

**The value of a variable resolution approach
to numerical weather prediction**

Roberto Buizza

**European Centre for Medium-Range Weather Forecasts, Reading UK
(www.ecmwf.int)**

Submitted to the AMS *Monthly weather Review*, 9 May 2009.

Revised version, 2 October 2009.

Words: 7292 (text without figure captions).

Figures: 16.

Key words: numerical weather prediction, variable resolution, ensemble prediction, error growth, predictability.

Corresponding author address: Dr R. Buizza, ECMWF, Shinfield Park, Reading, RG2-9AX, UK
(email: Buizza@ecmwf.int).

Abstract

It is shown that a numerical weather prediction system with variable resolution, higher in the early forecast range and lower afterwards, provides more skilful forecasts than a system with a constant resolution system. Results indicate that the advantage can be detected also beyond the time when the resolution is truncated (truncation time).

Forecasts generated with a T399 spectral truncation up to forecast day 3 and a T255 truncation from day 3 to day 8 (VAR3), are compared with forecasts generated with a constant T319 truncation. Firstly, forecasts are verified in an Idealized Model Error (IME) scenario against higher resolution, T799 simulations. In this scenario, VAR3 outperforms the T319 system beyond the day-3 truncation time for the entire 8-day forecast range, with differences statistically significant at the 5% level. Secondly, forecasts are verified in a realistic scenario against T799 analyses. In this case, although the advantage of VAR3 can still be detected beyond day-3, it is less evident and not statistically significant. Forecast error spectra indicate that using a higher resolution model during the first forecast days improves the forecasts of the large-scales, thus helping to maintain the advantage of the variable resolution system beyond the truncation time.

VAR3 and T319 ensembles are also compared with forecasts with a T255, T399 and T799 constant resolution. The predictability ‘gain’ of all ensemble configurations is measured with respect to the reference constant T255 configuration. Results show that, in the realistic scenario, VAR3 gives gains 50-75% higher than T319, and 50-75% lower than T799.

1 The variable resolution approach to weather prediction

Computing resources limit the configuration of operational weather forecasting systems. At the time of writing (September 2009), for example, the European Centre for Medium-Range Weather Forecasts (ECMWF) allow computing resources to run the following suite of operational systems twice-a-day, at 00 and 12 UTC: a T_L799L91 (spectral triangular truncation with total wave number 799 and 91 vertical levels, and linear grid in physical space) 12-hour 4-dimensional variational data assimilation system, a T_L799L91 10-day forecast, and a variable resolution ensemble prediction system (VAREPS). VAREPS (Buizza *et al* 2007) includes 51 members run with a T_L399L62 resolution from day 0 to day 10, and a lower T_L255L62 resolution from day 10 to day 15, with the ensemble starting at 00 UTC on every Thursday extended to 32 days to cover the monthly forecast range (Vitart *et al* 2008).

VAREPS was designed to resolve the smallest possible¹ scales up to the forecast time when their inclusion had a positive impact on the prediction of both the small and the synoptic scales, and not to resolve them afterwards, when including them had a smaller, less detectable impact on the synoptic scales. Buizza *et al* (2007) compared results based on a preliminary version of VAREPS, with a T_L399L40 resolution up to day 7 and T_L255L40 from day 7 to day 14, with forecasts generated using a constant resolution T_L319L40 EPS² (these two systems required a similar

¹ Possible in the sense that the CPU time required to complete the forecast integrations of the whole ensemble system would remain within acceptable limits.

² Note that in the VAREPS operational at the time of writing, resolution is truncated at day 10 instead of day 7, and the number of vertical levels is 62 instead of 40.

amount of computing resources). In the early forecast range, *Buizza et al (2007)* detected a clear advantage of VAREPS, especially in the prediction of mean-sea-level pressure in extreme weather conditions such as ones associated with tropical storms, but beyond the day 7 truncation time they found only limited evidence that VAREPS was providing better upper-level forecasts (say geopotential height at 500 hPa or temperature at 850 hPa). Although they could not find that VAREPS was statistically significantly better than the constant-resolution system beyond the truncation limit, they concluded that overall it was performing better than a constant resolution, T_L319 system with comparable cost. VAREPS became part of the ECMWF operational suite in September 2006.

It should be mentioned that a variable resolution approach had been used earlier at the National Centers for Environmental Prediction ensemble prediction system (NCEP, Washington).

Szunyogh & Toth (2002) provided evidence of the value of a variable resolution approach in the NCEP ensemble prediction system (*Tracton & Kalnay 1993, Toth & Kalnay 1997*). *Szunyogh & Toth (2002, see their discussion in section 2)*, concluded that the variable resolution approach was ‘... based on the experience that increased horizontal resolution for the first few days of model integration has significant positive impact on forecast quality for the entire forecast range’.

This work investigates in more details whether a variable resolution approach performs better than a constant resolution approach. Forecasts for different variables are verified both in a realistic framework against analyses, as in *Buizza et al (2007)*, and in an idealized framework. The comparison of the results obtained in the realistic and idealized frameworks will help understanding why *Buizza et al (2007)* found only limited statistical significance.

The key question that is addressed in this work is the following: ***which of two comparable-cost configurations gives the best medium-range forecasts: a constant resolution configuration, or one with a variable resolution, higher in the earlier forecast range and lower afterwards?***

Results based on upper-level fields diagnostics, will show that the benefit of using higher resolution in the early forecast range extends beyond the truncation time. The benefit will be more evident in an Idealized Model Error (IME) scenario, when a T_L799L62 forecast is used as verification, but less clear when analyses are used as verification. The comparison of forecast error spectra, and the use of a 3-parameter forecast error growth model will help understanding the reasons of this difference. Results will support our explanation that ‘model imperfections mask’ the positive signal of a variable resolution approach, and will help understanding why *Buizza et al (2007)* found differences with only a limited statistical significance.

Section 2 describes the experimental set-up and the verification measures used to assess the forecast performance. Sections 3 and 4 present average results obtained in the IME and in the realistic scenario, respectively. Section 5 discusses in more details the differences presented in sections 3 and 4, and explains the differences between the VAR3 and the T319 performances considering the time evolution of forecast error spectra, and using a 3-parameter model of forecast error growth. Section 6 compares the gains in predictability of different ensemble configurations. Finally, conclusions are drawn in section 7.

2 Experimental set-up and methodology

Ensemble forecasts have been run for an entire season, winter 2007-08 (from the 1st of December 2007 to the 28th of February 2008). However, to limit the amount of computer resources required to complete all experiments to a reasonable number, 5-member instead of 51-member ensembles

have been run for only 8 instead of the 15 days which are currently forecast in the operational ECMWF ensemble. More precisely, ensembles with one control member starting from the unperturbed analysis and 4 perturbed members have been run for up to 8 days in the following five configurations:

- *T255*: T_L255L62(day 0–8), with a 1800 second time step;
- *T319*: T_L319L62(day 0–8) with a 1800 second time step;
- *T399*: T_L399L62(day 0–8) with a 1200 second time step;
- *T799*: T_L799L62(day 0–8) with a 720 second time step;
- *VAR3*: T_L399L62(day 0–3) with a 1200 second time step and T_L255L62(day 3–8) with a 1800 second time step;
- *VAR5*: T_L399L62(day 0–5) with a 1200 second time step and T_L255L62(day 3–8) with a 1800 second time step;

Figure 1 shows the amount of CPU time that each configuration required to produce 8-day forecasts, relative to the T319 configuration: note that VAR3 requires about 25% more CPU than T319 (if these forecasts were to be extended to 10 days instead of 8, the difference would be 15%, while if they were to be extended to 15 days, which is the current forecast length of the ECMWF EPS, VAR3 would require 2% less CPU). It might be interesting for the reader to know that the experiments used in this work required a very large amount of CPU to be completed, equivalent to the CPU required running one and a half years of 10-day forecasts at T_L799L62 resolution.

All ensembles were run with the same model cycle (model cycle 33r2, which was operational at ECMWF between the 5th of June and the 6th of November 2007), all starting from the same initial

conditions, the control forecasts from the ECMWF high-resolution (unperturbed) operational analyses, and the perturbed forecasts from perturbed initial conditions defined using the operational singular vectors (*Buizza & Palmer 1995*). The perturbed initial conditions were generated by combining the leading 50 singular vectors computed over the Northern Hemisphere (NH) and the Southern Hemisphere (SH) extra-tropics, and the leading 10 singular vectors covering between 1 and 5 regions of the tropics where tropical depressions were detected in the analysis. Each perturbed forecast was also integrated in time using a stochastic scheme designed to simulate the effect of random model errors due to physical parameterisation schemes (*Buizza et al 1999*). The reader is referred to *Palmer et al (2007)* for a recent review of the ECMWF ensemble prediction system.

As mentioned in the Introduction, forecasts generated in configurations T255, T319, T399, VAR3 and VAR5 have been verified in the IME scenario against the T799 control forecast, and in a realistic scenario against ECMWF T_L799L91 operational analyses. Attention has been focused on three variables: the 500 and the 1000 hPa geopotential heights, and the 850 hPa temperature defined on a 2.5 degree regular latitude/longitude grid. None of the forecast fields have been biased corrected and/or recalibrated, i.e. forecasts have been used as produced by the model. It is worth to mention that *Buizza et al (2007)* found that although truncation from a high to a low resolution does not have any impact on upper air fields, it has an impact on some low-level variables such as divergence, vertical velocity, and precipitation. This led to the decision to implement the ECMWF operational VAREPS with a 24-hour overlap period. The variable resolution experiments analyzed in this work did not have any overlap period, and did not use any technique to reduce the impact of truncation.

Different forecast products have been assessed using a wide range of accuracy measures:

- Single forecasts defined by the ensemble control or by the ensemble-mean (defined as the average of the 5 ensemble members) have been verified using the root-mean-square error and the anomaly correlation coefficient.
- Probabilistic forecasts defined by the 5-member ensembles have been verified using the ranked probability skill score (RPSS, *Wilks 1995*) computed with respect to the sample climatology, the Brier skill score (BSS, *Brier 1950, Wilks 1995*) and the area under the relative operating characteristic curve (ROCA) computed in terms of the standard normal deviates (*Swets 1986*; see Appendix B in *Buizza et al 2007* for a detailed description of the method used to compute it).

The statistical significance of differences between two forecast systems has been assessed considering the non-parametric rank-sum Mann-Whitney-Wilcoxon test³ (*Wilks 1995*; see also Appendix A in *Buizza et al 2007* for a detailed discussion of the method used to compute it).

When there was the need to summarize the relative difference between two configurations, the two time-integrated indices that were first introduced in *Buizza et al (2007)* have been used. The indices have been defined as follows. Consider a forecast at time t given by two ensemble systems A and B, and verification measures $sc(A,t)$ and $sc(B,t)$, (e.g., the root-mean-square-error

³ Two key advantages of the rank-sum Mann-Whitney-Wilcoxon test are that (i) being non-parametric, it does not assume that the data distribution has any specific form, and (ii) it is 'resistant', i.e. its value is not affected by a few outliers. Given the distributions of scores of two different forecasts, the test assesses whether they belong to the same underlying distribution or not. The null hypothesis is that the two distributions of scores are from the same underlying distribution, and the test value is the probability that the two distributions are samples from the same underlying distribution.

of the control forecasts, or the RPSS of the probabilistic forecasts). The relative performance of system A compared with system B has been measured using the relative difference $resc(A,B;t)$:

$$resc(A, B; t) = \frac{sc(A; t) - sc(B; t)}{sc(B; t)},$$

The first index I_1 is defined as the T_1 -to- T_2 time-average relative difference

$$I_1 \equiv \langle resc(A, B) \rangle_{T_1, T_2} = \frac{1}{(T_2 - T_1)} \int_{T_1}^{T_2} \frac{sc(A; t) - sc(B; t)}{sc(B; t)} dt .$$

The second index I_2 expresses the same average difference in terms of the gain in forecast skill expressed in hours. More specifically, it gives the average difference in terms of hours of forecast that can be gained by using A instead of B:

$$I_2 \equiv \langle resc - h(A, B) \rangle_{T_1, T_2} = \frac{12}{(T_2 - T_1)} \int_{T_1}^{T_2} \frac{[sc(A; t) - sc(B; t)]}{sc(B; t) - sc(B; t - 12h)} dt .$$

This index will be used in section 6 to compare the average predictability gains between forecast day 3 and 8 of configurations T319, VAR3, VAR5, T399 and T799 compared with the reference, lower resolution T255 configuration.

3 Average performance of constant and variable resolution ensembles in the idealized model error (IME) and realistic scenarii

In this section, average (computed for the whole 90-day period) single and probabilistic forecasts from different ensemble configurations have been verified first within the IME scenario, and then in a realistic scenario. Although results have been produced for all three variables, for reasons of space only diagnostics relative to the 500 hPa geopotential height are shown, but similar conclusions could be drawn by considering the other two variables.

