be used to estimate the forecast skill horizon: a bias-corrected operational ensemble and a reference climatological ensemble based on past analyses. Section 2.1 reviews the main characteristics of these ensembles and introduces the data used to generate them. Section 2.2 summarises the configuration of the ECMWF medium-range/monthly ensemble forecasts used in this work. Section 2.3 defines the probabilistic score used to measure forecast accuracy. Section 2.4 defines the two ensembles used in this work, and finally section 2.5 gives the details on how the forecast skill horizon has been computed.

2.1 The ECMWF data-sets

Four datasets have been used in this work: ECMWF operational analyses and re-analyses to initialize and verify the forecasts and to define the reference climatological ensemble, and medium-range/monthly ensemble (ENS) forecasts and reforecasts. Table 1 lists a few key characteristics of these datasets (period, resolution, membership and forecast range).

Dataset	Period	#	Frequency	Horizontal Resolution		Vertical levels (TOA)	Central ICs	Number of 32-day forecasts in the dataset
				Day 0-10	Day 10-32			
AN	July 2012 to July 2013	1	Daily analyses	TL1279 (~16 km)		91 (0.1 hPa)	N/A	N/A
ERA-I	1992 to 2011	1	Daily ERA-Interim re-analyses	T_L255 (~80 km)		60 (0.1 hPa)	N/A	N/A
ENS	July 2012 to July 2013 (107 cases)	51	Twice weekly (00UTC Mondays & Thursdays)	T_L639 (~32 km)	T_L319 (~65 km)	62 (5 hPa)	AN	5457
ENS-refc	1992-2011	5	Once a week (00 UTC of Mondays)				ERA-I	5200

Table 1. Main characteristics of the four datasets used in this study: AN and ERA-Interim (analyses used as verification to compute the model biases), ENS forecasts (ENS) and reforecasts (ENS-refc), generated from the ECMWF operational medium-range/monthly ensemble.

The first dataset, AN, includes the operational analyses: they define the unperturbed (central) initial conditions of the operational ENS. The second dataset, ERA-I, includes the ERA-Interim re-analyses (Dee et al 2011): they are used as central (unperturbed) initial-conditions of the ENS reforecasts and as verification to compute the ENS model bias for the past 20 years, and are used to verify the forecasts for the period 2 July 2012 to 8 July 2013. They are also used to construct the climatological, reference ensemble (see section 2.4). As the bias correction is estimated from verifying the reforecasts with ERA-Interim, we decided to also verify the ENS forecasts with ERA-Interim to achieve consistency between training dataset and the actual forecast verification. The third and fourth data-sets are the operational ENS forecasts and reforecasts.

2.2 The ECMWF medium-range/monthly ensemble (ENS): key characteristics

The ECMWF ENS forecasts used in this work were generated between July 2012 and July 2013. At that time, ENS included 51 members, one unperturbed and 50 perturbed ones. Forecasts were run with a variable resolution (Buizza et al 2007): TL639L62 (spectral triangular truncation T639 with a linear grid, which corresponds to about 32 km grid spacing in physical space, and 62 vertical levels) during the first 10 days, and TL319L62 (i.e. about 65 km grid spacing) thereafter. ENS forecasts were run twice a day, with initial times at 00 and 12 UTC, up to 15 days. At 00 UTC on Mondays and Thursdays the forecasts were extended to forecast day 32 (Vitart et al 2008, 2014). ENS forecasts were coupled from

initial time to the WAM wave model (Janssen et al 2005, 2013) with 55 km resolution, 24 directions and 30 frequencies up to day 10, and 12 directions and 25 frequencies afterwards. From forecast day 10, the forecasts were also coupled to the NEMO (the Nucleus for European Ocean Modelling) ocean model, with the ORCA100z42 grid (1-degree horizontal resolution and 42 vertical layers). The reader is referred to Mogensen et al (2012a, b), and references therein, for a description of the ECMWF implementation of NEMO and NEMOVAR.

For both the ENS forecast and reforecast, each ensemble member is defined by the integration of the ECMWF model equations:

Eq. (1) $$e_j(d;t) = e_j(d;0) + \int_0^t [A_0(t') + P_0(t') + P'_j(t')]dt'$$

where A0 and P0 represents the 'unperturbed' model dynamical and physical tendencies (i.e. there is only one dynamical core and one set of parameterisations called with the same parameters), P'j represents the model uncertainty simulated using two model error schemes, the SPPT (Buizza et al 1999; Palmer et al 2009) and SKEB (Berner et al 2008, Palmer et al 2009) schemes, and the lead time and start date are denoted by t and d, respectively. For the atmosphere, the initial conditions of the ensemble starting at day 'd' are defined by adding perturbations to the unperturbed analysis:

Eq. (2) $$e_j(d,0) = e_0(d,0) + e'_j(d,0)$$

The unperturbed analysis is provided by the ECMWF high-resolution 4-dimensional variational assimilations (4DVAR), run at TL1279L91 resolution and with a 6-hour assimilation window. The analysis interpolated from the TL1279L91 resolution to the TL639L62 ensemble resolution.

The perturbations were generated by a linear combination of SVs and perturbations defined by the 10 members of the ECMWF Ensemble of Data Assimilations (EDA):

Eq. (3) $$e'_j(d,0) = \sum_{a=1}^{8}\sum_{k_a=1}^{50} \alpha_{j,k_a} SV_{k_a} + [f_{m(j)}(d-6,6) - < f_{m=1,10}(d-6,6) >]$$

The reader is referred to Leutbecher and Palmer (2008) for a description of how the SVs computed for the different areas are combined, and to Buizza et al (2008) and Isaksen et al (2010) for a description of how EDA-based perturbations are combined.

For the ocean component, the initial conditions were defined by the 5-member ensemble of ocean analysis, produced by NEMOVAR, the NEMO 3-dimensional variational assimilation system (Mogensen et al 2012a, b). Each ocean analysis was generated using all available in situ temperature and salinity data, an estimate of the surface forcing from ECMWF short range atmospheric forecasts, sea surface temperature analyses and satellite altimetry measurements. The 4 perturbed members were created using perturbed versions of the unperturbed wind forcing provided by the high-resolution 4DVAR.

Model uncertainties were simulated only in the atmosphere (i.e. not in the ocean or in the land), using two stochastic schemes (Palmer et al 2010). The stochastically perturbed parameterized tendency (SPPT, Buizza et al 1999) scheme simulates random model errors due to parameterized physical processes; the

current version uses 3 spatial and time level perturbations. The stochastic back-scatter (SKEB, Shutts 2005) scheme simulates the upscale energy transfer induced by the unresolved scales on the resolved scales. Shutts et al (2011) summarize the configuration of the versions of the stochastic schemes that are used in the ensemble forecasts and reforecasts used in this study.

Since March 2008, when the ECMWF medium-range and monthly ensembles were joined, a key component to the ECMWF ENS had been the reforecast suite (Vitart et al 2008; Hagedorn et al 2012). The suite included a 5-member ensemble run once a week with the operational configuration for the past 20 years. These reforecast ensembles start from the ECMWF re-analysis (ERA-Interim) instead of the operational one, use singular vectors of the day but EDA-based perturbations computed for the current year since the EDA has been running only since 2010 [see Buizza et al (2008) and Isaksen et al (2010) for more details]. The reforecasts are used to estimate the model climate, and to produce some operational calibrated products. One example is the Extreme Forecast Index (Lalaurette 2003, Zsoter 2006), which is defined for a range of variables (e.g. 2-meter temperature and precipitation) as a measure of the difference of forecast CDF and the model cimatological CDF. The forecast CDF is computed using the 51-member ENS forecasts, while the model climatological CDF is computed using 500 reforecasts (i.e. the 5-member ENS run for the past 20 years, once a week for the 5 weeks centred on today's week: thus 5*20*5=500).

It is worth mentioning that the operational ENS configuration at the time of finalizing this work (February 2015) is different from the one used in this study. Since November 2013, when a new model cycle was introduced, ENS uses 91 (instead of 62) vertical levels with the top of the atmosphere at 0.01 hPa (instead of 5 hPa), the number of EDA members has increased to 25 (from 10), and the ocean model is coupled from the initial time (instead of day 10).

2.3 Accuracy metrics: the root-mean-square error and the continuous ranked probability score

The accuracy metrics used in this work are the root-mean-square-error (RMSE) for single forecasts (ensemble control and ensemble-mean), and the continuous ranked probability score (CRPS, Brown 1974, Hersbach 2000) for PDF forecasts. The CRPS is an extension of the Ranked Probability Score (Epstein 1969, Murphy 1971). Compared to the RPS, the CRPS has two advantages: it is sensitive to the entire range of the parameter of interest and it does not rely on predefined intervals (i.e. it is not restricted to fixed intervals as the RPS is). The CRPS is the limit of the RPS for an infinite number of intervals, and is one of the most commonly used metrics of probabilistic forecast accuracy.

The CRPS has been computed as in Hersbach (2000). Consider a single grid point and a variable x (e.g. the 850 hPa temperature). Denote by p(x) the forecast PDF, defined by the ENS forecast members, and by xobs the observed value. The CRPS is defined as:

Eq. (4) $$CRPS = \int_{-\infty}^{+\infty} \left[CDF(x) - CDF_{obs}(x) \right]^2 dx$$

where CDF and CDFobs are the cumulative distribution functions of the forecast and the observation:

Eq. (5.a) $$CDF(x) = \int_{-\infty}^{x} p(y)dy$$

Eq. (5.b) $$CDF_{obs}(x) = H(x - x_{obs})$$

with:

Eq. (5.c) $$H(x) = \begin{cases} 0 & for \quad x < 0 \\ 1 & for \quad x \geq 0 \end{cases}$$

being the Heaviside function.

The CRPS is zero for a perfect forecast, when the forecast and observed CDFs are identical.

In this work we consider annual- and area-average CRPS values, computed by averaging the CRPS considering all cases covering one year of forecasts (jc=1, NC where NC, the number of cases, is 107, to include all Mondays and Thursdays from 2 July 2012 to 8 July 2013) , and considering all grid points inside a specific area:

Eq. (6) $$< CRPS >_{ND,G} = \frac{1}{G \times ND} \sum_{jc=1}^{NC} \sum_{g=1}^{G} w_g CRPS_{jd,g}$$

where the grid-point weights are proportional to the cosine of latitude. In principle, other metrics could have been considered as well. However, we think this a secondary aspect that can be left to future work.

2.4 Definition of the bias-corrected ensemble and the climatological ensemble

The operational ENS forecasts and reforecasts have been used to generate the bias-corrected ensemble (ENS-BC), and ERA-Interim re-analyses have been used to generate the reference, climatological ensemble (ENS-CLI), as explained hereafter.

2.4.1 The bias-corrected ensemble (ENS-BC)

For each forecast date between 2 July 2012 and 8 July 2013 for which the 32-day ensemble was produced (00 UTC of Mondays and Thursdays, i.e. 107 cases), the ENS-BC ensemble has been defined by the 51 ENS members after removing the bias computed from the ENS-refc dataset. More precisely, (following the same approach used to generate the operational ENS-based Extreme Forecast Index product) the bias has been calculated considering the 500 ENS reforecast members [5-member (j=1,5) ENS forecasts with initial dates defined by the 5 weeks centred on the forecast date (d=1,5) and for the past 20 years (y=1992-2011 for forecasts starting in 2012, and y=1993-2012 for forecasts starting in 2013)]:

Eq. (7.a) $$< bias >= \frac{1}{500} \sum_{y} \sum_{j=1}^{5} \sum_{d=1}^{5} \left[e_j(y;d,t) - a_j(y;d+t) \right]$$

For each date, each member of the ENS-BC forecast is defined as:

Eq. (7.b) $$e_j^{BC}(d,t) = e_j(d,t) - < bias >$$

Consider, e.g., 00UTC of 15 November 2012:

- ENS includes 51 forecasts, up to 32 days, started at 00 UTC of 15 November, 2012;
- The ENS reforecasts include 5-member forecasts starting at 00UTC of 1, 8, 15, 22 and 29 November for the past 20 years (1992-2011);
- The bias is computed using these 500 forecasts, verified against ERA-Interim analyses;
- ENS-BC is defined by the 51 bias-corrected ENS forecasts.