3.1 Single control and ensemble-mean forecasts verified in the IME scenario

In the IME scenario it has assumed that the model is ‘perfect’ apart for the fact that it is lacking resolution, and accuracy measures have been computed using the T799 control forecast as verification. Figure 2 shows the root-mean-square-error (rmse) of the control and the ensemble-mean forecasts of the 500 hPa geopotential height over the Northern Hemisphere (NH) for four ensemble configurations, VAR3, T319, T799 and T399, verified against the T799 control forecasts. Since differences are rather small, Fig. 2 also shows the relative difference $resc(A,B;t)$ computed using the T399 forecasts as reference. The comparison of the performance of the different ensembles allows us to understand the impact of using an increased resolution (T399 versus T319) or of using a variable resolution approach (VAR3 versus T319).

The top-left panel of Fig. 2 shows that the rmse of the T319 control forecast is the largest (the rmse of the T799 control forecast is zero by construction). The bottom-left panel of Fig. 2 shows that, compared with the T399 control, the T319 control has initially a 22% larger rmse, with the difference decreasing gradually to $\sim 15\%$ at forecast day 8. Note that the VAR3 control performs very similarly to the T399 control even after the day-3 truncation from T399 to T255 (differences are of the order of 2-3%). The right panels of Fig. 2 show the corresponding results for the ensemble-mean forecasts. The top-right panel of Fig. 2 shows that the ensemble-mean of the T799 ensemble has the lowest rmse, as expected, starting from zero since the EPS initial perturbations are symmetric and thus at $t=0$ the ensemble-mean coincides with the control. The rmse of the T799 ensemble-mean gives a lower bound that should be expected from an idealized 5-member ensemble system with initial perturbations scaled to perfectly represent the initial uncertainty. The difference between the rmse of the ensemble-mean of the T799 ensemble and the

other ensembles gives a measure of the relative impact on forecast error of using a too coarse resolution (T399 or T319 instead of T799). The bottom-right panel of Fig. 2 shows that, compared with the T399 ensemble-mean, the T319 ensemble-mean has initially a 20% larger rmse, with the difference decreasing gradually to $\sim 5\%$ at forecast day 8. Again, compared with the T399 ensemble-mean, the VAR3 ensemble-mean has only a slightly larger rmse, while the T799 ensemble-mean has initially a 50% smaller rmse, which increases gradually to being $\sim 20\%$ smaller at forecast day 8. Overall, the differences between the ensemble-mean forecasts (Fig. 2, right panels) are smaller than the differences between the control forecasts (Fig. 2, left panels).

Results for the Southern Hemisphere (SH), shown in Fig. 3, lead to similar conclusions, with differences between the T319 and the other ensemble configurations been slightly larger.

3.2 Probabilistic forecasts verified in the IME scenario

Figure 4 shows the rank probability skill score (RPSS, which is the equivalent of the rmse for probabilistic forecasts, see *Wilks* 1995 for a definition) over both the NH and the SH for ensemble configurations VAR3, T319, T799 and T399 verified in the IME scenario against the T799 control forecasts. Figure 4 also shows the relative difference $resc(A,B;t)$ computed using T399 as reference. The top panels of Fig. 4 show that, over both hemispheres, unsurprisingly, the T799 has the highest and that the T319 the lowest score. The bottom panels of Fig. 4 show that the differences between the four configurations are smaller than the differences detected for the single control or ensemble-mean forecasts. The T799 ensemble has RPSS values that are initially very close to the others, becoming $\sim 4\%$ higher than the others by forecast day 8. The differences between the T319 ensemble and the others is very small, with RPSS values being less than 1% smaller than the ones of the T399 ensemble.

3.3 Single control and ensemble-mean forecasts verified in the realistic scenario

The performance of the single control and ensemble-mean forecasts is now assessed in the realistic scenario against ECMWF T799 analyses. Figure 5 shows the rmse of the control and the ensemble-mean forecasts of the 500 hPa geopotential height over the NH, and the relative difference $resc(A,B;t)$ computed using the T399 forecasts as reference. Compared with the results obtained in the IME scenario (Fig. 2), differences are about 10-times smaller. The bottom panels of Fig. 5 shows that the T319 control and ensemble-mean forecasts has rmse values which are only few percentages higher than the T399 ensemble, and that the T799 control and ensemble-mean forecasts have rmse values only few percentages lower than the T399 ensemble. Similar results are obtained for the SH (Fig. 6).

3.4 Probabilistic forecasts verified in the realistic scenario

Figure 7 shows that, considering the probabilistic forecasts of 500 hPa geopotential height verified against ECMWF T799 analyses, the four ensembles have very similar RPSS, with differences remaining smaller than 1% for the whole forecast range for both hemispheres.

4 Statistical significance of VAR3-T319 forecast differences

The results discussed in section 3 have indicated firstly that differences between ensemble performances that are large in the IME scenario, almost disappear completely in the realistic case. Secondly, they have shown that in the IME scenario differences between the ensemble configurations are larger if one considers single than probabilistic forecasts. Thirdly, considering

the two ensemble configurations with comparable CPU requirements, VAR3 and T319, they have shown that the VAR3 configuration performs better. In this section, attention is focused on these two configurations: their average performance is compared, and the statistical significance of their differences is assessed by computing the value of the non-parametric rank-sum Mann-Whitney-Wilcoxon test (RMWW).

4.1 Statistical significance of the differences between VAR3 and T319 single control and ensemble-mean forecasts verified in the IME scenario

The top-left panel of Fig. 8 shows the relative difference $resc(A,B;t)$ (see Section 2 for its definition) between the rmse of the control forecast of the VAR3 ensemble over NH, computed using the T319 control forecasts as reference. The bottom-left panel shows the corresponding values for the SH, and the right-panels show the corresponding results for the ensemble-mean forecasts. Figure 8 shows that the relative differences are negative, indicating that the rmse of the T319 forecasts are larger, in agreement with the results discussed in section 3. Figure 8 also shows that the differences are all statistically significant at the 3% level (the RMWW value is always lower than 3%, more precisely lower than 0.5% for the control forecasts and lower than 3% for the ensemble-mean forecasts, indicating that there is less than 3% probability that the scores of the VAR3 and the T319 ensembles were drawn from the same distribution). Figure 8 also confirms that the differences are larger for the single control forecasts than for the single ensemble-mean forecasts. These results indicate that for single control and ensemble-mean forecasts, VAR3 outperforms the constant resolution T319 system even beyond the day-3 truncation time.

4.2 Statistical significance of the differences between VAR3 and T319 probabilistic forecasts verified in the IME scenario

The top-left panel of Fig. 9 shows the relative difference $resc(A,B;t)$ between the RPSS of the probabilistic forecasts of the VAR3 ensemble computed over NH, using the T319 control forecasts as reference. The bottom-left panel shows the corresponding values for the SH, and the right-panels show the corresponding results for another measure of the accuracy of a probabilistic forecast, the area under the relative operating characteristics (ROCA, which is a measure of the ability of a forecasting system to discriminate between occurrence and non-occurrence of a forecast event). Figure 9 shows that the relative differences are small but positive, indicating that the RPSS and the ROCA of the VAR3 forecasts are larger, i.e. better, in agreement with the results discussed in section 3. Figure 9 also shows that the differences are all statistically significant at the 4% level for the whole forecast range (the RMWW value is always lower than 4%, more precisely lower than ~1% for the RPSS and lower than ~4% for the ROCA). The comparison between Figs. 8 and 9 confirms that differences are smaller for probabilistic than single forecasts. These results indicate that also for probabilistic forecasts, although differences are smaller than for single forecasts, the variable resolution VAR3 system outperforms the constant resolution T319 system beyond the day-3 truncation time.

4.3 Statistical significance of the difference between single VAR3 and T319 control and ensemble-mean forecasts verified in the realistic scenario

Figure 10 shows the differences in the skill of single forecasts verified against ECMWF T799 analyses. Figure 10 shows that the relative differences are still negative, thus indicating that the VAR3 forecasts have lower rmse, but the value of the RMWW test is higher than the 5% level obtained for forecasts verified in the IME scenario (Fig. 8). This indicates that although there is

an advantage of the variable resolution system even when forecasts are verified against analyses, the differences are not statistically significant at the 5% level.

4.4 Statistical significance of the difference between VAR3 and T319 probabilistic forecasts verified in the realistic scenario

Figure 11 shows the differences in the skill of probabilistic forecasts verified against ECMWF T799 analyses. Figure 11 shows that the differences are much smaller than in the IME scenario (Fig. 9). Only for the RPSS over SH between forecast day 4 and 8 (see bottom-left panel of Fig. 11), the relative differences are between 0.5% and 1%. As it was the case for the single forecasts verified in the realistic scenario, the RMWW test values are always above 5%, indicating that the differences are not statistically significant at the 5% level.

5 Interpretation of the differences in VAR3 and T319 forecast error

The results discussed in sections 3 and 4 indicate that VAR3 forecasts are more accurate than T319 forecasts for the whole forecast range, with larger and statistically significant differences detected in the IME scenario, when forecasts are verified against T799 control forecasts, but with smaller and not statistically significant differences when forecasts are verified in the realistic scenario against ECMWF T799 analyses.

The interpretation of these results is the following:

- In the IME scenario, the only source of forecast error is due to the fact that forecasts use a resolution lower than the one used in the verification (T319, or T399-T255 in VAR3,

- compared with T799). Error propagates upscale more rapidly in the T319 than in the VAR3 configuration in the short forecast range, say up to forecast day 3, thus making the T319 error larger for all waves. At day 3, when the VAR3 resolution is reduced from T399 to T255, the error difference is already rather large, and even if the VAR3 forecast is performed with a lower resolution than the T319 one, the VAR3 forecast error remains lower. This interpretation is further discussed in section 5.1, where ensemble-mean forecast error spectra are compared, and in section 5.2, where a 3-parameter error growth model proposed by *Simmons & Holligsworth (2001)* is fitted to the forecast errors.
- In the realistic scenario, forecasts errors are due not only to T799 versus T399/T319/T255 resolution differences, but also to the fact that, independently of resolution, the ‘model’ describes only approximately reality (model ‘imperfection’). In this realistic scenario, the differences induced by using a T319 or a T399 resolution in the short forecast range are much smaller than in the idealized scenario because they are ‘masked’ by the contribution to the forecast error due to model imperfection. In other words, the contribution to the forecast error due to the resolution difference is not any more dominant. As a result, the VAR3 and T319 configurations perform similarly even during the first 3 days. The difference between the results obtained in the IME and the realistic scenario are confirmed by the discussion reported in section 5.2.

5.1 Comparison of the spectra of the ensemble-mean forecast errors

Figure 12 shows the spectra of the ensemble-mean forecast error computed in the IME case at the truncation time ($t+72h$), and 36 hours before and after the truncation ($t+36h$ and $t+108h$). The left panels display the full spectra, and the right panels show the relative difference between the VAR3 and the T319 spectra, $diff(n) = (sp_{VAR3}(n) - sp_{T319}(n)) / sp_{T319}(n)$. Figure 12 explains

the time evolution of the forecast error spectrum: at t+36h the spectra has maximum value at around total wave number 12 for VAR3 and 9 for T319. The top-right panel of Fig. 12 points out that although the difference between the two forecast errors comes mainly from scales with total wave number larger than 30 (~25%), there is already a detectable contribution coming from scales with total wave number smaller than 30. This explains why the difference in the spectra (top-left panel of Fig. 12) is evident for all scales, and explains the shift of the peak of the T319 spectrum towards the larger scales. As time progresses, the relative contribution from the small scales decreases: at t+72h (middle-right panel of Fig. 12), the error differences for the scales with total wave number between 10 and 30 is similar to the difference for the scales with total wave number bigger than 30, both at ~10%. After t+72h, due to the truncation of the VAR3 resolution to T255, the forecast error difference decreases even further. At t+108h, the relative difference between the two forecast errors (bottom-right panel) for the small scales is only ~ 5%, down from ~25% at t+36h. At this time, the forecast error of the VAR3 forecast becomes more similar to the error of the T319 forecasts for the small scales, due to the fact that now VAR3 has a lower resolution than T319. But for the large scales, the VAR3 forecast remains overall better, as shown by the fact that the forecast error difference peaks at about total wave number 10.