2.4.2 The climatological ensemble (ENS-CLI)

The climatological ensemble includes 100 members, each defined by a 32-day chronological sequence of ERA-Interim re-analyses. More precisely, it consists of the re-analyses that verify the 500-member ENS reforecasts, which are used to compute the model bias. Consider, for example as in section 2.4.1, 00UTC of 15 November 2012: this is how the 100-member ENS-CLI is defined. For the 20 reforecast years 1992-2011, and for the 5 dates for which the reforecasts have been run (1, 8, 15, 22 and 29 November), we can construct a chronological sequence of analyses:

- ENS-CLI(m=1)={AN(1991.11.01.00),AN(1991.11.01.12),…,AN(1991.12.01.12)};
- ENS-CLI(m=2)={AN(1991.11.08.00),AN(1991.11.08.12),…,AN(1991.12.09.12)};
- … ;
- ENS-CLI(m=100)={AN(2011.11.29.00),AN(2011.11.29.12),…,AN(2011.12.31.12)}.

This defines a 100-member ensemble used as a reference forecast to be compared with the operational ENS. The same procedure is used for each of the 107 initial conditions (00UTC of Mondays and Thursdays from 2 July 2012 to 8 July 2013). The ENS-CLI control is defined as a randomly selected member of ENS-CLI.

2.5 Definition of the Forecast Skill Horizon

As mentioned in the Introduction and illustrated schematically in Fig. 1, the forecast skill horizon is defined as the forecast time when the forecast PDF ceases to be distinguishable, in a statistical sense, from the climatological (reference) PDF. The forecast skill horizon is computed as the forecast time when the average CRPS of the bias-corrected ensemble ENS-BC ceases to be statistically significantly lower, at the 99th-percentile level, than the CRPS of the climatological ensemble ENS-CLI. More precisely, this is defined as the forecast lead time when at least 1% of the sample estimate of the distribution of the mean difference of the CRPS of the forecast distribution and the CRPS of the climatological distribution is positive. Mean score differences are assumed to be distributed by Student's t-distribution. Serial correlation of score differences between start dates is accounted for. The CRPS of a perfectly reliable ensemble should converge to the CRPS of the climatological ensemble. However,

for the actual ECMWF ensemble it is possible that the CRPS reaches the value of the climatological ensemble at some lead time and after that exceeds it. For instance, this could happen if the ensemble is over, or under, dispersive. This will not limit our ability of determining the forecast skill horizon for the ECMWF ensemble.

The forecast skill horizon has been estimated for seven variables and three different areas:

- Variables: geopotential height at 500 hPa (Z500); temperature at 850 and 200 hPa (T850, T200); wind components at 850 and 200 hPa (U850, V850, U200 and V200);
- Areas: northern hemisphere extra-tropics (20°N-90°N; NH); southern hemisphere extra-tropics (20°S-90°S; SH); tropics (20°S-20°N; TR).

As motivated in the Introduction, the sensitivity of the forecast skill horizon to the temporal and spatial scale of weather events has been assessed by considering fields with increasingly longer spatial scale and lower frequency variability. This was obtained by applying spatial smoothing and time-averaging operators:

- Spatial smoothness was obtained by spectrally truncating the fields from T120 (spectral triangular truncation with total wave number 120, corresponding to about 170 km) to T60 (~ 330km), T30 (~670 km) and T15 (~1300km), T7 (3000km);
- Temporal low-frequency variability was achieved by applying time averaging, thus considering not only instantaneous grid-point values, but also values averaged over 1, 2, 4, 8 and 16 days. The forecasts for time averages have been assigned a lead time equal to the mid-point of the averaging window.

Figures 2 and 3 illustrate, for one case of December 2012, the impact of spatial truncation on instantaneous (i.e. H0) fields and the impact of time-averaging on local (i.e. T120) fields. Figure 2 shows that as more severe spectral filtering is applied, the small scales start disappearing and the local maxima and minima decrease in amplitude, with the impact being more detectable once truncations beyond T15 are applied. Figure 3 shows the impact of time averaging on local (i.e. T120) fields. As for the case of spatial truncation, as the time averaging period lengthens the field becomes smoother, with local maxima and minima decreasing in amplitude, and the gradients becoming less pronounced. Note that, for this case, the effect of spatial or time averaging on the fields is similar apart for the strongest truncation.

Once the analysis and forecast fields have been retrieved from the ECMWF archive, this is how the forecast skill horizon has been computed for a Txx spatial truncation and 2H-hour time-average for each variable and area considered:

a) extract all the j-th members spectral fields at the Txx truncation, and define them on a regular 1.5°-degree latitude-longitude grid;

b) apply a 2H-hour time averaging over the interval (t-H;t+H), e.g. for the j-th member; thus, for example, for the H12 averaging, the +144h field is defined, at each grid point, by the time average between +132h and the +156h;

c) compute the bias from the ENS reforecasts, remove the bias from each member, and construct the ENS-BC forecasts;

d) construct the reference climatological ensemble ENS-CLI using truncated and time-average analyses;

e) evaluate the performance of the two ensembles (ENS-BC and ENS-CLI) using RMSE and CRPS;

f) define the forecast skill horizon as the forecast time when CRPS(ENS-BC), defined as in Eq. (6), is not significantly different from the CRPS(ENS-CLI), at the 99th-percent level.

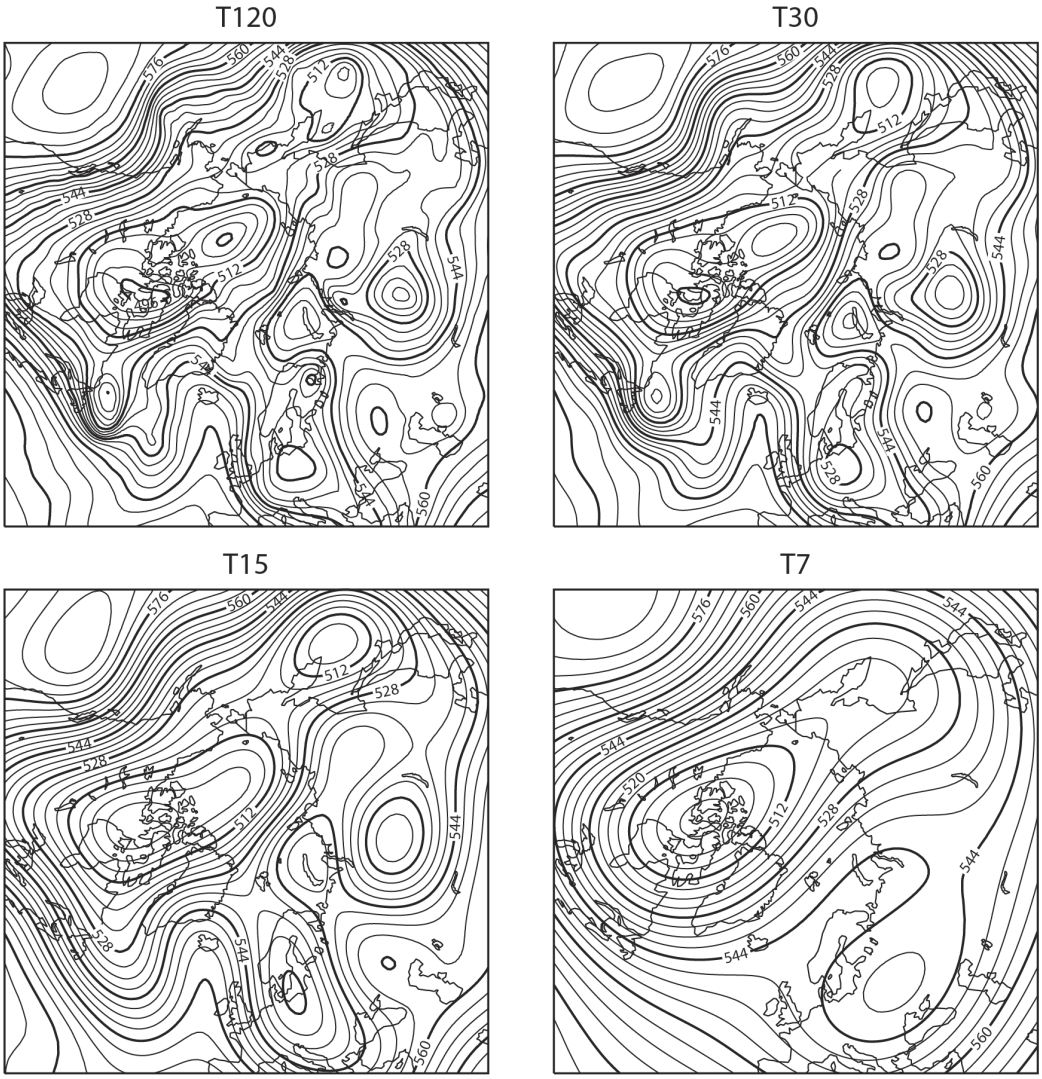

Figure 2. Impact of spectral truncation on an instantaneous (H0) 500 hPa geopotential height field. Top-left panel: original T120 field valid for 12UTC of the 10th of December 2012; top-right panel: T30 truncation; bottom-left panel: T15 truncation; bottom-right panel: T7 truncation. The contour interval is 80m.

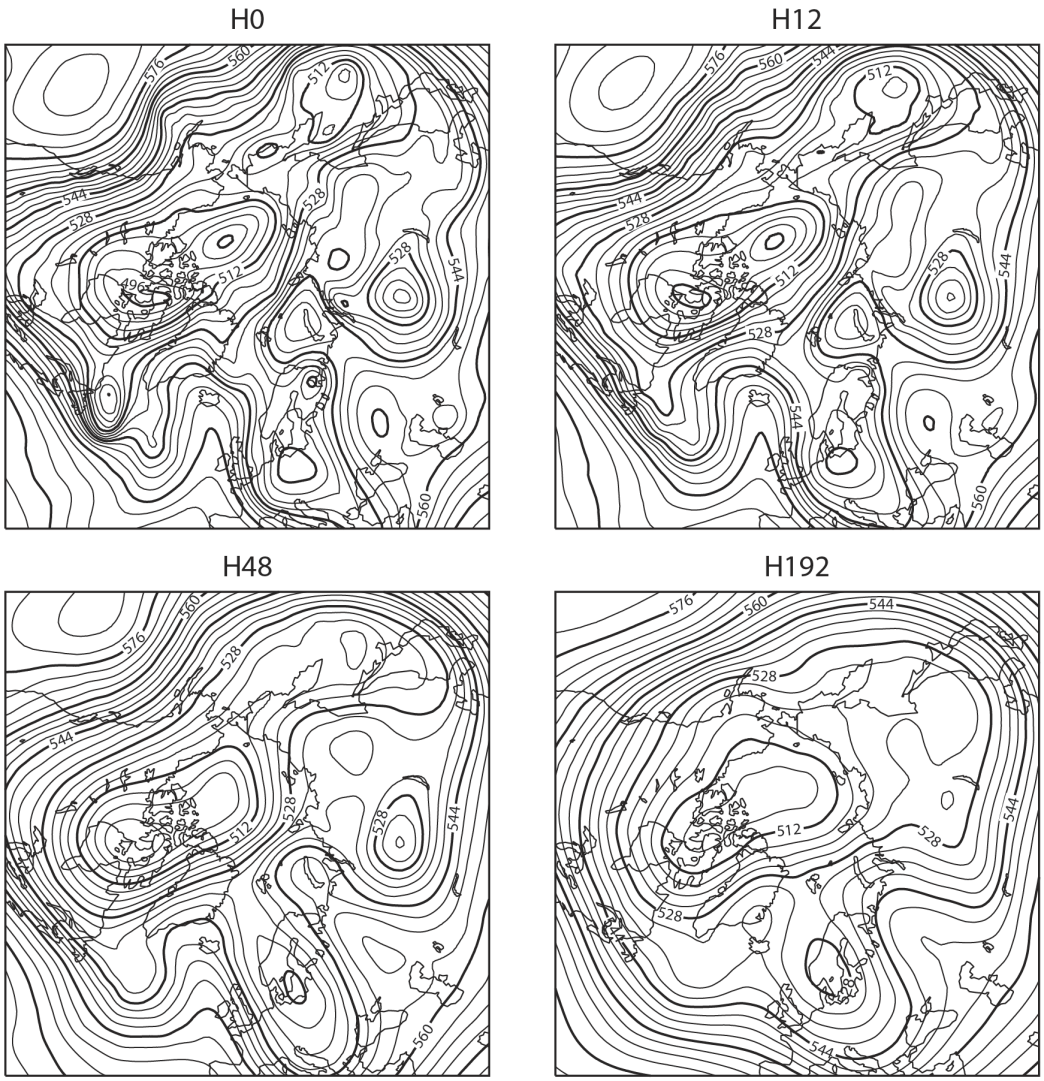

Figure 3. Impact of time-averaging on local (T120) 500 hPa geopotential height field. Top-left panel: instantaneous field valid for 12UTC of 10th of December 2012; top-right panel: 1-day (H12) time-average field centred on 12UTC of the 12th (defined by averaging the fields from 12UTC of the 9th to the 12 UTC of the 11th of December); bottom-left panel: 4-day (H48) time average field centred on 12UTC of the 12th; bottom-right panel: 16-day (H192) time average field centred on 12UTC of the 12th. The contour interval is 80m.

It should be noted that, since we have been using 32-day forecasts, for some fields we might only be able to indicate that the forecast skill horizon is beyond the available forecast length. It should also be taken into account that the available forecast length depend on the time-averaging: it is 32 days for instantaneous (H0) fields, 31.5 days for 24-hour average (H12) fields, …, 28 days for 8-day average (H96) fields and 24 days for 16-day (H192) average fields. When the forecast skill horizon is estimated to be beyond the available forecast length, it will be preceded by the symbol '>' (e.g., for H96, it will be indicated as '> 26.0').

3 Estimates of the forecast skill horizon for the ECMWF ensemble

In the first part of this section we are going to discuss results based on instantaneous, grid-point fields (i.e. without any spectral filtering and/or time averaging), both for the single control forecast and the full ensemble forecast. Secondly, we will discuss the sensitivity of the forecast skill horizon to the spatial and temporal scale of the forecast fields. Thirdly, we will discuss the impact of the bias-correction. Finally, we will discuss the seasonal variation of the forecast skill horizon.

3.1 Forecast skill horizon for instantaneous, grid-point forecasts

Initially, consider instantaneous, grid-point fields given by the control forecast, so that we can compare our results with earlier estimates mentioned in Section 1.2 such as the one of Lorenz (1969a). Figure 4 shows the annual average (for the 107 cases considered in this work; see section 2.1) RMSE of the ENS-BC control forecast and the ENS-CLI reference, for the 500 hPa geopotential height (Z500) over the Northern and the Southern Hemispheres (NH, SH). It also shows the difference of the RMSEs of ensemble forecast and the climatological ensemble with the confidence intervals (1st to 99th percentile of the estimated distribution of score differences). Results indicate that the confidence bars touch the zero line at forecast day 17 for Z500 over NH, and at day 21.5 for Z500 over the SH. Fortuitously, the NH forecast skill horizon agrees with Lorenz (1969a)'s estimate of 16.8 days.

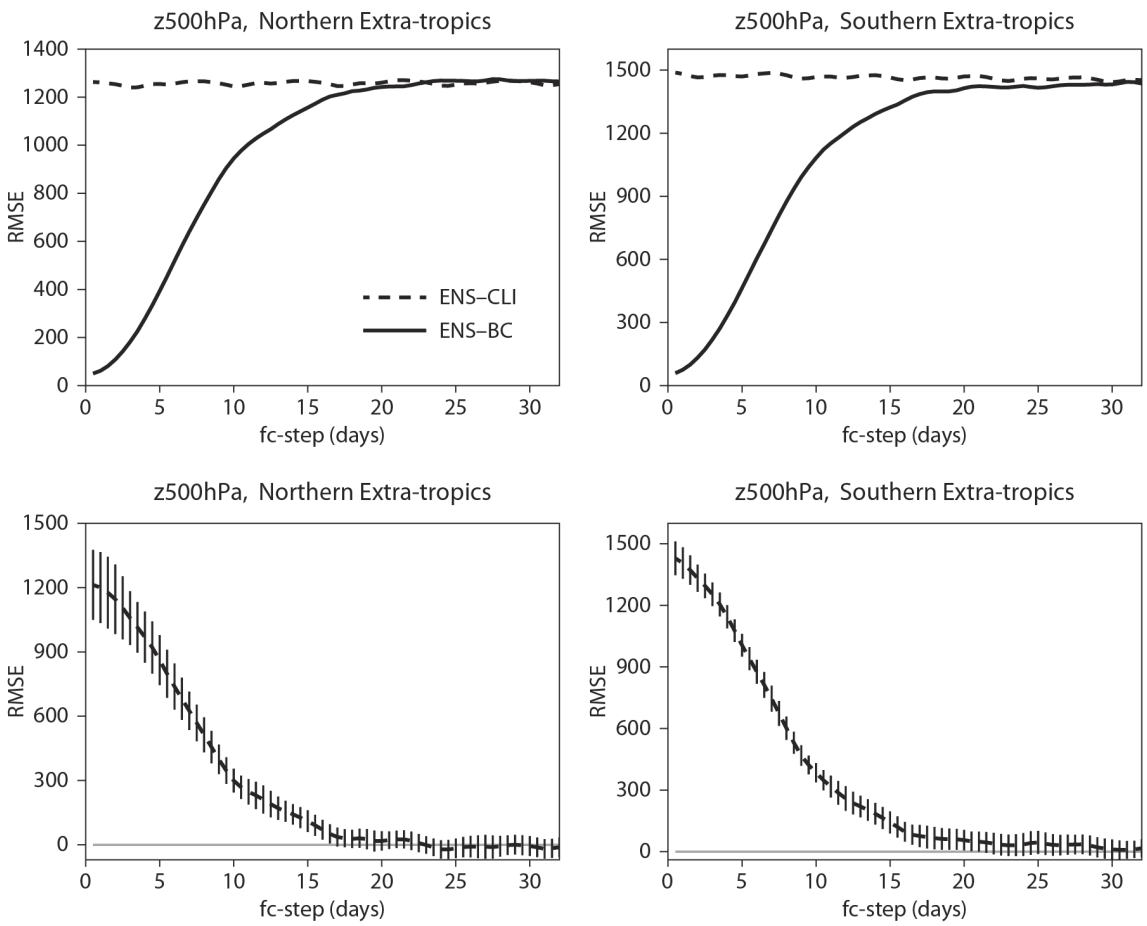

Figure 4. Top-left panel: NH annual-average (107 cases) root-mean-square-error (RMSE) of the ENS-BC control forecast (solid line) and the ENS-CLI control forecast (dashed line), for instantaneous 500 hPa geopotential fields truncated at T120 (H0-T120). Bottom-left: difference between the RMSE of the ENS-CLI and the ENS-BC control forecasts (dashed line) with 98th percentile confidence intervals (bars) for NH. Right panels: as left panels but for SH. The forecast skill horizon (see text for definition) is 17 days for Z500 over NH and 21.5 days for Z500 over SH. Units in all panels are m^2/s^2.

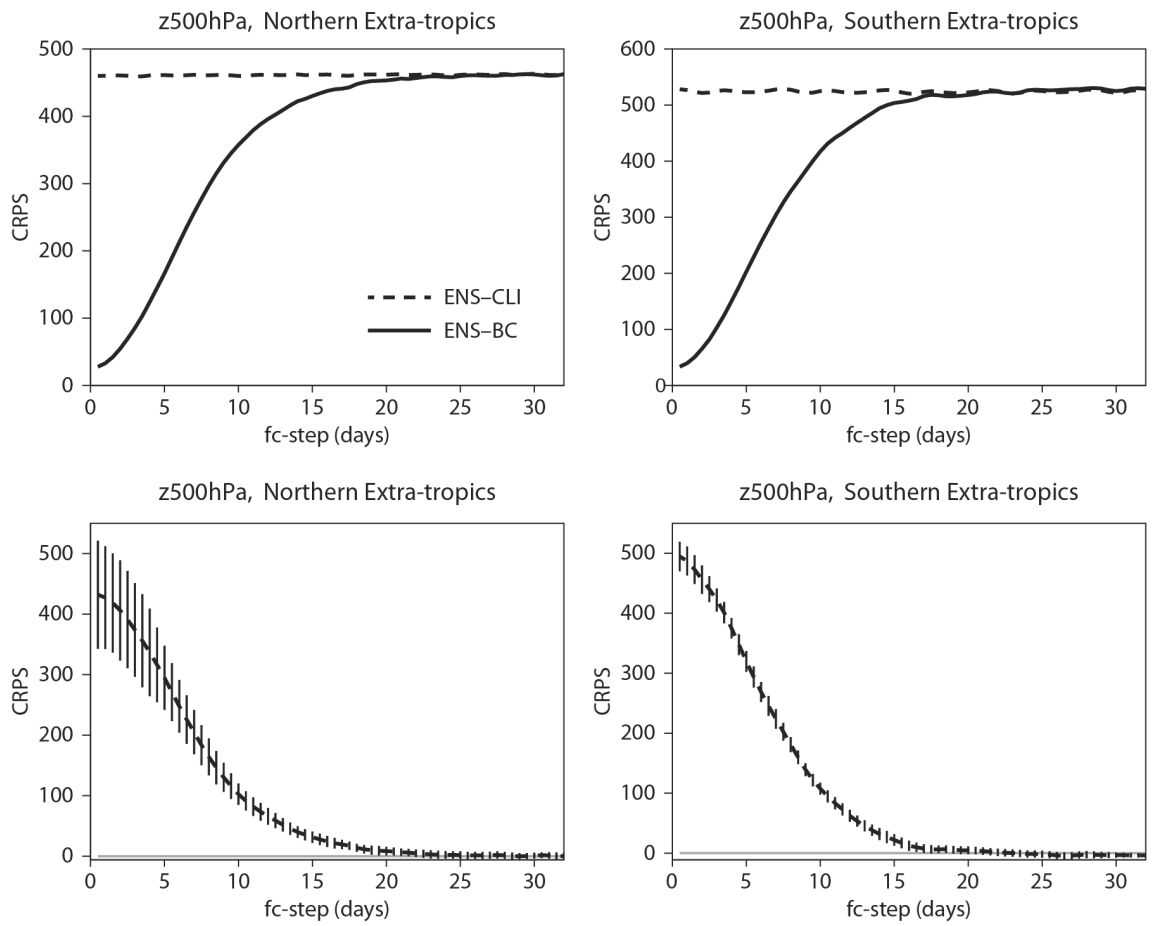

Figure 5. Top-left panel: NH annual-average (107 cases) CRPS of ENS-BC (solid line) and ENS-CLI (dashed line), for instantaneous 500 hPa geopotential height fields truncated at T120 (H0-T120). Bottom-left: difference between the CRPS of ENS-CLI and ENS-BC (dashed line) with 98th percentile confidence intervals (bars) for NH. Right panels: as left panels but for SH. The forecast skill horizon (see text for definition) is 22 days for Z500 over NH and 19 days for Z500 over SH. Units in all panels are m^2/s^2.

Figure 5 shows the corresponding annual average CRPS and the difference of the climatological and forecast CRPS scores for instantaneous, grid-point ensemble forecasts of Z500 over NH and SH. The overall shape of the curves is similar to the RMSE curves, but the confidence intervals have shorter lengths for the CRPS, thus suggesting that there is less variability in the CRPS than the RMSE difference between ENS-BC and ENS-CLI. The confidence intervals hit the zero line at forecast day 22 for the probabilistic prediction of Z500 over NH, and at day 19 for the probabilistic prediction of Z500 over SH. Figure 6 shows the annual average RMSE (left panels) and the CRPS (right panels) for instantaneous, grid-point 850 hPa temperature forecasts, computed over the NH, the SH and the tropics. It shows that the confidence intervals hit the zero line at between forecast day 16.5 (for T850 CRPS over SH) and day 23 (for T850 CRPS over NH).