Figure 13 shows the corresponding spectra of the forecast error of the VAR3, the T319 and the T399 ensemble-mean forecasts in the realistic scenario: the left panels display the full spectra, and the right panels the relative difference between the spectra of the VAR3 and T319 control forecast errors. First of all, note that compared with the IME case, the forecast errors are larger for both configurations, and they exhibit most of the power in the large scales already at t+36h. Secondly, note that the relative difference between the two forecast errors is smaller than in the IME case. At t+36h, as it was detected from the IME results, the relative difference is larger for the small scales (top-right panel of Fig. 13), but it is only ~ 2% larger compared with ~25% in the

IME case (please note that, to make the differences more visible, the vertical axis of figs. 12 and 13 have been set to different values). Between t+36h and t+72h, as in the IME case the difference propagates to the larger scales but remains rather small (~1.5%). At t+108h (bottom panels of Fig. 13), VAR3 shows smaller errors for all scales, with differences slightly larger than at t+72h (~2%) for all scales with total wave number larger than 10.

5.2 Interpretation of the forecast error differences using a 3-parameter error growth model

A 3-parameter, forecast-error growth model has been used to investigate the gross features of the growth of the root-mean-square error of the 500 hPa geopotential height forecasts over the Northern Hemisphere. The model has been applied to the error of the T319 and the VAR3 ensemble-mean forecasts. The 3-parameter model is a modification of a 2-parameter model proposed by *Dalcher & Kalnay* (1987) and *Reynolds et al* (1994), and modified into a 3-parameter model with the inclusion of a linear term by *Simmons & Hollingsworth* (2001):

$$\frac{dE}{dt} = \gamma + \alpha E - \beta E^2$$

Simmons & Hollingsworth (2001) related the linear term γ to the growth of a component of the analysis error that is more rapid over the first day or two, and the exponential term to error growth due to initial condition errors. This parametric model can be used to estimate the forecast error doubling time $dblt(j)$ at the j -th forecast day:

$$dblt(j) = \frac{\ln 2}{\alpha + \gamma/E_j} .$$

Apart for the short forecast range, when the initial-time uncertainty is important and the linear term γ has a dominant effect and make the error growth super-exponential, doubling times can be used to understand the time evolution of the forecast error.

As in Lorenz (1982), Simmons *et al* (1995) and Simmons & Holligsworth (2001), the error model has been written in its finite-difference form

$$\frac{\Delta E}{\Delta t} = \gamma + \alpha \bar{E} - \beta \bar{E}^2$$

and the three parameters α , β and γ have been derived for both the VAR3 and the T319 configurations by a least-square fit of the differences of the root-mean-square errors

$\Delta E_j = E_{j+1} - E_j$ and $\bar{E}_j = 0.5(E_{j+1} + E_j)$, where E_j is the j -day forecast error.

Results indicate that the model is capable to describe the error growth of the T319 and VAR3 ensemble-mean forecast, as can be seen by the scatter plot (top panels of Figure 14) of the error increments versus the estimated increments both in the IME and the realistic cases (in both cases the correlation coefficients between the actual and the estimated forecast error increments is above 99%). The model has been used to estimate the initial-time uncertainty and the forecast error doubling times at each forecast day. Results indicate that:

- In the IME case (Fig. 14, left panels), the VAR3 and T319 doubling times are similar up to forecast day 3, but afterwards the VAR3 doubling times are smaller. At initial time, VAR3 forecasts start ‘closer’ to the verification (the T799 forecast) than the T319 forecasts, since they have a higher, T399 resolution. The T319 parametric curve has $\gamma_{T319}=0.95$ while the VAR3 curve has $\gamma_{VAR3}=0.84$, i.e. a 10% smaller value. Thus, the VAR3 forecast starts with a smaller initial error, which also grows less quickly during the

- super-exponential phase and similarly up to day 3 than the T319 forecast. The end result is that the VAR3 error remains much smaller than the T319 error up to the truncation time (day 3). After the truncation, the VAR3 doubling times are shorter, i.e. the forecast error grows faster, and this makes the difference between the VAR3 and T319 gradually smaller.
- In the realistic scenario, the parameter γ of the forecast error curve of all configurations is almost 3-times larger than the corresponding parameters computed in the IME scenario. This reflects the fact that the forecast error growth in the short forecast range is much faster in the realistic than in the IME scenario. The T319 parametric curve has $\gamma_{T319}=2.83$ (compared to 0.95 in IME) while the VAR3 curve has $\gamma_{VAR3}=2.75$ (compared to 0.84 in IME) for Z500. Thus, in the realistic scenario γ_{VAR3} is only 3% smaller than γ_{T319} , while it was 10% smaller in the IME scenario, which indicates that resolution has a smaller impact on the overall forecast error growth rate. The end result is that the VAR3 error is only slightly smaller than the T319 up to the truncation time. After the truncation time, the VAR3 and T319 forecast error doubling times are almost identical (Fig. 14, bottom-right panel).
 - To summarize, the fact that in the realistic scenario (a) γ_{VAR3} and γ_{T319} and (b) the VAR3 and T319 doubling times are closer than in the IME scenario explains why the differences between the VAR3 and the T319 forecasts errors are smaller. Furthermore, the fact that for each configuration the parameter γ is almost 3-times larger in the realistic than in the IME scenario explains why the forecast error is larger in the former scenario.

Simmons & Hollingsworth (2001) has shown that the model can be used to extrapolate the ensemble-mean forecast error evolution beyond 8-days for both configurations up to forecast day 40, when the curves reach their asymptotic limit

$$E_{\infty} = \frac{\alpha}{2\beta} + \sqrt{\frac{\alpha^2}{4\beta^2} + \frac{\gamma}{\beta}}.$$

Results indicate that the asymptotic limit is equal to 94 m² and 88 m², respectively, for the T319 and the VAR3 ensemble-mean forecasts in the IME scenario, and to 112 m² for both in the realistic scenario. In the IME scenario, during the first days (say up to forecast day 10) the VAR3 curve stays below the T319 curve, due to a combination of a smaller initial error and a slower forecast error growth in the short forecast range (reflected in a smaller γ). The fact that the VAR3 asymptotic value is lower than the T319 value is a result of the larger difference in model activity between T255 and T799, than between T319 and T799. In other words, the (VAR3) T255 model is less active, and this makes the asymptotic level smaller (on the impact of forecast activity on the asymptotic value, see also the discussion at the end of section 4 of *Simmons & Holligsworth* 2001). By contrast, in the realistic scenario the two curves asymptote to the same level: this indicates that the difference in activity between the two forecast resolutions is ‘masked’ by the effect of model imperfections.

6 Comparison of the gains in forecast skill of configurations T319, VAR3, VAR5, T399 and T799 with respect to the reference T255 system

The results discussed in sections 3, 4 and 5 have indicated that VAR3 is a better system than T319. To further document the quality of the VAR3 system, forecasts from the VAR3 and T319 configurations have been compared with forecasts generated using three constant resolutions, T255, T399 and T799. Furthermore, the sensitivity of the performance of the variable resolution configuration to the time when resolution is truncated from T399 to T255 has been assessed by comparing the scores of configurations VAR3 and VAR5 (see section 2 for its definition). The

performance of all these configurations has been summarized using the index I_2 , computed between forecast day 3 and 8. This index I_2 , which has been computed over the NH for all the three forecast variables analyzed in this study, measures the gains that each configuration can bring compared to the reference T255 configuration.

Figure 15 shows the predictability gains, measured by I_2 , computed in the IME scenario. Results are in line with what has been discussed in section 3-5. The top-panel of Fig. 15 indicates, for example, that for Z500 single control forecasts, using a T319 instead of a T255 configuration brings a ~13 hour gain. The gain is higher if measured in terms of T850 (~18 hours) and Z1000 (~14 hours). The other two panels of Fig. 15 show that the gains are lower if one considers single ensemble-mean forecasts, or probabilistic forecasts, instead of the control forecasts. Figure 15 explains that, compared with T319, the VAR3 gains are ~50-75% larger: thus, VAR3, which costs ~25% more than T319 (see Fig. 1), brings gains which are ~50-75% larger. Figure 15 also shows that the differences between the gains of VAR3 and VAR5 are very small, indicating that moving the truncation further into the forecast range does not bring any large improvement.

Figure 16 shows the corresponding predictability gains, measured by I_2 , computed in the realistic scenario. Results are in line with what has been shown earlier, and confirm that the differences between the ensemble configurations are much smaller in the realistic scenario. Compared with the IME case, gains are reduced by at least 50 % (from 10-25 hours down to 2-12 hours). Results still confirm that VAR3 outperforms T319 for all variables, with gains still between 50-75% larger. Note that Fig. 16 also shows the gains of a constant T799 configuration compared with T255. The comparison between the gains of the T399 and the T799 resolutions indicate that doubling the resolution increases the gains by between 25-50%.

7 Conclusions

The VARIable Resolution Ensemble Prediction System (VAREPS) was designed to resolve the smallest possible scales up to the forecast time when their inclusion had a positive impact on the prediction of both the small and the synoptic scales, and not to resolve them afterwards, when including them had a smaller, less detectable impact on the synoptic scales. *Buizza et al* (2007) compared results based on a preliminary version of VAREPS with forecasts from a constant resolution system that required a similar amount of computer resources, and concluded that VAREPS provided better forecasts in the early forecast range without losing accuracy in the long forecast range. Although they found a clear advantage of VAREPS in the early forecast range up to the truncation time, they detected only some limited evidence that VAREPS was providing better upper-level forecasts than a constant resolution T_L319 system beyond the truncation time.

This work investigated in more details and using different diagnostic approaches whether a variable resolution approach to numerical weather prediction would bring better forecasts *beyond the truncation time* than an equivalent, constant resolution system. To address this question, ensembles with 5 members have been run with an 8-day forecast length in 7 different configurations for a whole season (winter 2007-08). In particular, a T319 constant resolution configuration has been compared with a VAR3 configuration with a resolution T399 up to forecast day 3 and T255 from day 3 to day 8. The performance of these ensembles have been assessed both in an idealized model error (IME) scenario, with forecasts verified against T799 control forecasts, and in a realistic scenario, with forecasts verified against ECMWF T799 analyses.

Results have indicated that VAR3 forecasts are more accurate than T319 forecasts for the whole forecast range. VAR3 forecasts are definitely more accurate in the IME case, when forecasts are verified against T799 control forecasts, but less so when forecasts are verified against ECMWF T799 analyses. In the IME case, the only source of forecast error is due to the fact that forecasts use a resolution lower than the one used in the verification (T319 or T399-T255 in VAR3, compared with T799). In the short forecast range up to the truncation time, error propagates upscale more rapidly in the T319 forecast than in the VAR3 (T399) forecast, thus making the T319 error larger for all waves. At day 3, when the VAR3 resolution is reduced from T399 to T255, the error difference is already large, and even if the VAR3 (T255) forecast is performed with a lower resolution than the T319 one, its error remains lower. The rank-sum, non-parametric Mann-Whitney-Wilcoxon test indicates that differences are statistically significant to the 5% level for up to forecast day 8.

When forecasts are verified against analyses in the realistic scenario, differences in the performance between the T319 and the VAR3 configurations are masked by ‘model imperfections’, i.e. by the effect of model errors not represented by the IME assumption. As a result, the difference induced by using a T319 or a T399 resolution during the first 3 days is much smaller. The contribution to the forecast error due to model resolution is not any more dominant, and thus it is not surprising that in this scenario the T319 and the VAR3 configurations perform more similarly. In this case, the rank-sum, non-parametric Mann-Whitney-Wilcoxon test has indicated that differences are not statistically significant to the 5% level. These results obtained in the realistic scenario are consistent with the findings of *Buizza et al (2007)*, who did not detect any statistically significant difference between the performance of a variable resolution and a constant resolution system.

The analysis of the time evolution of the forecast error spectra and the use of a 3-parameter forecast error growth model has helped understanding these results. For example, the comparison of the 3-parameter curves of VAR3 and T319 have indicated that the key difference in the forecast error growth can be detected in γ parameter, which describes the super-exponential forecast error growth in the short forecast error. The fact that, for each forecast configuration, γ is almost 3-times larger in the realistic scenario indicates that the effect of ‘model imperfection’ dominates over the effect of using a 319 or a 399 forecast resolution.

Considering the key question posed in the introduction, these results should provide enough evidence of the fact that *of two comparable-cost configurations, one with a constant resolution, and one with a variable resolution, higher in the earlier forecast range and lower afterwards, the latter gives the best medium-range forecasts even beyond the truncation time, but differences might not be detectable, or might have a low statistical significance, when forecasts are verified against analyses, because the benefit of using this configuration is masked by the effect of ‘model imperfections’*. The comparison of the results obtained in the IME and the realistic scenario suggests that future model error reductions might lead to more evident differences.

Acknowledgements

The development, operational implementation, and continuous improvement of the ECMWF Ensemble Prediction System would not have been possible without the contribution of many staff members and consultants: their work is acknowledged. Tim Palmer is acknowledged for having provided valuable comments and suggestions to an earlier version of this paper.