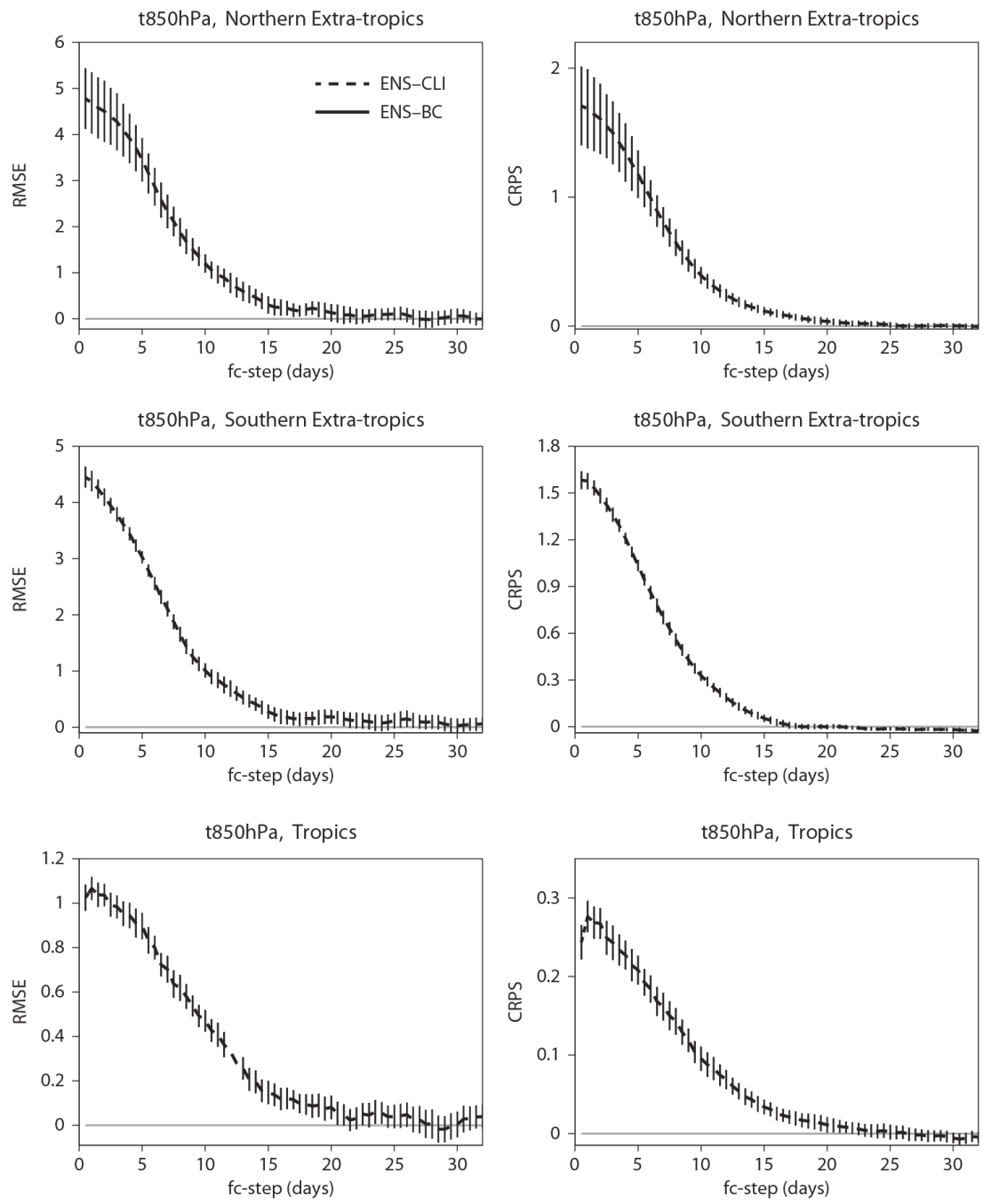

Figure 6. Left panels: differences between the RMSE of the ENS-BC and the ENS-CLI control forecasts, computed over NH (top-left panel), SH (middle-left panel) and the tropics (bottom-left panel), for instantaneous 850 hPa temperature fields truncated at T120 (H0-T120). Right panels: as left panels but for the CRPS. In all panels, solid lines show the annual (107 cases) averages, and the bars the 98th percentile confidence interval of the difference. The forecast skill horizon(see text for definition) for the single control forecasts are 19.5 days for T850 over NH, 21.5 days for T850 over SH and the tropics; the forecast skill horizon for the probabilistic predictions is 23 days for T850 over NH, 16.5 days for T850 over SH and 22 days for T850 over the tropics. Units in all panels are K.

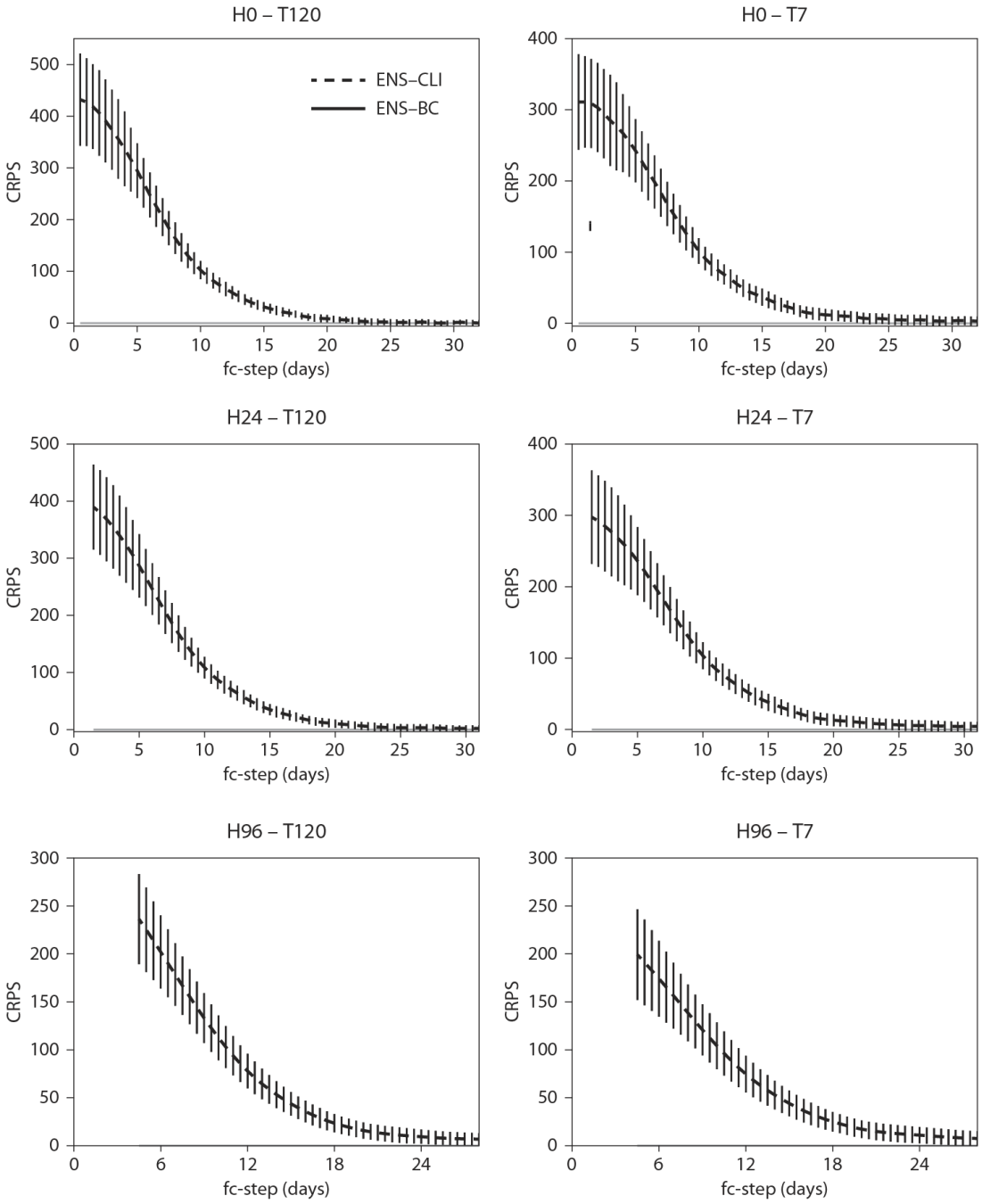

Figure 7. As Fig. 6 but for Z500 over NH for spatially/temporally filtered fields. Top-left panel: instantaneous fields truncated at T120 (H0-T120: 21d forecast skill horizon). Middle-left panel: 48-hour averaged fields truncated at T120 (H24-T120: 22d forecast skill horizon). Bottom-left panel: 192-hour averaged fields truncated at T120 (H96-T120: 26d forecast skill horizon). Top-right panel: instantaneous fields truncated at T7 (H0-T7: 23d forecast skill horizon). Middle-right panel: 48-hour averaged fields truncated at T7 (H24-T7: 24d forecast skill horizon). Bottom right panel: 192-hour averaged fields truncated at T7 (H96-T30: >28d forecast skill horizon). Units in all panels are m^2/s^2.

Thus the first conclusion that we can draw from these results is that Lorenz (1969a)'s estimates for instantaneous, grid-point forecasts of Z500 and T850 fields, were not too different from our latest estimates, which indicate slightly longer values of about 17-22 days. The second one is that the precise value depends on the variable and the area. The third one is that different forecast skill horizons (of up to 3 days) are obtained if one considers probabilistic versus single forecasts. For the reasons discussed in the introduction, hereafter we will base our analysis and considerations on ensemble-based estimates.

3.2 Sensitivity of the forecast skill horizon to spatial and temporal scales

Figure 7 shows the sensitivity of CRPS differences between ENS-BC and ENS-CLI for Z500 over NH to spatial filtering and time averaging. For reason of space, results are shown only for six of the 36 configurations that have been considered:

- H0 T120: instantaneous values, extracted at T120 spectral truncation (equivalent to about 180 km resolution, which is about 6 times the ENS grid spacing, which is 32 km);
- H0 T7: instantaneous values, truncated at T7 spectral resolution (about 660 km);
- H24, T120 and T7: as above but for 48-hour time averaged values (for each time step t, fields have been averaged from t-12 hours to t+12 hours);
- H96, T120 and T7: as above but for 192-hour time averaged values (for each time step t, fields have been averaged from t-96 hours to t+96 hours).

Z500	**H0**		**H24**		**H96**	
	NH	**SH**	**NH**	**SH**	**NH**	**SH**
T120	**21.0**	**19.0**	22.0	20.5	26.0	21.5
T30	22.0	19.0	22.0	20.5	26.0	21.5
T7	23.0	21.0	24.0	21.0	> **28.0**	**22.5**

Table 2. Forecast skill horizons for the probabilistic prediction of the 500 hPa geopotential height (Z500) over NH and SH, for fields with increasingly smother spatial scales (T120, T30 and T7 spectral triangular truncation) and longer time average [instantaneous (H0), 2d (H24) and 8d (H96)].The arrow symbol (>) indicates that the forecast skill horizon was beyond the last time step that could have been verified (i.e. 32d for H0, 31 for H24 and 28d for H96).

Table 2 lists the forecast skill horizons for the 6 configurations, computed for Z500 over NH and SH. Results indicate that as the spectral filtering makes the fields smoother and the fields are time-averaged, the forecast skill horizon becomes longer. The sensitivity to time averaging is stronger than the

sensitivity to spatial filtering. Results depend on the area: for Z500 over NH, the forecast skill horizon for instantaneous (H0) high-resolution (T120) fields is 21 days, while it is more than 28 days for 8d-averaged (H96) lower-resolution (T7) fields. For Z500 over SH, the impact of smoothing and averaging is less evident, with forecast skill horizon differences of only 3.5 days (from 19 to 22.5 days). If one compares the forecast skill horizon for the instantaneous, grid point values (H0-T120) with the spatially filtered and 8-day averaged (H96-T7) ones, results indicate that the low-frequency, larger-scale phenomena are between 3.5 and 7 days more predictable.

T850	**H0**			**H24**			**H96**		
	NH	**SH**	**TR**	**NH**	**SH**	**TR**	**NH**	**SH**	**TR**
T120	**23.0**	**16.5**	**22.0**	25.0	18.0	26.0	> 28.0	25.0	> 28.0
T30	24.0	17.0	23.0	25.0	18.0	27.0	> 28.0	25.5	> 28.0
T7	> 32.0	23.0	26.5	> 31.0	23.5	28.0	**> 28.0**	**> 28.0**	**> 28.0**

Table 3a. Forecast skill horizons for the probabilistic prediction of the 850 hPa temperature (T850) over NH, SH and the tropics (TR), for fields with increasingly smother spatial scales (T120, T30 and T7 spectral triangular truncation) and longer time average [instantaneous (H0), 2d (H24) and 8d (H96)].The arrow symbol (>) indicates that the forecast skill horizon is larger than the last time step that could have been verified (i.e. 32d for H0, 31 for H24 and 28d for H96).

U850	**H0**			**H24**			**H96**		
	NH	**SH**	**TR**	**NH**	**SH**	**TR**	**NH**	**SH**	**TR**
T120	**17.0**	**15.5**	**21.5**	19.0	18.0	23.0	23.0	20.0	26.0
T30	18.0	15.5	22.5	20.0	18.5	24.0	24.0	20.0	26.0
T7	27.5	19.5	25.0	28.0	20.5	25.5	**> 28.0**	**21.5**	**28.0**

Table 3b. As Table 3.a but for the probabilistic prediction of the 850 hPa zonal wind (U850).

Tables 3 and 4, which show forecast skill horizons for the temperature and wind at 850 hPa and at 200 hPa, computed over three areas (NH, SH and the tropics), confirm the Z500 results. For NH and SH, the forecast skill horizons for T850 are 1-2 days longer than the Z500 forecast skill horizons. Overall, for

temperature and wind at these two pressure levels the tropics have forecast skill horizons similar to the extra-tropics. As for Z500, the low-frequency, larger-scale phenomena (H96-T7) have forecast skill horizons about one week longer than the instantaneous, grid point values (H0-T120).

T200	**H0**			**H24**			**H96**		
	NH	**SH**	**TR**	**NH**	**SH**	**TR**	**NH**	**SH**	**TR**
T120	**28.0**	**19.0**	**19.5**	30.5	20.0	20.0	> 28.0	22.5	20.5
T30	28.0	19.0	19.5	30.5	20.0	20.0	> 28.0	22.5	20.5
T7	31.5	24.5	19.5	> 31.0	25.0	20.5	**> 28.0**	**26.0**	**20.5**

Table 4a. As Table 3a but for the probabilistic prediction of the 200 hPa temperature (T200).

U200	**H0**			**H24**			**H96**		
	NH	**SH**	**TR**	**NH**	**SH**	**TR**	**NH**	**SH**	**TR**
T120	**23.5**	**21.0**	**25.5**	25.5	23.0	26.5	> 28.0	27.0	28.0
T30	24.0	21.5	26.0	26.0	23.0	26.5	> 28.0	27.0	28.0
T7	32.0	23.0	26.5	> 31.0	23.5	27.0	**> 28.0**	**25.0**	**27.5**

Table 4b. As Table 3a but for the probabilistic prediction of the 200 hPa zonal wind (U200).

The results discussed above for the bias-corrected ensemble (ENS-BC) can be summarized by comparing 'grand-averages' of the forecast skill horizons, computed by averaging the values for different variables, areas and scales.

To highlight the sensitivity to the time averaging, the top panel of Fig. 8 shows, for six temporal scales (H0, H12, H24, H48, H96 and H192), the range of the 36 annual average forecast skill horizons computed for three spatial scales (T120, T30 and T7), four fields (Z500, T850, U850, V850) and three areas (NH, SH and TR). For example, the '0 day' bar shows the average plus and minus the standard deviation, of the 36 forecast skill horizons computed for the instantaneous, local fields (H0). A simple linear regression of the average forecast skill horizons with the averaging period (99% correlation coefficient) confirms that there is a strong sensitivity to the time averaging. Average forecast skill

horizons range between 16-24 days for instantaneous (H0) fields, while they range between 23 days and the maximum available forecast length (32 days) for 16-day average fields (H192). The slope of the linear fit indicates a predictability gain of about 0.5 days per day of averaging.

To highlight the sensitivity to the spatial filtering, the bottom panels shows, for three spatial scales (T120, T30 and T7), the range of the 72 annual average forecast skill horizons computed for the six temporal scales (H0, H12, H24, H48, H96 and H192), four fields (Z500, T850, U850, V850) and three areas (NH, SH and TR). Also in this case the average results confirm a linear relationship (83% correlation coefficient), with spatially-filtered fields being more predictable than local fields. In this case, a linear regression of the average forecast skill horizons with the spatial filtering indicates a weaker sensitivity to spatial filtering, with a gain of about 0.1 day for each extra 10-wave spectral filtering.

Figure 8 highlights in few key numbers, our estimates of the length of the forecast skill horizon computed using one year of ECMWF bias-corrected, monthly ensemble forecasts, to be contrasted with the early estimates of about 2 weeks. Values depend on the scales, the variable and the area considered. As already mentioned, please note that for some filter settings the skill horizon exceeds the forecast length up to which we had data, and thus we can only say that the skill horizon is longer than the maximum available forecast step (this is indicated by the symbol '>' in front of the forecast length). The forecast skill horizon ranges between 16 and 24 days for instantaneous fields, with the longest lengths obtained if one spatially filters the field. This range includes, in the lower end, Lorenz (1969a, b)'s estimates of about 2 weeks obtained for the 500 hPa geopotential field over the NH. It also clearly shows that the estimates are much longer, between 23 and the maximum available forecast length (i.e. 32 days for H0, 31.5 days for H12, ..) for time-average fields.

It is worth to point out that the increase in predictability due to time-averaging is not entirely a result of assigning the forecast to the middle of the time period, so that the earlier and more accurate forecasts are included. We compared the skill of the time-average forecast with the skill based on time-averages of the score of the instantaneous forecast, assigning the average skill measure to the middle of the time window. Results indicate that the skill of the time-average fields is statistically significantly higher than the skill of time averaged scores.

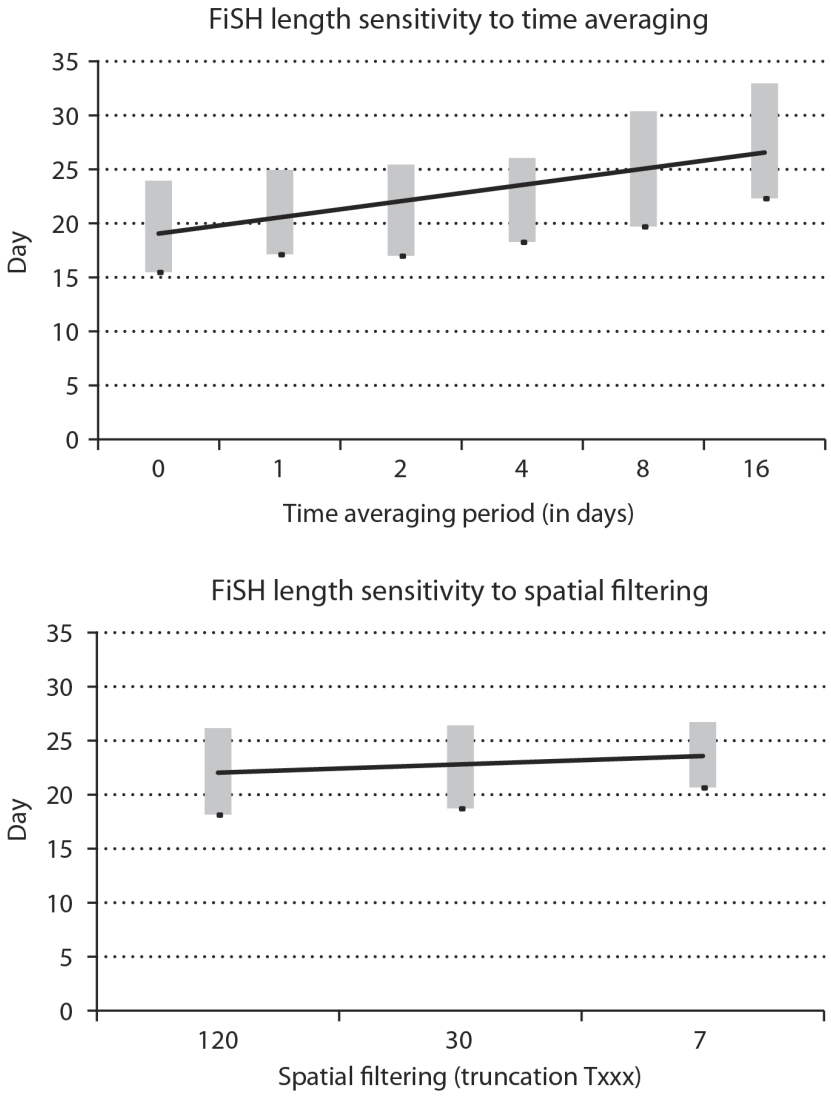

Figure 8. Forecast skill horizon sensitivity to time-averaging (top panel) and to spectral filtering (bottom panel). Each panel shows the 'grand-average' (see section 3.4 for details) forecast skill horizon computed considering four variables (Z500, T850, U950, V850) and three regions (NH, SH, TR).