References

- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.
- Buizza, R., & T.N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**, 1434-1456.
- Buizza, R., M. Miller, & T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **125**, 2887-2908.
- Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., & Vitart, F., 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). *Q. J. Roy. Meteorol. Soc.*, **133**, 681-695.
- Dalcher, A., & Kalnay, E., 1987: Error growth and predictability in operational ECMWF forecasts. *Tellus*, **39**, 474-491.
- Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505-513.
- Palmer, T N, Buizza, R., Leutbecher, M., Hagedorn, R., Jung, T., Rodwell, M, Virat, F., Berner, J., Hagel, E., Lawrence, A., Pappenberger, F., Park, Y.-Y., van Bremen, L., Gilmour, I., & Smith, L., 2007: The ECMWF Ensemble Prediction System: recent and on-going developments. A paper presented at the 36th Session of the ECMWF Scientific Advisory Committee. *ECMWF Research Department Technical Memorandum n. 540*, pp 55 (available from ECMWF, Shinfield Park, Reading RG2-9AX, UK, or from <http://www.ecmwf.int/publications/library/>).
- Reynolds, C. A., Webster, P. J., & Kalnay, E., 1994: Random error growth in NMC's global forecasts. *Mon. Wea. Rev.*, **122**, 1281-1305.

Simmons, A. J., & Hollingsworth, A., 2001: Some aspects of the improvement in skill of numerical weather prediction. ECMWF Research Department Technical Memorandum No. 342, pp. 33 (available from ECMWF, Shinfield Park, Reading, RG29AX, UK, or from <http://www.ecmwf.int/publications/library/>).

Simmons, A. J., Mureau, R., & Petroliaigis, T., 1995: Error growth and predictability estimates for the ECMWF forecasting system. *Q. J. R. Meteorol. Soc.*, **121**, 1739-1771.

Swets, J. A., 1986: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychol. Bull.*, **99**, 181-198.

Szunyogh, I., & Toth, Z., 2002: The effect of increased horizontal resolution on the NCEP Global Ensemble Mean Forecasts. *Mon. Wea. Rev.*, **130**, 1115-1143.

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

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

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

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences*. Academic Press, Inc., San Diego, pp. 467 (ISBN 0-12-751965-3).

Figures' and tables' captions

Figure 1. CPU time required to complete one 8-day integration, expressed in terms of the CPU time required by the T319 configuration.

Figure 2. IME results for the 500 hPa geopotential height over NH verified against T799 control forecasts. (a) rmse of the control (CON) forecast and (b) rmse of the ensemble-mean (EM) forecast, for the ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed). (c) As (a) but for the relative difference (in percentage) of the rmse of the control forecast $[rmse(CON) - rmse(CON_{T399})]/rmse(CON_{T399})$. (d) As (b) but for the relative difference of the rmse of the EM forecast $[rmse(EM) - rmse(EM_{T399})]/rmse(EM_{T399})$.

Figure 3. As Fig. 2 but for the SH.

Figure 4. IME results for the probabilistic forecast of the 500 hPa geopotential height verified against T799 control forecasts. (a) rank-probability-skill-score (RPSS) computed over the NH for ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed). (b): as (a) but for the SH. (c) As (a) but for the relative difference (in percentage) $[RPSS - RPSS_{T399}]/RPSS_{T399}$.computed over NH. (d) As (c) but for the SH.

Figure 5. Realistic results for the 500 hPa geopotential height over NH verified against ECMWF T799 analyses. (a) rmse of the control (CON) forecast and (b) rmse of the ensemble-mean (EM) forecast, for the ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399

(chain-dashed). (c) As (a) but for the relative difference of the rmse of the control forecast $[rmse(CON) - rmse(CON_{T399})]/rmse(CON_{T399})$. (d) As (b) but for the relative difference of the rmse of the EM forecast $[rmse(EM) - rmse(EM_{T399})]/rmse(EM_{T399})$.

Figure 6. As Fig. 5 but for the SH.

Figure 7. Realistic results for the probabilistic forecast of the 500 hPa geopotential height verified against ECMWF T799 analyses. (a) RPSS computed over NH for ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed). (b): as (a) but for the SH. (c) As (a) but for the relative difference (in percentage) $[RPSS - RPSS_{T399}]/RPSS_{T399}$ computed over NH. (d) As (c) but for the SH.

Figure 8. IME results for the 500 hPa geopotential height verified against T799 control forecasts. (a) percentage difference $[rmse(CON_{VAR3}) - rmse(CON_{T319})]/rmse(CON_{T319})$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for $[rmse(EM_{VAR3}) - rmse(EM_{T319})]/rmse(EM_{T319})$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

Figure 9. IME results for the probabilistic prediction of the 500 hPa geopotential height verified against T799 control forecasts. (a) percentage difference $[RPSS_{VAR3} - RPSS_{T319}]/RPSS_{T319}$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted)

computed over NH. (b): as (a) but for $[ROCA_{VAR3} - ROCA_{T319}] / ROCA_{T319}$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

Figure 10. Realistic results for the 500 hPa geopotential height verified against ECMWF T799 analyses. (a) percentage difference $[rmse(CON_{VAR3}) - rmse(CON_{T319})] / rmse(CON_{T319})$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for $[rmse(EM_{VAR3}) - rmse(EM_{T319})] / rmse(EM_{T319})$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

Figure 11. Realistic results for the probabilistic prediction of the 500 hPa geopotential height verified against ECMWF T799 analyses. (a) percentage difference $[RPSS_{VAR3} - RPSS_{T319}] / RPSS_{T319}$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for $[ROCA_{VAR3} - ROCA_{T319}] / ROCA_{T319}$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

Figure 12. IME results for the 500 hPa geopotential height, verified against T799 control forecasts over NH. Left panels: 30-day average (1-31 January 2008) total wave-number spectra (between wave number 1 and 63) of the error of the ensemble-mean forecast of configurations T319 (dashed) and VAR3 (solid) at forecast step $t+36h$ (top, in m^2 , multiplied by $0.4 \cdot 10^{-1}$), $T+72h$ (middle panel, in m^2 , multiplied by $0.8 \cdot 10^{-2}$) and $t+108h$ (bottom panel, in m^2 , multiplied by $0.2 \cdot 10^{-2}$). Right panels (in percentages): as left panels but for the relative difference (in percentage) between the spectra $diff(n) = (sp_{VAR3}(n) - sp_{T319}(n)) / sp_{T319}(n)$.

Figure 13. Realistic results for the 500 hPa geopotential height, verified against ECMWF T799 analyses over NH. Left panels: 30-day average (1-31 January 2008) total wave-number spectra (between wave number 1 and 63) of the error of the ensemble-mean forecast of configurations T319 (solid) and VAR3 (dashed). at forecast step $t+36h$ (top, in m^2 , multiplied by $0.5 \cdot 10^{-3}$), $T+72h$ (middle panel, in m^2 , multiplied by $0.5 \cdot 10^{-3}$) and $t+108h$ (bottom panel, in m^2 , multiplied by $0.25 \cdot 10^{-3}$). Right panels (in percentages): as left panels but for the relative difference (in percentage) between the spectra $diff(n) = (sp_{VAR3}(n) - sp_{T319}(n)) / sp_{T319}(n)$.

Figure 14. 3-parameter forecast error model applied to the 500 hPa geopotential height forecast errors computed over NH. Top panels: scatter plot of the actual versus estimated ensemble-mean rmse increments computed in the IME (left panel) and the realistic (right panel) scenario for T319 (black full circles) and VAR3 (grey crosses). Bottom panels: estimated ensemble-mean forecast error doubling times computed in the IME (left panel) and the realistic (right panel) scenario for T319 (solid) and VAR3 (dash).

Figure 15. IME results. Predictability gains, measured using the 'gain' index I_2 , for configurations T319 (black), VAR3 (dark grey), VAR5 (white bars) and T399 (light grey) computed for the 500 and the 1000 hPa geopotential height (Z500, Z1000) and the 850 hPa temperature (T850) control forecasts (top panel), ensemble-mean forecasts (middle panel) and probabilistic forecasts (bottom panel). Single forecasts' accuracy has been measured using rmse over NH, and probabilistic forecast accuracy using RPSS over NH.

Figure 16. Realistic results. Predictability gains, measured using the 'gain' index I_2 , for configurations T319 (black), VAR3 (dark grey), VAR5 (white), T399 (light grey) and T799 (striped) computed for the 500 and the 1000 hPa geopotential height (Z500, Z1000) and the 850 hPa temperature (T850) control forecasts (top panel), ensemble-mean forecasts (middle panel) and probabilistic forecasts (bottom panel). Single forecasts' accuracy has been measured using rmse over NH, and probabilistic forecast accuracy using RPSS over NH.

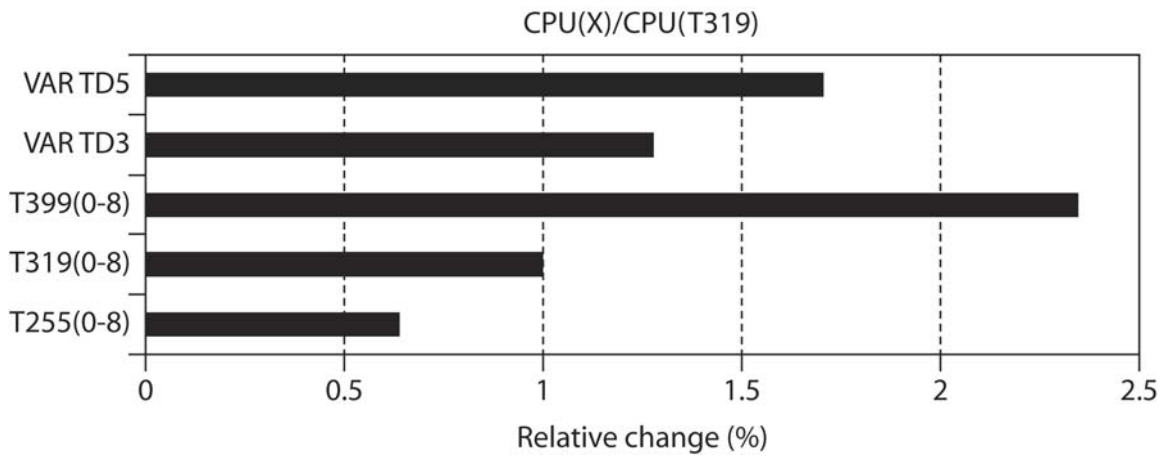


Figure 1. CPU time required to complete one 8-day integration, expressed in terms of the CPU time required by the T319 configuration.

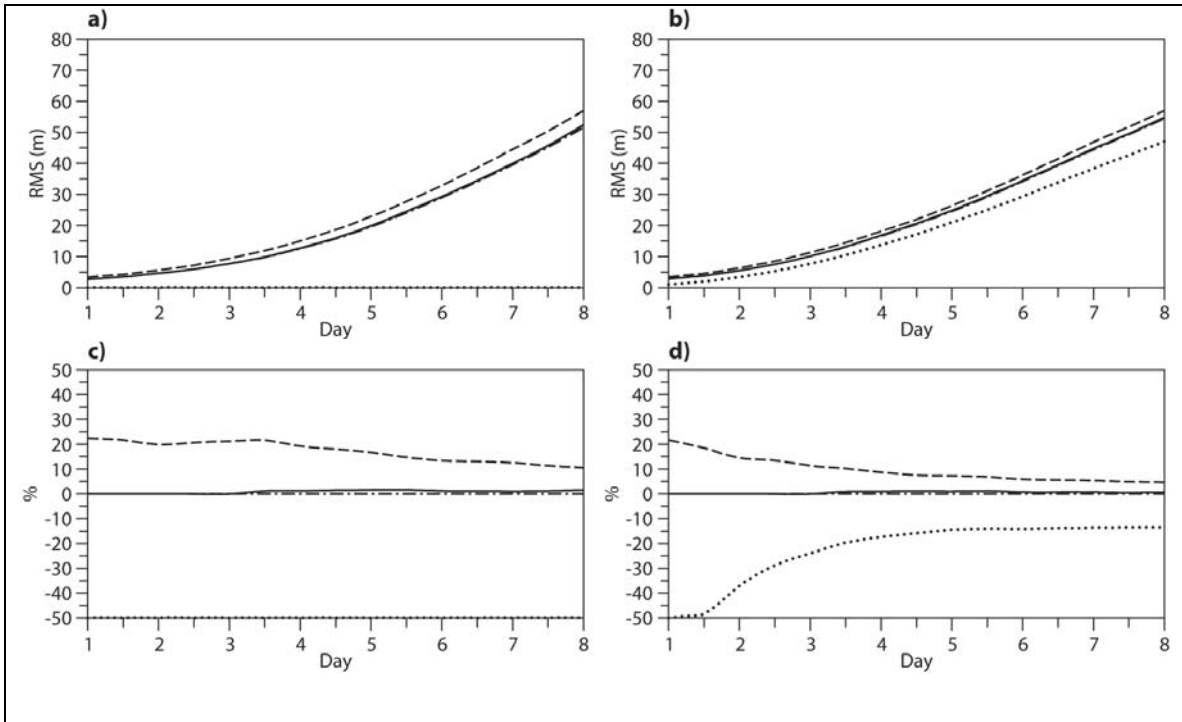


Figure 2. IME results for the 500 hPa geopotential height over NH verified against T799 control forecasts. (a) rmse of the control (CON) forecast and (b) rmse of the ensemble-mean (EM) forecast, for the ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed). (c) As (a) but for the relative difference (in percentage) of the rmse of the control forecast $[rmse(CON) - rmse(CON_{T399})]/rmse(CON_{T399})$. (d) As (b) but for the relative difference of the rmse of the EM forecast $[rmse(EM) - rmse(EM_{T399})]/rmse(EM_{T399})$.