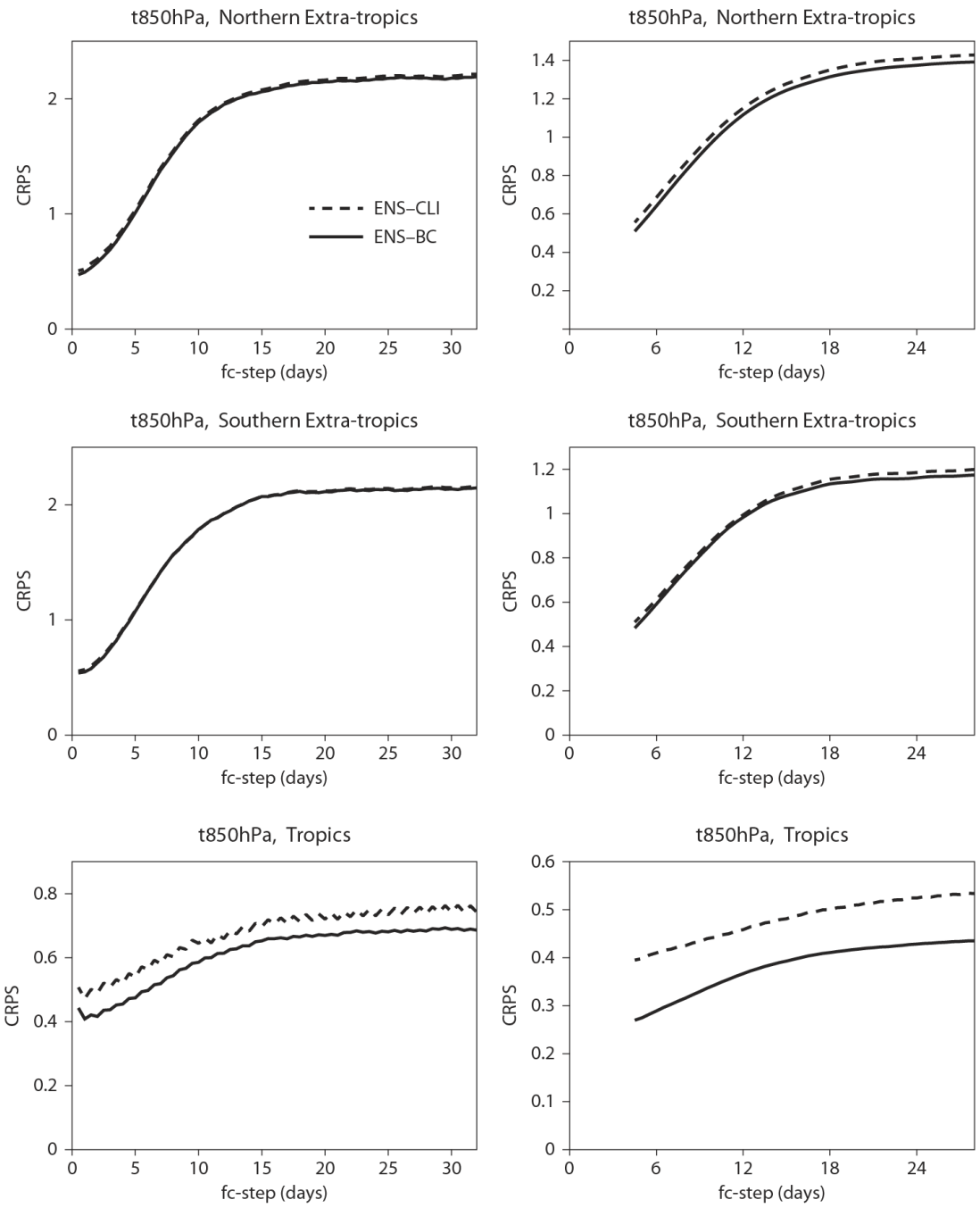

Figure 9. Sensitivity to bias correction of annual-average (107 cases) CRPS for the 850 hPa temperature over NH (top panels), SH (middle panels) and tropics (bottom panels) for instantaneous fields truncated at T120 (H0-T120, left panels) and for 192-hour averaged fields truncated at T120 (H96-T120: right panels). Solid lines refer to bias-corrected ENS-BC and dashed lines to ENS without bias-correction. Units in all panels are K.

3.3 Sensitivity of the forecast skill horizon to bias correction

Since all the ensemble forecasts have been bias-corrected, a relevant question to ask is what the impact of bias-correction is. To assess it, two configurations have been scored also without bias-correction: instantaneous (H0) and 8-day average (H96) fields, with a T120 truncation.

Figure 9 shows the CRPS for ENS (i.e. without bias-correction; dashed lines) and ENS-BC (solid lines), for the 850 hPa temperature over NH, SH and TR for instantaneous (H0) and 8-day average fields (H96). Table 5 lists the forecast skill horizons for two fields over three areas. Results indicate that the bias-correction has a clear impact, especially over the tropics for temperature. Over the extra-tropics, bias-correction increases the forecast skill horizon for instantaneous fields by about 3 days, while over the tropics it extends it by about 3 weeks.

	H0			**H96**		
	NH	**SH**	**TR**	**NH**	**SH**	**TR**
Z500 - ENS-BC	22.0	19.0	n/a	> 28.0	21.5	n/a
Z500 - ENS	19.0	16.5	n/a	20.5	18.5	n/a
T850 - ENS-BC	23.0	16.5	22.0	> 28.0	25.0	> 28.0
T850 - ENS	19.5	16.5	12.5	21.5	18.5	6.5

Table 5. Sensitivity to bias-correction. Forecast skill horizon for the probabilistic prediction of the 500 hPa geopotential height (Z500) and the 850 hPa temperature (T850) over NH, SH and the tropics (TR), for instantaneous (H0) and 8-day average (H96) fields, truncated at T120. The arrow symbol (>) indicates that the forecast skill horizon was beyond the last time step that could have been verified (i.e. 32d for H0 and 28d for H96).

3.4 Seasonal variations of the forecast skill horizon

Figure 10 shows that the forecast skill horizon depends on the season. The figures shows the average CRPS of ENS-BC and of ENS-CLI, computed selecting 28 cases in JJA, DJF and the whole year, for the 500 hPa geopotential height truncated at T120. The CRPS of both ENS-BC and ENS-CLI vary, with the cold season showing the largest asymptotic values CRPS(ENS-CLI) and also the longest forecast skill horizon. The variation can be detected both if one considers instantaneous fields or time-averaged ones. Considering the NH, the forecast skill limit for instantaneous fields (top-left panel) are about 2 days longer for the cold season (DJF) (Table 6), with annual average values (considering only 28 cases) being somewhere in between the DJF and the JJA values. By contrast, over the SH values are about 5 days longer for the warm season (DJF) than the cold one (JJA). Thus these results confirm an inter-annual variation of the forecast skill horizon, and indicate that in the NH (SH) the cold (warm) season,

at least for the period considered in this work, is more predictable. However, care should be taken in generalizing these results, since the seasonal averages include only 28 cases spanning one specific year.

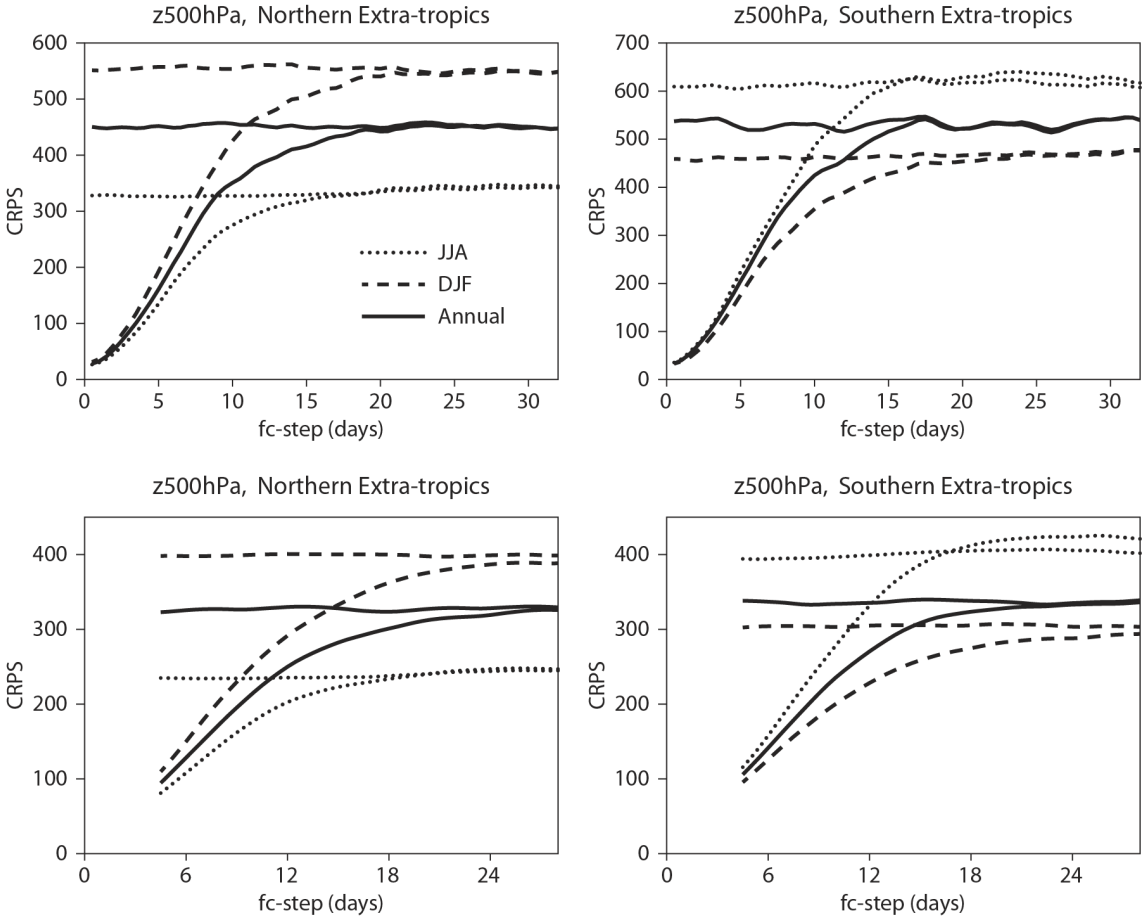

Figure 10. CRPS for Z500 averaged considering 28 cases in JJA (dotted lines), DJF (dashed lines) and the whole year (solid lines), for ENS-BC (lines growing with forecast time) and for the reference ENS-CLI (quasi-horizontal lines). Top-left panel: average CRPS for instantaneous fields truncated at T120 (H0-T120) over NH. Top-right panel: as left panel but for SH. Bottom panels: as top panels but for 192-hour average fields truncated at T120 (H96-T120). Units in all panels are m^2/s^2.

Z500	H0		H96	
	NH	SH	NH	SH
DJF	18.5	19.5	20.0	26.0
JJA	16.0	14.0	18.0	15.0
YEAR	18.5	16.5	19.5	20.0

Table 6. Forecast skill horizons for the probabilistic prediction of the 500 hPa temperature (T850) over NH and SH, averaged considering only 28 cases in DJF, JJA or spanning the whole year, for H0-T120 and H96-T120 fields.

4 Discussion

Section 3 showed that our average forecast skill horizon estimates are sensitive to the spatial and temporal scale of the event of interest, consistent with other published results that there are large-scale, low-frequency phenomena that can be predicted weeks ahead. Our new contributions to previous results are twofold: firstly, the range of predictability has been quantified for the first time in a 'consistent and integrated (seamless) way', applying the same methodology (i.e. considering forecasts generated from the same ensemble, looking at the same fields, using the same metric and applying the same definition of forecast skill limit) to phenomena with different scales. Secondly, these new estimates are based on the ECMWF state-of-the-art medium-range/monthly ensemble forecasts for a whole year (107 cases), bias-corrected using bias estimates computed with 20-years of ensemble reforecasts. In other words, these updated estimates are based on one of (and perhaps) the best available (Buizza 2014) medium-range/monthly ensembles. Results have indicated that the forecast skill horizon ranges from about 2 weeks for smaller-scale, higher-frequency fields, to 4 weeks, and even beyond, for larger-scale, lower-frequency fields.

4.1 How can we reconcile the results with Lorenz (1969) estimate?

A number of factors need to be taken into account to explain why these recent estimates extend the range of predictive skill into several weeks while early estimates by Lorenz (1969a) suggested that perhaps there is a finite time of predictability of about 2 weeks regardless how small the initial uncertainty is in the atmosphere. The mechanisms at work in a fluid of many scales of motion that explains this finite predictability time scale is the upscale growth of errors and the fact that the time scale of error growth decreases with the scale of motion. Our work does not dispute the presence of this mechanism. In fact, later work by Rotunno and Snyder (2008) and Durran and Gingrich (2014) confirmed Lorenz's (1969a) work that we should expect a finite predictability time scale regardless of the amplitude of initial error.

This is linked to the fact that in the meso-scales the kinetic energy spectrum is shallow and depends on horizontal wavenumber as $K^{-5/3}$. On the other hand, in the synoptic and planetary scales the energy spectrum in the atmosphere is known to be steeper, close to K^{-3}. Furthermore, Lorenz (1969a) used a very simple assumption to represent the saturation of errors while Durran and Gingrich (2014) introduced a more refined slowed growth as the errors near saturation. While the latter work clearly explains that butterflies or sea-gulls do not matter in practical terms in present-day numerical weather prediction, no attempt had been made to refine the estimate of the finite time scale of atmospheric predictability. It may be a worthwhile exercise to attempt this with the refined error saturation model of Durran and Gingrich (2014) and using the mixed K^{-3} and $K^{-5/3}$ kinetic energy spectrum observed in the atmosphere. It would be of interest whether this exercise would yield predictability time scales more consistent with Kleeman's (2008) estimate of 45 days for the mid-latitude atmosphere.

We should also expect that details of the range of predictive skill should depend on the processes represented by the numerical model. Some of the more idealized work starting with that of Lorenz (1969a) is based on two-dimensional dynamics only. Most of the previous studies based on more complete global circulation models did not have a convection scheme that could simulate realistically the small-scale variability (Bechtold et al 2004) and the diurnal cycle (Bechtold et al 2013). The inclusion of moist processes is crucial for the predictability in the tropical atmosphere. Phenomena such as blocking and the MJO cannot be simulated by dry dynamics. In the case of the MJO, Vitart et al (2014) showed that improvements in the prediction of the propagation of organized convection in the tropics led to improvements in the skill over Europe. They also concluded that '.. Based on the ability of the IFS to simulate the impact of the MJO on tropical cyclone activities and the skill of the model to predict MJO events, sub-seasonal forecasts of tropical cyclone activity over weekly periods .. are reliable up to week 4 over some basins.' (see their discussion in section 3.1.3).

This result implies that predictive skill estimated using forecasts based on a model that can describe the MJO propagation and its links with the extra-tropics and smaller scale waves, are going to be longer than estimates based on a model that cannot describe them. Convection is just an example and the representation of other processes also plays their role. The simplified predictability studies are useful for understanding particular aspects of the error growth but they cannot provide a complete picture as they lack a state-of-the-art representation of radiation (Morcrette et al 2007), the land-surface and its interaction with the free atmosphere (Balsamo et al 2014), and they are also not coupled to a dynamical ocean and ocean wave model (Vitart et al 2007, 2014; Janssen et al 2013). Similarly, a model that can predict regime transitions, e.g. the onset, evolution and decay of low-frequency events such as Euro-Atlantic blocking events, will have higher predictive skill (Pelly and Hoskins, 2003).

Even though errors will completely saturate in the atmospheric flow over some time scale τ regardless how small the initial error is, there is the possibility that the distribution of weather is still modulated in a predictable manner beyond this time-scale τ by those components of the climate system that have a slower error growth. The components of the earth system that are believed to have a slower growth of errors include the land-surface, sea-ice and the ocean. Recent numerical weather and climate prediction models either include these components or are likely to include these components in the coming years. There is quite a range of predictability studies of the coupled atmosphere-ocean system ranging from highly idealized low-order systems (Vannitsem 2014) to intermediate-complexity systems (Goswami and Shukla 1991; Kleemann and Power 1994) and finally to full atmosphere-ocean global circulation

models (e.g. Collins 2002). While there is no generally established theoretical predictability horizon for coupled atmosphere-ocean system, there is sufficient evidence that the modulation of the atmospheric PDF of weather by the slower ocean extends to months and seasons. Furthermore, recent results indicate that sea-ice prediction may induce predictable signals up to a lead time of three years (Tietsche et al 2014). In those regions and seasons with a sufficiently strong coupling between land surface and atmosphere, the land surface can also carry predictable signals (e.g. Guo et al 2012).

4.2 Why should ensembles be used to estimate predictability?

For this work, the forecast skill horizon has been estimated using an ensemble rather than single forecasts, since they provide a more complete estimate of the future forecast states (Palmer et al 2007, Buizza 2008). We selected a simple metric for measuring the quality of forecasts of scalars, CRPS. The CRPS is a reasonable choice as it does not require a discretisation of the state space and can be computed easily with the available ensemble sizes. However, we cannot exclude that it leads to an underestimation of the actual range of predictive skill that one might be able to establish in theoretical work with much larger ensembles and information theory based approaches such as relative entropy. Further evidence of the concept of the propagation of a 'predictive signal' from the large to the small scales, that counteracts the upscale error propagation, comes from recent work at ECMWF. Vitart et al (2014) reviewed the link between the MJO and Euro-Atlantic predictability, and showed that improvements in the prediction of the propagation of organized convection in the tropics led to improvements in the skill over Europe. This result implies that predictive skill estimated using forecasts based on a model that can describe the MJO propagation and its links with the extra-tropics and smaller scale waves, are going to be longer than estimates based on a model that cannot describe them.

With ensembles, the predictable signal in the modulation of the distribution of weather can be extracted better than with single forecasts as the RMS error of a single forecast will exceed that of the RMS error of the climatological mean long before errors are completely saturated. Ensembles should be able in principle to extract even small modulations in the probability distribution of weather induced by some predictable component of the earth system. In addition, time and space averaging is a basic way to focus on the more predictable components and to filter the less predictable components. The fact that larger-scale/lower-frequency phenomena are more predictable than smaller-scale/higher-frequency phenomena can help us understanding why the forecast skill horizon is now longer than the two weeks estimated in the 1970s-1980s. Shukla (1998) talked about 'predictability in the midst of chaos' to explain how skilful long-range predictions of phenomena like El Nino were possible despite fast error-growth rates from small to large scales. Hoskins (2013) talked about 'discriminating between the music and the noise', and introduced the concept of a predictability chain, whereby, for example, 'a large anomaly in the winter stratospheric vortex gives some predictive power for the troposphere in the following months'.

4.3 What are the main limitations of our CRPS-based definition of skill?

The definition of the forecast skill horizon as done in this work via the CRPS of the bias corrected ensemble (ENS-BC) and a sample from the climatological distribution (ENS-CLI) imply certain limitations that we would like to discuss now.

We measure the forecast skill horizon by using a statistical significance test that determines when the CRPS of ENS-BC is not smaller any more than the CRPS of ENS-CLI. The distribution of CRPS differences in the sample of O(100) cases determines the width of the confidence intervals together with the probability threshold of 0.99 that the actual CRPS difference falls within the confidence interval. The length of the confidence intervals derived from the significance test will depend on the sample size. For a larger (smaller) sample the confidence intervals will shrink (expand). In consequence, the forecast skill horizon might increase if a larger sample were available. Ideally, one would like to have an estimate of the skill in the limit of very large sample sizes. However, there is no obvious way in which this can be achieved.

The skill horizon also depends on the quality of the climatological distribution. In theoretical predictability studies such as that of Kleeman (2008), one can generate very large samples from the model climatology. This permits us to accurately estimate the distribution in principle. Here we are limited by 20 years of re-analysis and the fact that we require chronological sequences while at the same time we need to preserve the annual cycle. This resulted in a climatological ensemble with Mc =100 members. We assume that the sensitivity to our definition of the climatological ensemble is small. For theoretical work using a perfect model context, one could actually imagine situations with infinite forecast skill horizon simply due to the fact that the size of the forecast ensemble Mf is larger than the size Mc of the climatological ensemble. In principle, this systematic bias of the CRPS due to the finite ensemble size can be corrected for (Ferro et al, 2008).

In addition to the finite size of the climatological ensemble, the non-stationarity of the current climate implies that in principle one might detect much larger forecast skill horizons if the signal due to anthropogenic climate forcing is correctly predicted and the reference for the estimation of the forecast skill horizon is based on a climatological distribution from the past 20 years. If one uses a large set of reforecasts to estimate skill, it should be possible to distinguish the skill arising from the ability to predict the average climate change signal from the skill that is due to predicting sub-seasonal to interannual variability. This could be achieved by defining a climate distribution based on data centred on the year of interest (but excluding it of course).

Here, we use a bias correction, which is a very basic ensemble calibration technique. More detailed statistical corrections of the raw model output may be able to further extend the forecast skill horizon, consider e.g. the multivariate regression forecast proposed by DelSole (2005). The degree to which this is possible will be limited obviously by the size of the available training data and their consistency with the real-time forecasts. In the context of a changing climate and a changing observing system there will be ultimately limitations on what can be achieved with more elaborate calibration techniques. Weigel et al (2008) make the point that the skill increase due to time averaging is partly due to the simple fact that longer time windows will include more accurate shorter lead time forecasts in the average. One could accurately quantify the skill increase that is due to this aspect. Our results are consistent with this statement but show that some increase in skill is due to time-averaging the fields (see Sec. 3.2).

We evaluate the average predictability over one year. This has two implications. Firstly, inter-annual variations in predictability could lead to some uncertainty in estimating the multi-year average forecast skill horizon. Secondly, users may also be interested in flow-dependent variations in predictability. It would be of interest to quantify the variations of the forecast skill horizons with flow regimes. Establishing these variations is a significant task beyond the scope of this work. The challenge is that conditioning on particular flow regimes will reduce the sample size and make it harder again to reliably estimate the forecast skill horizon.

Finally, we consider ensemble forecasts skilful if their CRPS is statistically different from the CRPS of the reference forecast, even if the skill level is, in absolute terms, very small. Although this definition has the limitation that the number of users who could exploit this very low level of skill might be very limited, it is an objective way to quantify predictability, usually followed to estimate predictability by comparing the accuracy of a forecast to the one of a well-defined reference.

5 Conclusions

Numerical weather prediction has seen, in the past 25 years, a shift from a 'deterministic' approach, based on single numerical integrations, to a probabilistic one, with ensembles of numerical integrations used to estimate the probability distribution function of forecast states. This shift made it meaningful to extend the forecast length beyond 10 days, and to establish ensemble seasonal forecasting. Our work, which has shown that the probabilistic forecast skill horizon extends beyond two weeks, complements and refines the results of the 1970s-1980s that suggested that errors of instantaneous local forecasts will saturate on a time-scale of about 2 weeks. The assessment of the predictive skill of ECMWF monthly ensembles has indeed indicated that the forecast skill horizon is sensitive to the spatial and temporal scale of the predicted phenomena.

This work has shown that by applying the same methodology (in our case a CRPS-based metric) to measure ensemble skill of forecasts with different spatial and temporal scales, it is possible to make 'seamless' quantitative considerations on what scales can be predicted at different forecast ranges.

Our results, obtained applying the same methodology and skill metric to ECMWF ensemble forecasts with increasingly coarser spatial and temporal scales, indicate that while instantaneous, grid-point fields have forecast skill horizons of between 16-23 days, large-scale, low-frequency filtered fields have forecast skill horizons of beyond 23 days. The forecast skill horizon depends not only of the field's spatial-temporal scale, but also on the variable and area, and on the season. It is worth stressing the fact that, since our work is based on forecasts with a maximum forecast length of 32 days, in some cases we could only state that the skill horizon was beyond the longest available forecast length (which is also a function of the time-averaging period).

Forecast skill horizons beyond 2 weeks are now achievable thanks to major advances in numerical weather prediction. More specifically, they are made possible by the synergies of better and more complete models, which include more accurate simulation of relevant physical processes (e.g. the coupling to a dynamical ocean and ocean waves), improved data-assimilation methods that allowed a more accurate estimation of the initial conditions, and advances in ensemble techniques.

6 Acknowledgements

We would like to thank Erland Källén, Franco Molteni, Alan Thorpe and Frederic Vitart, for discussing the content of this work during its execution, and for providing very valuable comments to an earlier version of this manuscript. Anabel Bowen is thanked for her work that led to improved figures. We would also like to thank the Chief Editor and Prof. D. J. Parker for their editorial work, and Prof. E. Kalnay and Prof. Sir Brian Hoskins for their very valuable reviews that have helped us improving it.

7 References

Balsamo, G., A. Agustì-Panareda, C. Albergel, A. Beljaars, S. Boussetta, E. Dutra, T. Komori, S. Lang, J. Muñoz-Sabater, F. Pappenberger, P. de Rosnay, I. Sandu, N. Wedi, A. Weisheimer, F. Wetterhall, and E. Zsoter, 2014: Representing the Earth surfaces in the Integrated Forecasting System: Recent advances and future challenges. ECMWF Research Department Technical Memorandum n. 729, pp 50 (available from ECMWF, Shinfield Park, Reading RG2-9AX, U.K.; see also www.ecmwf.int).

Bechtold, P., M. Köhler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. Rodwell, F. Vitart, and G. Balsamo, 2008: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. Q. J. R. Meteorol. Soc., 134, 1337–1351.

Bechtold, P., N. Semane, P. Lopez, J.-P. Chaboureau, A. Beljaars, and N. Bormann, 2013: Representing equilibrium and non-equilibrium convection in large-scale models. ECMWF Research Department Technical Memorandum n. 705, pp 27 (available from ECMWF, Shinfield Park, Reading RG2-9AX, U.K.; see also www.ecmwf.int).

Berner J., G. Shutts, M. Leutbecher, and T.N. Palmer, 2008: A Spectral Stochastic Kinetic Energy Backscatter Scheme and its Impact on Flow-dependent Predictability in the ECMWF Ensemble Prediction System. J. Atmos. Sci., 66, 603-626.