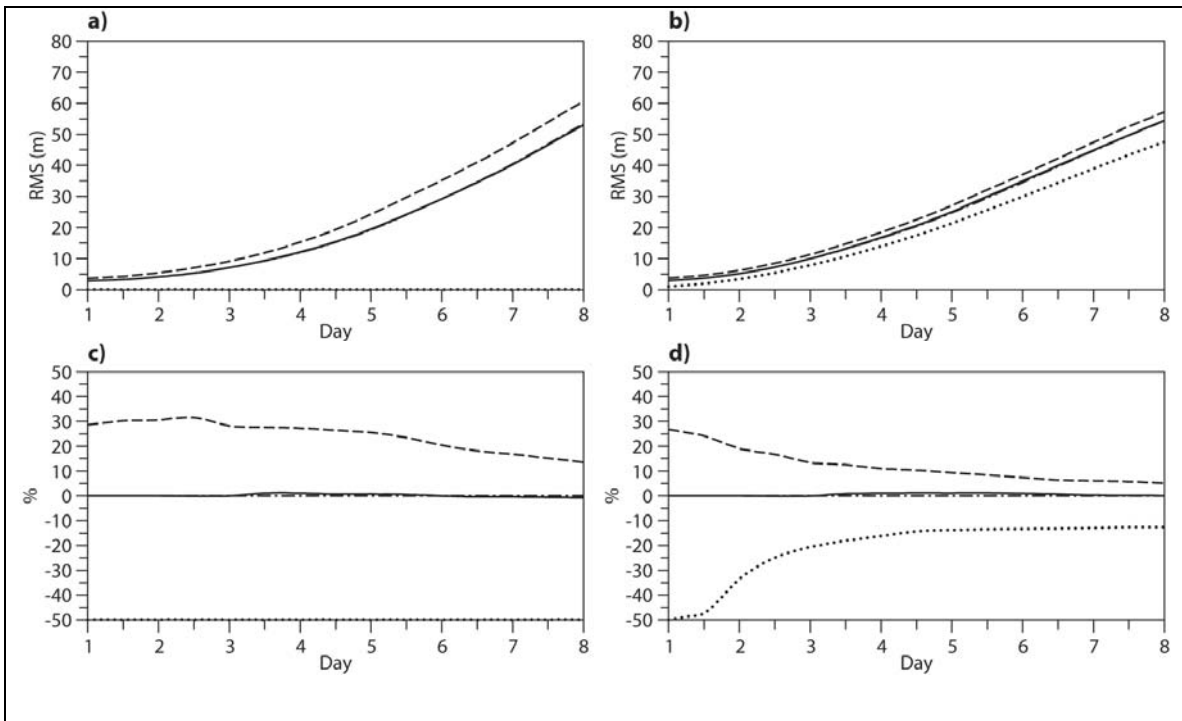


Figure 3. As Fig. 2 but for the SH.

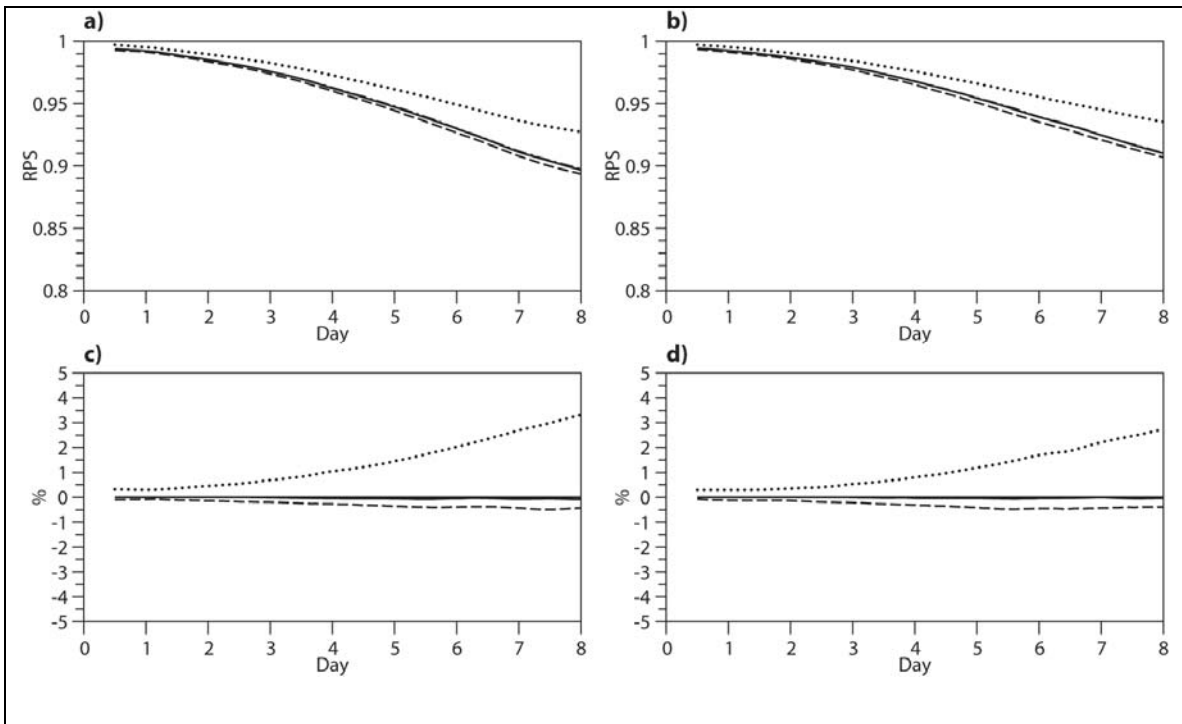


Figure 4. IME results for the probabilistic forecast of the 500 hPa geopotential height verified against T799 control forecasts. (a) rank-probability-skill-score (RPSS) computed over the NH for ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed).

(b): as (a) but for the SH. (c) As (a) but for the relative difference (in percentage)

$\left[RPSS - RPSS_{T399}\right] / RPSS_{T399}$.computed over NH. (d) As (c) but for the SH.

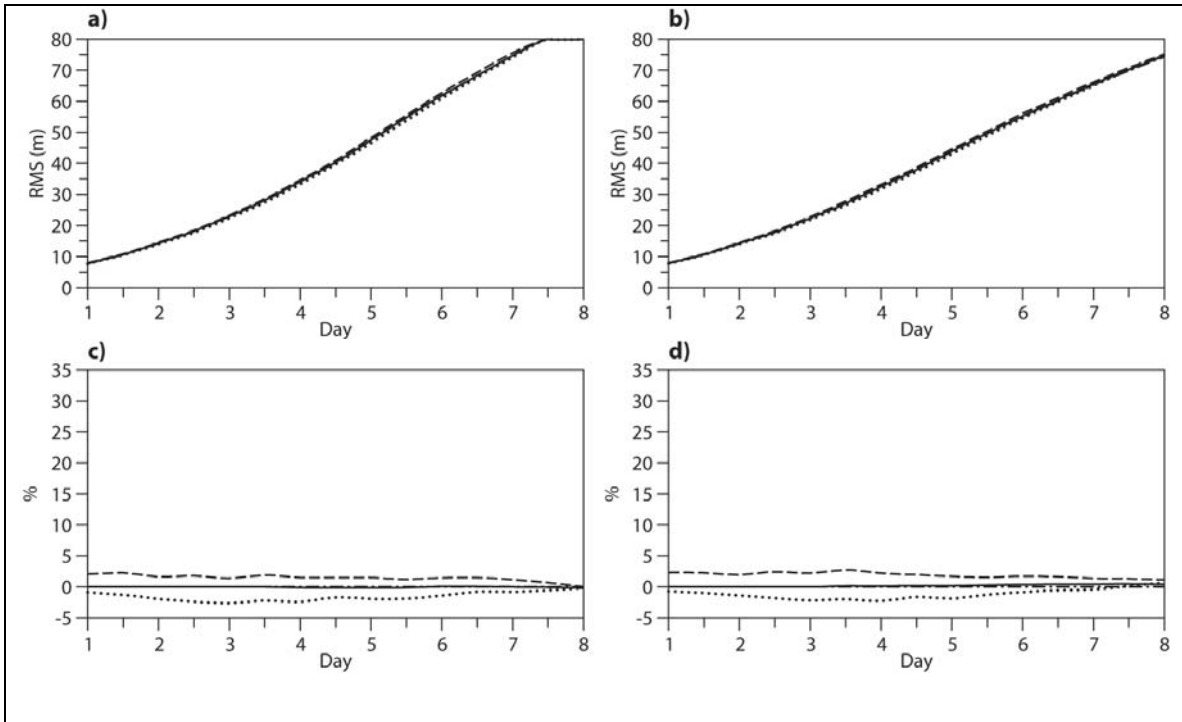


Figure 5. Realistic results for the 500 hPa geopotential height over NH verified against ECMWF T799 analyses. (a) rmse of the control (CON) forecast and (b) rmse of the ensemble-mean (EM) forecast, for the ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed). (c) As (a) but for the relative difference of the rmse of the control forecast $[rmse(CON) - rmse(CON_{T399})]/rmse(CON_{T399})$. (d) As (b) but for the relative difference of the rmse of the EM forecast $[rmse(EM) - rmse(EM_{T399})]/rmse(EM_{T399})$.

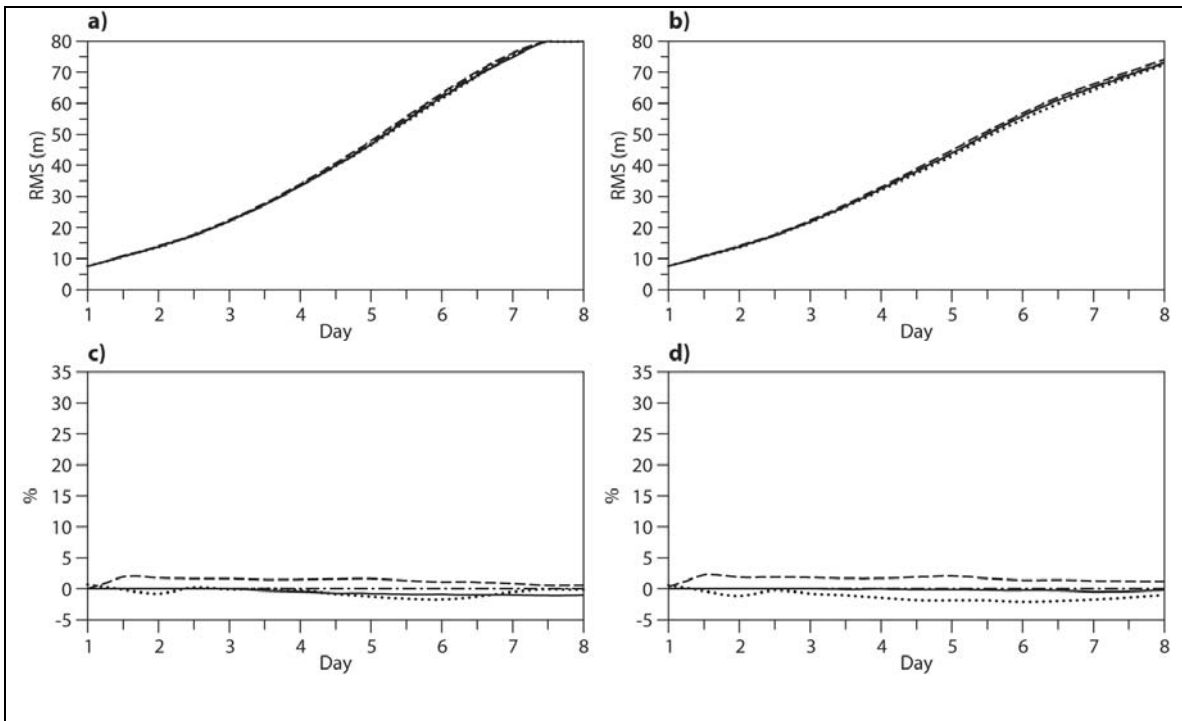


Figure 6. As Fig. 5 but for the SH.

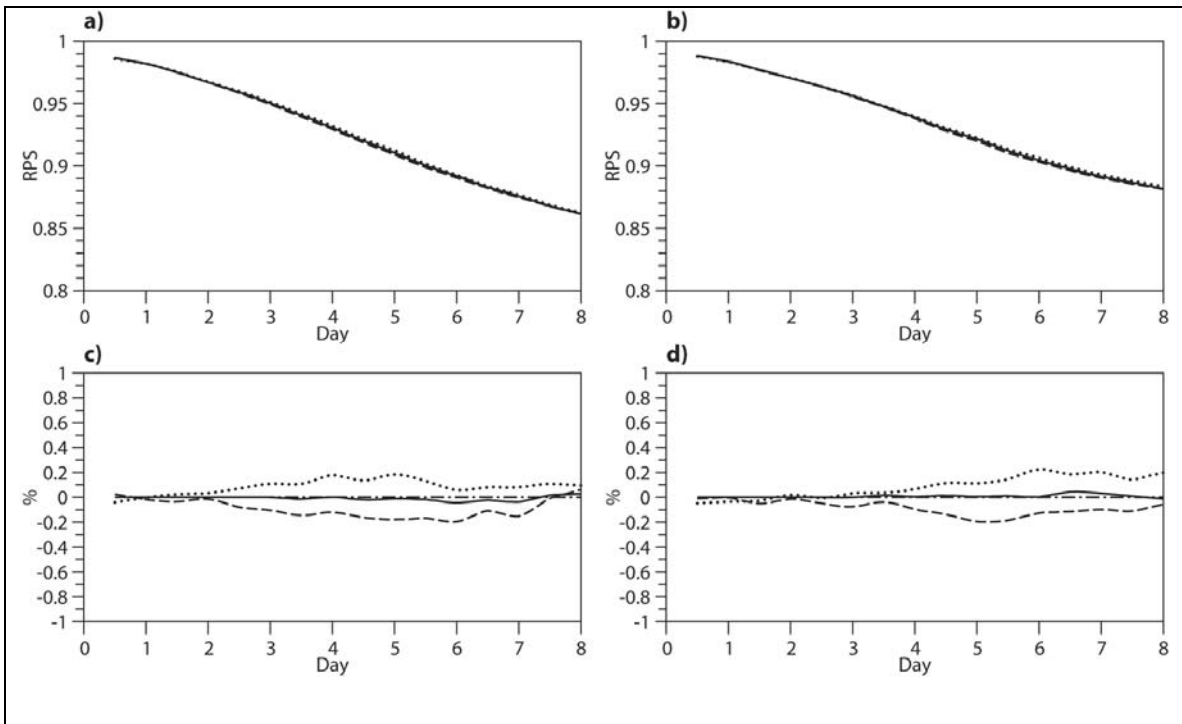


Figure 7. Realistic results for the probabilistic forecast of the 500 hPa geopotential height verified against ECMWF T799 analyses. (a) RPSS computed over NH for ensemble configurations VAR3 (solid), T319 (dashed), T799 (dotted) and T399 (chain-dashed). (b): as (a) but for the SH. (c) As (a) but for the relative difference (in percentage) $[RPSS - RPSS_{T399}] / RPSS_{T399}$ computed over NH. (d) As (c) but for the SH.

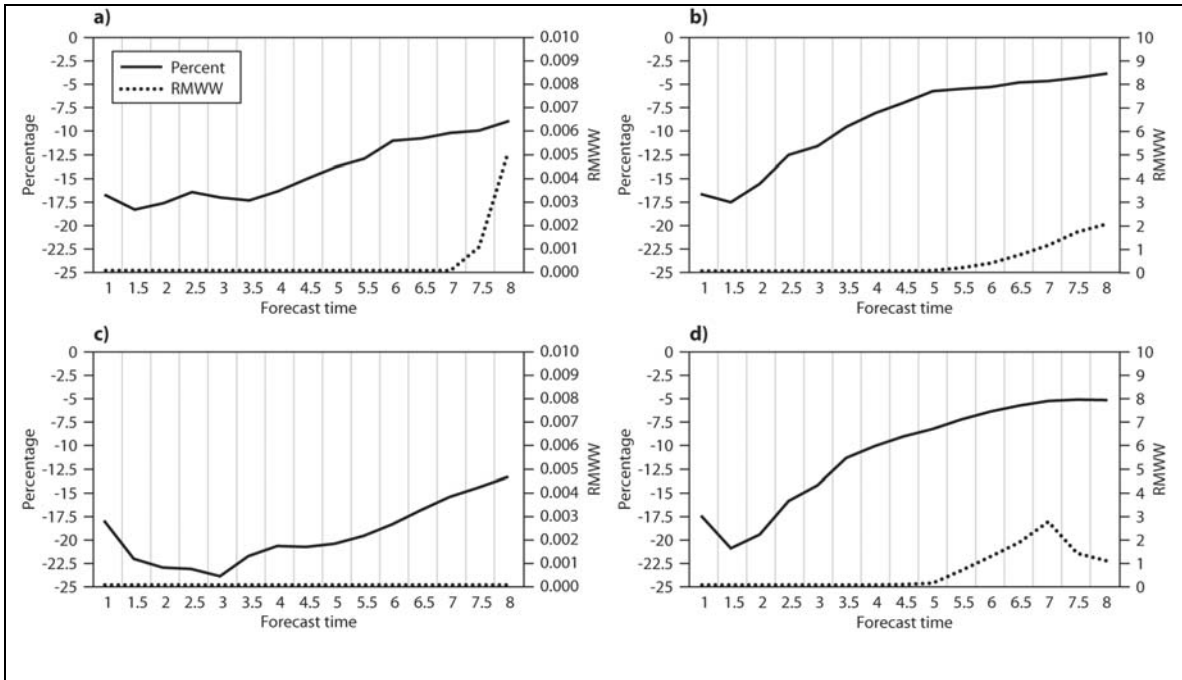


Figure 8. IME results for the 500 hPa geopotential height verified against T799 control forecasts.

(a) percentage difference $[rmse(CON_{VAR3}) - rmse(CON_{T319})] / rmse(CON_{T319})$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for $[rmse(EM_{VAR3}) - rmse(EM_{T319})] / rmse(EM_{T319})$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

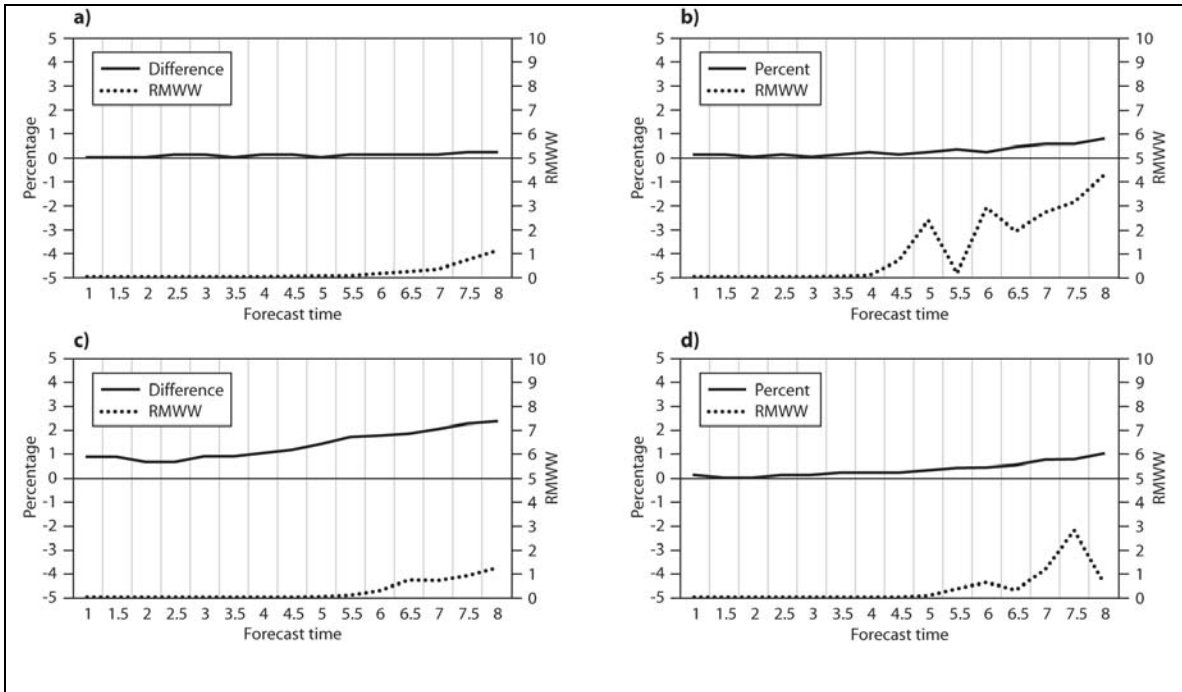


Figure 9. IME results for the probabilistic prediction of the 500 hPa geopotential height verified against T799 control forecasts. (a) percentage difference $[RPSS_{VAR3} - RPSS_{T319}] / RPSS_{T319}$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for $[ROCA_{VAR3} - ROCA_{T319}] / ROCA_{T319}$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

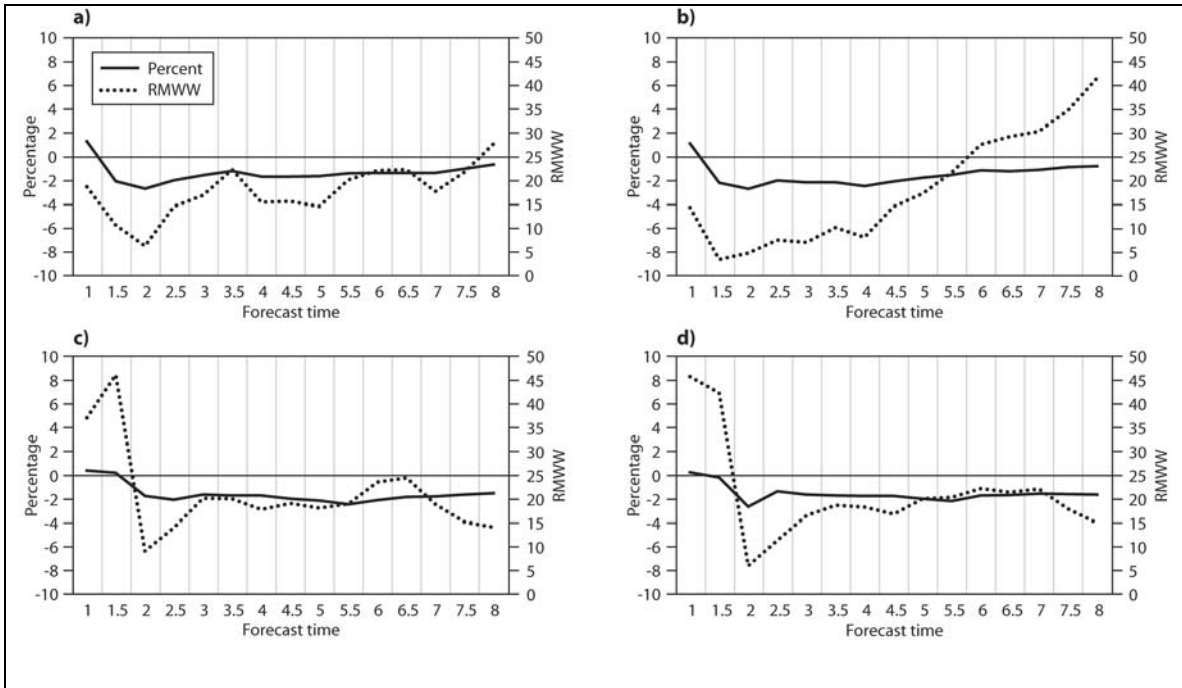


Figure 10. Realistic results for the 500 hPa geopotential height verified against ECMWF T799 analyses. (a) percentage difference $[rmse(CON_{VAR3}) - rmse(CON_{T319})] / rmse(CON_{T319})$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon (RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for $[rmse(EM_{VAR3}) - rmse(EM_{T319})] / rmse(EM_{T319})$ and corresponding RMWW test over NH. (c-d): as (a-b) but over SH.