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. Q. J. Roy. Meteorol. Soc., 116, 867-912.

Brown, T. A., 1974: Admissible scoring systems for continuous distributions. Manuscript P-5235, The Rand Corporation, Santa Monica, CA, pp. 22. Available from The Rand Corporation, 1700 Main Street, Santa Monica, CA 90470-2138.

Buizza R., and T. N. Palmer, 1995: The singular vector structure of the atmospheric general circulation. J. Atmos. Sci., 52, 1434-1456.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., and Vitart, F., 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). Q. J. Roy. Meteorol. Soc., 133, 681-695.

Buizza, R., 2008: The value of Probabilistic Prediction. Atmos. Sci. Lett., 9, 36-42.

Buizza, R., Leutbecher, M., and Isaksen, L., 2008: Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. Q. J. R. Meteorol. Soc., 134, 2051-2066.

Buizza R., M. Miller, and T. N. Palmer: 1999: Stochastic representation of model uncertainties in the ECMWF EPS. Q. J. Roy. Meteor. Soc., 125, 2887-2908.

Buizza, R., 2014: The TIGGE ensembles. ECMWF Research Department Technical Memorandum n. 739, pp 52 (available from ECMWF, Shinfield Park, Reading RG2-9AX, U.K.; see also www.ecmwf.int).

Charney, J. G., et al, 1966: The feasibility of a global observation and analysis experiment. Bull. Amer. Meteorol. Soc., 47, 200-220.

Collins, M., 2001: Climate predictability on interannual to decadal time scales: the initial value problem. Clim. Dyn., 19, 671–692.

Dalcher, A., and Kalnay, E., 1987: Error growth and predictability in operational ECMWF forecasts. Tellus, 39A, 474-491.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F., 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc., 137, 553–597.

DelSole, T., 2005: Predictability and information theory: Part II: Imperfect forecasts. J. Atmos. Sci., 62, 3368–3381.

DelSole, T. and M. K. Tippet, 2007: Predictability: Recent insights from information theory. Rev. Geophys., 45, RG4002.

Durran, D. R. and M. Gingrich, 2014: Atmospheric predictability: Why butterflies are not of practical importance. J. Atmos. Sci., 71, 2476–2488.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. J. Appl. Meteorol., 8, 985-987.

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

Gneiting, T. and A. E. Raftery, 2007: Strictly proper scoring rules, prediction and estimation. J. Amer. Stat. Assoc., 102, 359–378.

Goswami, B.N. and J. Shukla, 1991: Predictability of a coupled ocean-atmosphere model. J. Clim., 4, 3–22.

Guo, Z., P. A. Dirmeyer, T. DelSole, and R. D. Koster, 2012: Rebound in Atmospheric Predictability and the Role of the Land Surface. J. Clim., 25, 4744–4749.

Hagedorn, R. R. Buizza, T. M. Hamill, M. Leutbecher, and T.N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. Q. J. R. Meteorol. Soc., 139, 1814–1827.

Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. Weather and Forecasting, 15, 559-570.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. Mon. Wea. Rev., 124, 1225-1242.

Hoskins, B. J., 2013: Review article: the potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. Q. J. R. Meteorol. Soc., 139, 573-584.

Hurrell, J.W, Kushnir, Y., Ottersen, G., and Visbeck, M., 2003: An overview of the North Atlantic Oscillation". Geophysical Monograph 134, American Geophysical Union.

Janssen, P., J.-R. Bidlot, S. Abdalla and H. Hersbach, 2005: Progress in ocean wave forecasting at ECMWF. ECMWF Research Department Technical Memorandum n. 478. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

Janssen, P.A.E.M., O. Breivik, K. Mogensen, F. Vitart, M. Balmaseda, J.-R. Bidlot, S. Keeley, M. Leutbecher, L. Magnusson and F. Molteni, 2013: Air-sea interaction and surface waves. ECMWF Research Department Technical Memorandum n. 712. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

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

Kleeman, R., 2008: Limits, variability, and general behaviour of statistical predictability of the midlatitude atmosphere. J. Atmos. Sci., 65, 263–275.

Kleeman, R. and S. B. Power, 1994: Limits to predictability in a coupled ocean-atmosphere model due to atmospheric noise. Tellus, 46A, 529–540.

Lalaurette, F. 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. Q. J. Roy. Meteorol. Soc., 129, 3037-3057.

Leutbecher, M, and Palmer T. N., 2008: Ensemble forecasting. J. Comp. Phys. 227, 3515–3539.

Lorenz, E. N., 1969a: The predictability of a flow which possess many scales of motion. Tellus, XXI, 3, 289-307.

Lorenz., E. N., 1969b: Atmospheric predictability as revealed by naturally occurring analogues. J. Atmos. Sci., 26, 636-646.

Lorenz, E. N., 1969c: How much better can weather prediction become? Technology Review, 39-49. Accessible from the MIT library (http://eaps4.mit.edu/research/Lorenz/publications.htm).

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. Tellus, 34, 505–513.

Lorenz, E. N., 1984: Estimates of atmospheric predictability at medium-range. Predictability of Fluid Motions (G. Holloway and B. West, eds.), New York, American Institute of Physics, 133-139.

Madden, R. A. and P.R. Julian, 1971: Detection of a 40-50 day oscillation in the zonal wind in the tropical Pacific. J. Atm. Sci., 5, 702-708.

Mogensen, K., M. Alonso Balmaseda, A. Weaver, 2012a: The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for System 4. ECMWF Research Department Technical Memorandum n. 668, pp 59. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

Mogensen, K., S. Keeley and P. Towers, 2012b: Coupling of the NEMO and IFS models in a single executable. ECMWF Research Department Technical Memorandum n. 673, pp. 23. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

Molteni, F., U. Cubasch and S. Tibaldi, 1987: 30 and 60-day forecast experiments with the ECMWF spectral models. Proceedings of the first ECMWF Workshop on Predictability in the medium and extended range (ECMWF, Reading, U.K., 17-19 Mar. 1986), pg. 51-108. Available from ECMWF, Shinfield Park, Reading RG2-9AX.

Molteni, F., T. Stockdale, M. Balmaseda, G. Balsamo, R. Buizza, L. Ferranti, L. Magnusson, K. Mogensen, T. Palmer and F. Vitart, 2011: The new ECMWF seasonal forecast system (System 4). ECMWF Research Department Technical Memorandum n. 656, pp. 49. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

Morcrette, J.J., Bechtold, P., Beljaars, A., Benedetti, A., Bonet, A., Doublas-Reyes, F., Hague, J., Hamrud, M, Haseler, J., Kaiser, J. W., Leutbecher, M., Mozdzynsky, G., Razinger, M., Salmond, D., Serrar, S., Suttie, M., Tomkins, A., Untch, A., and Weisheimer, A., 2007: Recent advances in radiation transfer parameterizations. . ECMWF Research Department Technical Memorandum n. 593, pp. 50. Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also http://old.ecmwf.int/publications/).

Murphy, A. H., 1971: A note on the ranked probability score. J. Appl. Meteorol., 10, 155-156.

Palmer, T.N. and D.L.T. Anderson, 1994: The prospects for seasonal forecasting - A review paper. Q. J. Roy. Meteorol. Soc., 120, 755-793.

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., and Tribbia, J., 1993: Ensemble prediction. Proceedings of the ECMWF Seminar on Validation of models over Europe: vol. 1. ECMWF, Shinfield Park, Reading RG2-9AX, UK, 285 pp (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Palmer, T N, Buizza, R., Leutbecher, M., Hagedorn, R., Jung, T., Rodwell, M, Vitart, F., Berner, J., Hagel, E., Lawrence, A., Pappenberger, F., Park, Y.-Y., van Bremen, L., Gilmour, I., and Smith, L., 2007: The ECMWF Ensemble Prediction System: recent and on-going developments. ECMWF Research Department Technical Memorandum n. 540, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pp. 53.

Palmer, T.N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer and A Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Research Department Technical Memorandum n. 598, pp. 42 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Pelly, J. L, and Hoskins, B. J., 2003: How well does the ECMWF ensemble prediction system predict blocking? Q. J. Roy. Meteorol. Soc., 129, 1683-1702.

Rotunno, R. and C. Snyder, 2008: A generalization of Lorenz's model for the predictability of flows with many scales of motion. J. Atmos. Sci., 65, 1063–1076.

Shukla, J., 1981: Dynamical predictability of monthly means. J. Atmos. Sci., 38, 12, 2547-2572.

Shukla, J., 1998: Predictability in the midst of chaos: a scientific basis for climate forecasting. J. Atmos. Sci., 38, 2547-2572.

Shutts, G. 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. Q. J. Roy. Meteor. Soc., 131, 3079-3100.

Shutts, G., Leutbecher, M., Weisheimer, A., Stockdale, T., Isaksen, L., and Bonavita, M, 2011: Representing model uncertainty: stochastic parametrizations at ECMWF. ECMWF Newsletter, 129, 19-24.

Simmons, A., J., and Hollingsworth, A., 2002: Some aspects of the improvement in skill of numerical weather prediction. Q. J. Roy. Meteorol. Soc., 128, 647-677.

Smagorinsky, J, 1969: Problems and promises of deterministic extended-range forecasting. Bull. Amer. Meteor. Soc., 50, 286–311.

Stockdale, T.N., D. L. T. Anderson, M. A. Balmaseda, F. Doblas-Reyes, L. Ferranti, K. Mogensen, T.N. Palmer, F. Molteni and F. Vitart, 2011: ECMWF seasonal forecast system 3 and its prediction of sea surface temperature. Clim. Dyn., 37, 455-471.

Tietsche, S., J.J. Day, V. Guemas, W.J. Hurlin, S.P.E. Keeley, D. Matei, R. Msadek, M. Collins, and E. Hawkins, 2014: Seasonal to interannual Arctic sea ice predictability in current global climate models. Geophys. Res. Lett., 41, 1035–1043.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. Bull. Amer. Meteor. Soc., 74, 2317-2330.

Toth, Z., and Kalnay, E., 1997: Ensemble Forecasting at NCEP and the breeding method. Mon. Wea. Rev., 125, 3297-3319.

Tracton, M. S., and Kalnay, E., 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. Weather and Forecasting, 8, 379-398.

Vannitsem, S, 2014: Dynamics and predictability of a low-order wind-driven ocean-atmosphere coupled model. Clim. Dyn., 42, 1981–1998.

Vitart, F, 2004: Monthly forecasting at ECMWF. Mon. Wea. Rev., 132, 2761-2779.

Vitart, F., S. Woolnough, M. A. Balmaseda and A. Tompkins, 2007: Monthly forecast of the Madden-Julian Oscillation using a coupled GCM. Mon. Wea. Rev., 135, 2700-2715.

Vitart, F., Buizza, R., Alonso Balmaseda, M., Balsamo, G., Bidlot, J. R., Bonet, A., Fuentes, M., Hofstadler, A., Molteni, F., and Palmer, T. N., 2008: The new VAREPS-monthly forecasting system: a first step towards seamless prediction. Q. J. Roy. Meteorol. Soc., 134, 1789-1799.

Vitart, F., 2013: Evolution of ECMWF sub-seasonal forecast skill scores over the past 10 years. ECMWF Research Department Technical Memorandum n. 694, pp. 28 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Vitart, F., G. Balsamo, R. Buizza, L. Ferranti, S. Keeley, L. Magnusson, F. Molteni, and A. Weisheimer, 2014: Sub-seasonal predictions. ECMWF Research Department Technical Memorandum n. 734, pp. 47 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK).

Weigel, A., D. Baggenstos, M. A. Liniger, F. Vitart, and C. Appenzeller, 2008: Probabilistic verification of monthly temperature forecasts. Mon. Wea. Rev., 136, 5162–5182.

Zsoter, E., 2006: Recent developments in extreme weather forecasting, ECMWF Newsletter, 107, 8-17.

Zsoter, E., Buizza, R., and Richardson, D., 2009: 'Jumpiness' of the ECMWF and UK Met Office EPS control and ensemble-mean forecasts'. Mon. Wea. Rev., 137, 3823-3836.