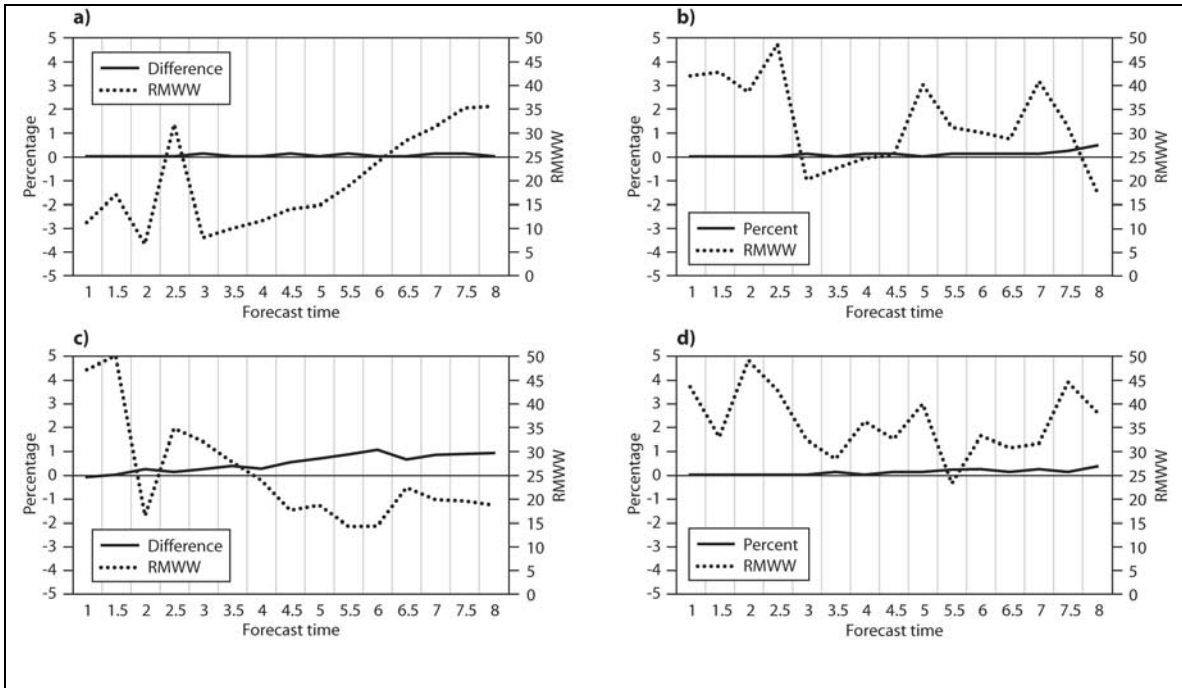


Figure 11. Realistic results for the probabilistic prediction of the 500 hPa geopotential height

verified against ECMWF T799 analyses. (a) percentage difference

$[RPSS_{VAR3} - RPSS_{T319}] / RPSS_{T319}$ (solid) and corresponding Rank Mann-Whitney-Wilcoxon

(RMWW) statistical test value (dotted) computed over NH. (b): as (a) but for

$[ROCA_{VAR3} - ROCA_{T319}] / ROCA_{T319}$ and corresponding RMWW test over NH. (c-d): as (a-b)

but over SH.

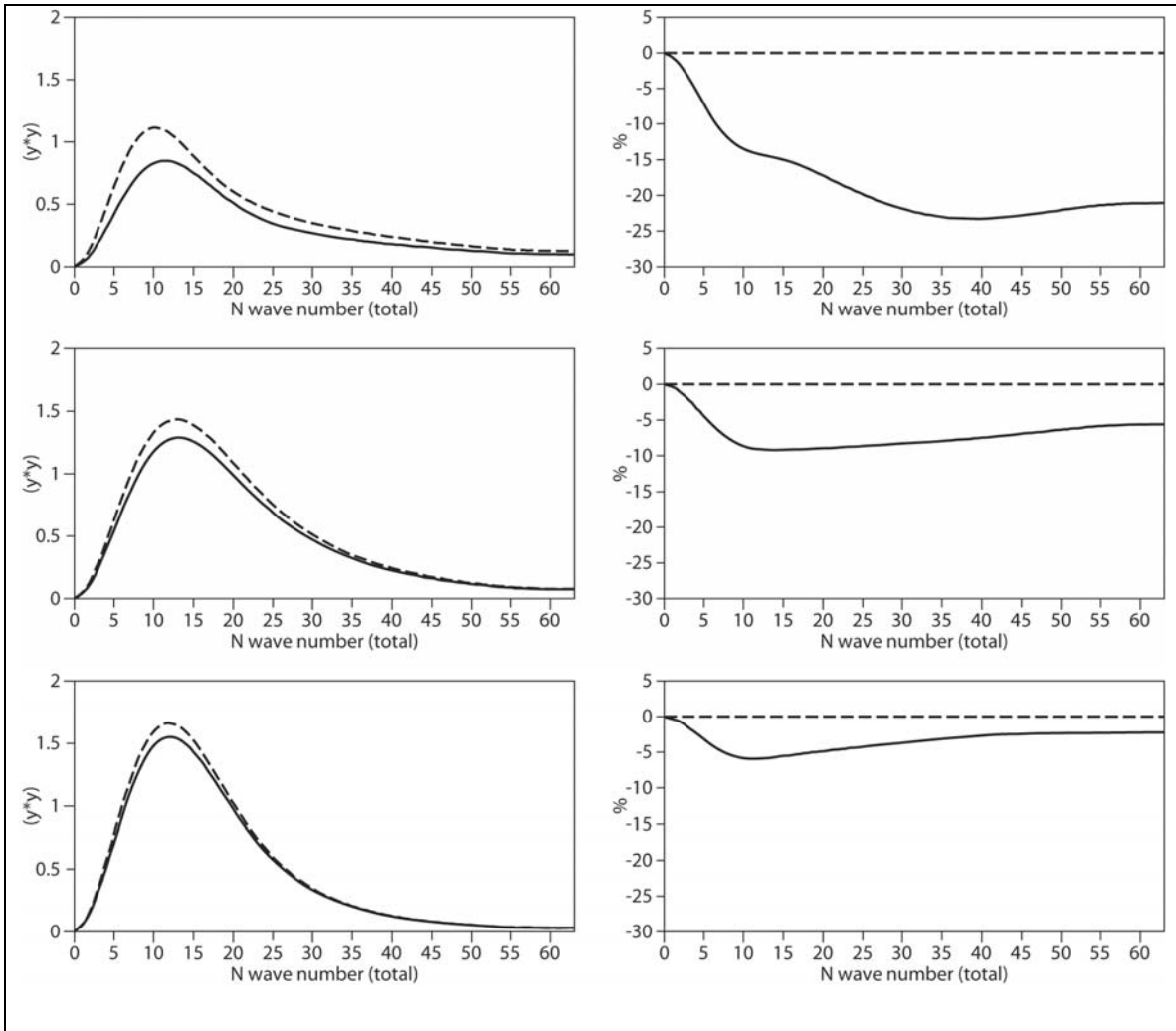


Figure 12. IME results for the 500 hPa geopotential height, verified against T799 control

forecasts over NH. Left panels: 30-day average (1-31 January 2008) total wave-number spectra (between wave number 1 and 63) of the error of the ensemble-mean forecast of configurations T319 (dashed) and VAR3 (solid) at forecast step $t+36h$ (top, in m^2 , multiplied by $0.4 \cdot 10^{-1}$), $T+72h$ (middle panel, in m^2 , multiplied by $0.8 \cdot 10^{-2}$) and $t+108h$ (bottom panel, in m^2 , multiplied by $0.2 \cdot 10^{-2}$). Right panels (in percentages): as left panels but for the relative difference (in percentage) between the spectra $diff(n) = (sp_{VAR3}(n) - sp_{T319}(n)) / sp_{T319}(n)$.

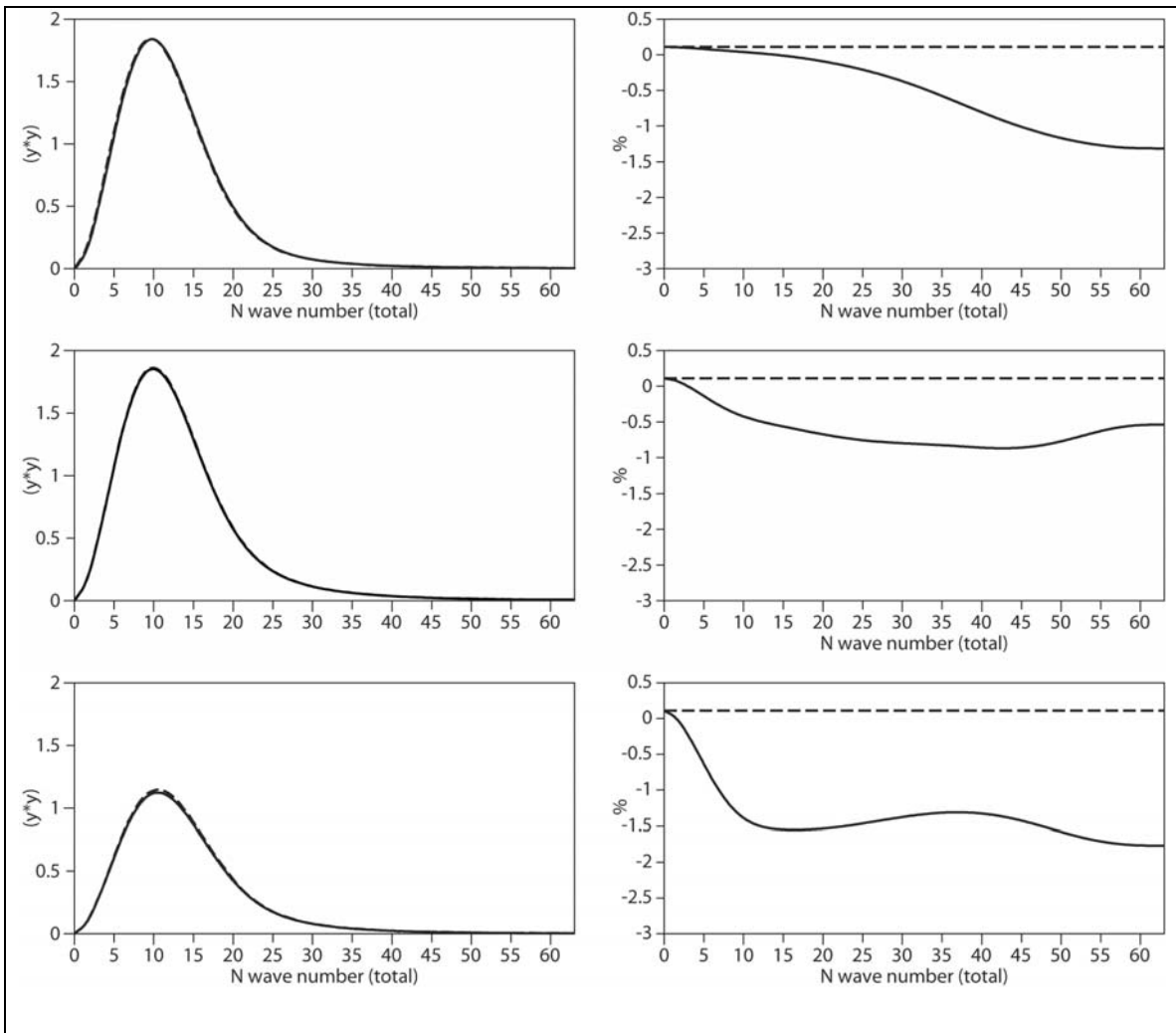


Figure 13. Realistic results for the 500 hPa geopotential height, verified against ECMWF T799 analyses over NH. Left panels: 30-day average (1-31 January 2008) total wave-number spectra (between wave number 1 and 63) of the error of the ensemble-mean forecast of configurations T319 (solid) and VAR3 (dashed) at forecast step $t+36h$ (top, in m^2 , multiplied by $0.5 \cdot 10^{-3}$), $T+72h$ (middle panel, in m^2 , multiplied by $0.5 \cdot 10^{-3}$) and $t+108h$ (bottom panel, in m^2 , multiplied by $0.25 \cdot 10^{-3}$). Right panels (in percentages): as left panels but for the relative difference (in percentage) between the spectra $diff(n) = (sp_{VAR3}(n) - sp_{T319}(n)) / sp_{T319}(n)$.

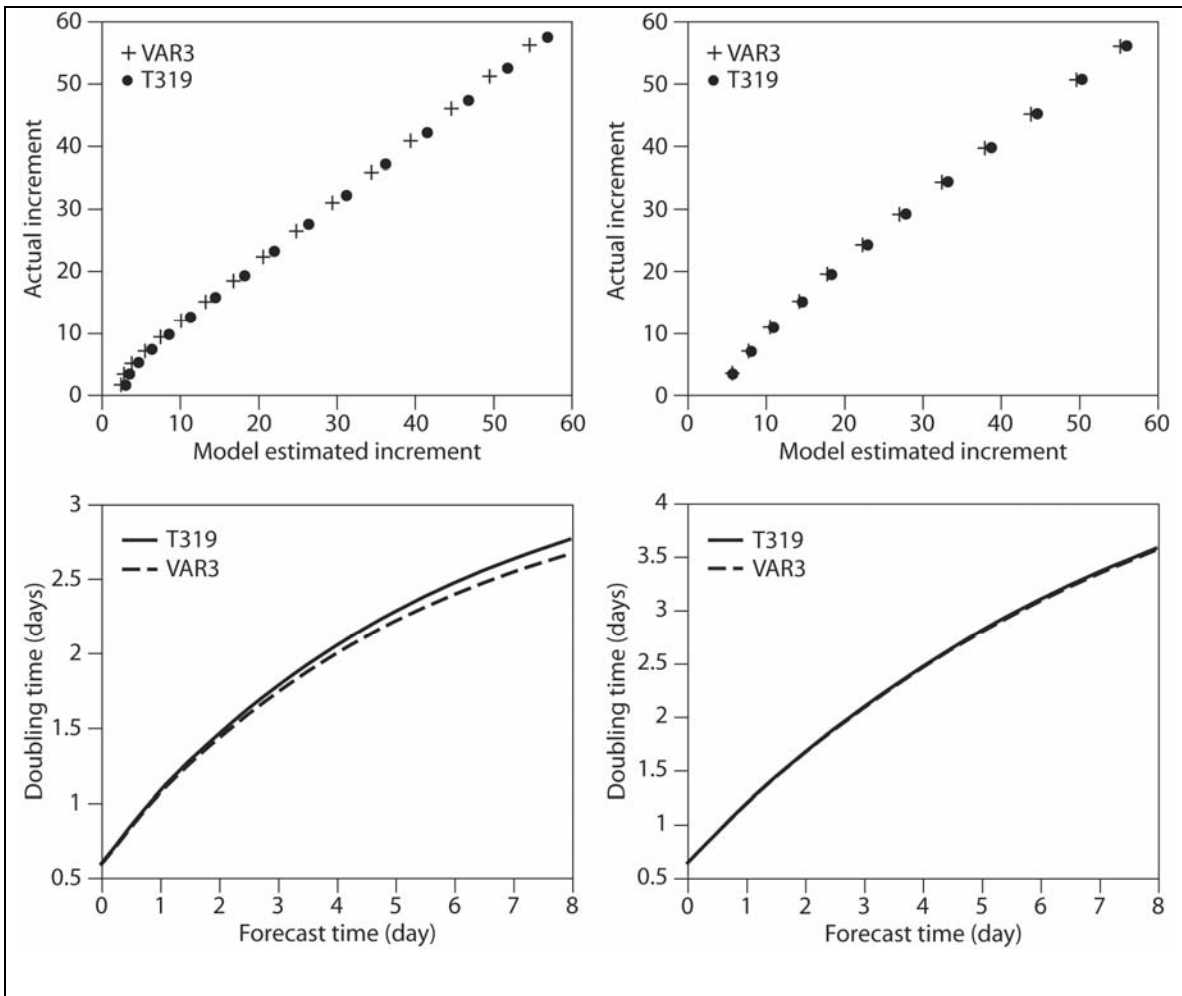


Figure 14.3-parameter forecast error model applied to the 500 hPa geopotential height forecast errors computed over NH. Top panels: scatter plot of the actual versus estimated ensemble-mean rmse increments computed in the IME (left panel) and the realistic (right panel) scenario for T319 (black full circles) and VAR3 (grey crosses). Bottom panels: estimated ensemble-mean forecast error doubling times computed in the IME (left panel) and the realistic (right panel) scenario for T319 (solid) and VAR3 (dash).

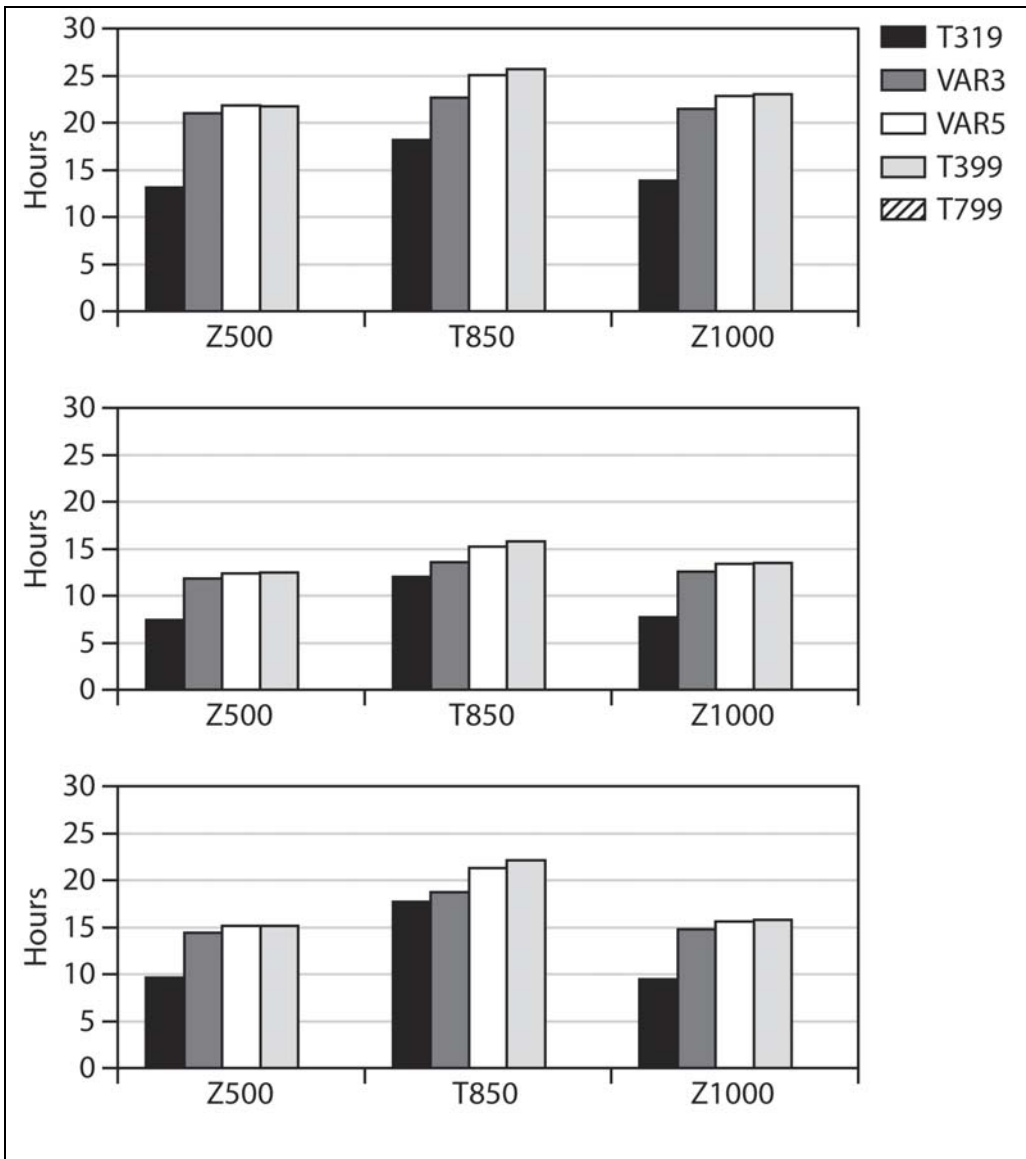


Figure 15. IME results. Predictability gains, measured using the ‘gain’ index I_2 , for configurations T319 (black), VAR3 (dark grey), VAR5 (white bars) and T399 (light grey) computed for the 500 and the 1000 hPa geopotential height (Z500, Z1000) and the 850 hPa temperature (T850) control forecasts (top panel), ensemble-mean forecasts (middle panel) and probabilistic forecasts (bottom panel). Single forecasts’ accuracy has been measured using rmse over NH, and probabilistic forecast accuracy using RPSS over NH.

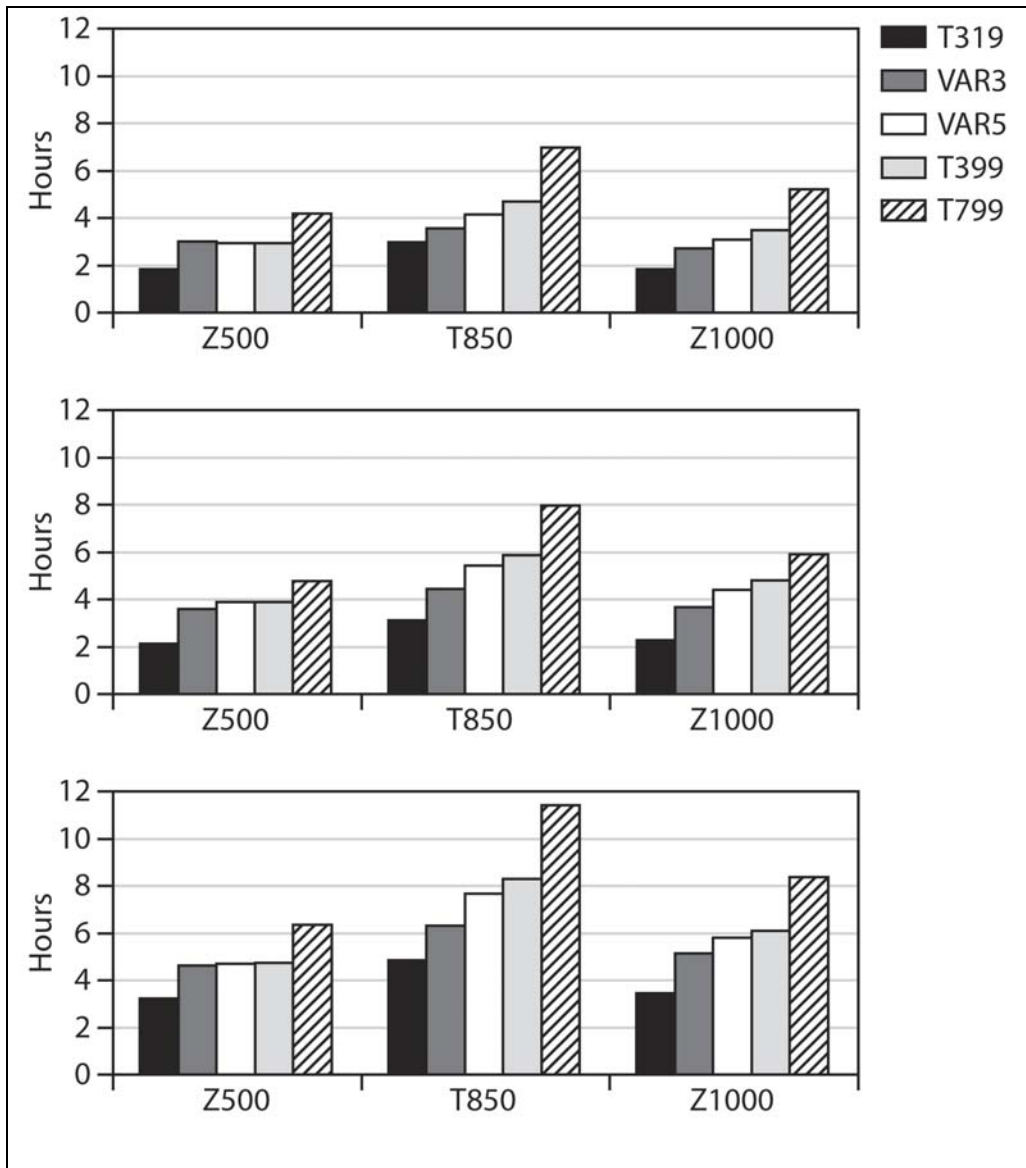


Figure 16. Realistic results. Predictability gains, measured using the ‘gain’ index I_2 , for configurations T319 (black), VAR3 (dark grey), VAR5 (white), T399 (light grey) and T799 (striped) computed for the 500 and the 1000 hPa geopotential height (Z500, Z1000) and the 850 hPa temperature (T850) control forecasts (top panel), ensemble-mean forecasts (middle panel) and probabilistic forecasts (bottom panel). Single forecasts’ accuracy has been measured using rmse over NH, and probabilistic forecast accuracy using RPSS over NH.