WORLD METEOROLOGICAL ORGANIZATION

COMMISSION FOR BASIC SYSTEMS
OPAG on DPFS

MEETING OF THE CBS (DPFS) TASK TEAM ON SURFACE VERIFICATION

GENEVA, SWITZERLAND
20-21 OCTOBER 2014

DPFS/TT-SV/Doc. 4.1c

(15.10.2014)

Agenda item : 4.1

ENGLISH ONLY

REPORT ON A SENSITIVITY STUDY IN SURFACE VERIFICATION AT ECMWF

(Submitted by Thomas Haiden)

Summary and purpose of document

This document provides background information on the results of the sensitivity study performed at ECMWF on the effect on scores of some choices in surface verification.

Action Proposed

The meeting is invited to take note of the results presented and to consider them where appropriate in the discussion of open items of the draft document on standard procedures.

Summary. The sensitivity of error scores to some choices in surface verification against SYNOP observations is investigated. It is tested how a simple quality control procedure for 2-m temperature, which is based on temporal consistency, affects the results. Use of the nearest land grid-point instead of the nearest grid-point in verification of 2-m temperature and 10-m wind speed is studied, as well as the application of a height-correction for 2-m temperature. It is found that the height correction has the strongest, and overall beneficial, effect. Use of the nearest land grid-point mostly affects the mean errors (biases), not necessarily reducing them, while the effect on mean absolute and root mean square errors is just a few %. The smallest (actually negligible) effect is found for the quality control of 2-m temperature. However, a more sophisticated quality control, which possibly detects more of the erroneous observations, may have a somewhat larger effect.

1. Introduction

The work of the World Meteorological Organization (WMO) Commission for Basic Systems (CBS) is accomplished through a Management Group and four Open Programme Area Groups (OPAGs). As part of the OPAG on the Data-Processing and Forecasting System (DPFS), the Expert Team on the Operational Weather Forecasting Process and Support (ET-OFPS) has set up a task team (TT) on standard procedures for surface verification. The main objective of the TT is to come up with recommendations for surface verification, similar to what exists for upper-air verification (WMO, 2012).

During the work of the TT, a number of open questions were identified, some of which are addressed in this study. Specifically, the issues of quality control (QC), the question of whether to use the nearest grid-point or the nearest land grid-point, and whether to use a height correction for 2-m temperature verification, were noted as requiring further study.

2. Effect of quality control on 2-m temperature verification

The SYNOP data distributed via the Global Telecommunication System (GTS) is generally not quality controlled before dissemination and may contain errors. Since these errors may distort verification results it is desirable to apply some sort of QC. In order not to favour any model in inter-comparison studies, the QC should be model-independent. This excludes methods which identify erroneous observations by comparison with model values and basically leaves techniques which test spatial or temporal consistency among observations.

Given the widely varying density of the SYNOP network across the globe, spatial consistency does not appear to be a viable option to be applied everywhere. This is why temporal consistency is the method tested here. Specifically, erroneous 2-m temperature observations are identified by comparing each observation to the one 24 hours before and after at the same station. If the temperature rises and drops (or drops and rises) over the three days by an amount larger than a given threshold ΔT_{max}, the observation is considered erroneous. Note that strong temperature changes due to air-mass changes and fronts are generally not reversed after just one day, so they would not be wrongly identified as erroneous (Figure 1).

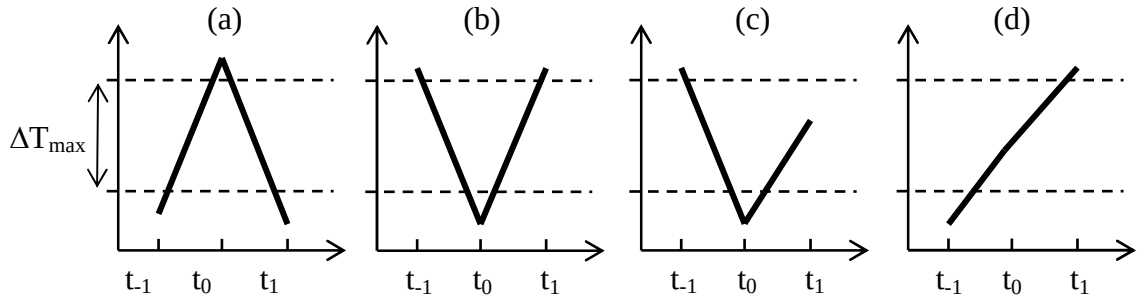

Figure 1. Schematic illustration of the temporal consistency check. Temperature observations at t_0 would be identified as erroneous in cases (a) and (b), but not in (c) and (d).

Proper choice of the threshold ΔT_{max} is important in order to have an effective filter and at the same time not exclude actual cases of strong up-down or down-up changes in temperature. In order to see what magnitude of change could occur naturally, ECMWF analysis data at grid-points nearest to SYNOP locations has been statistically evaluated. Note that ECMWF analyses are used here only to inform the choice of threshold. No forecasts are used in the actual quality control. Figure 2 shows the tails of the distributions of up-down and down-up episodes at 00 and 12 UTC obtained for the period 1 Jan to 31 Dec 2013.

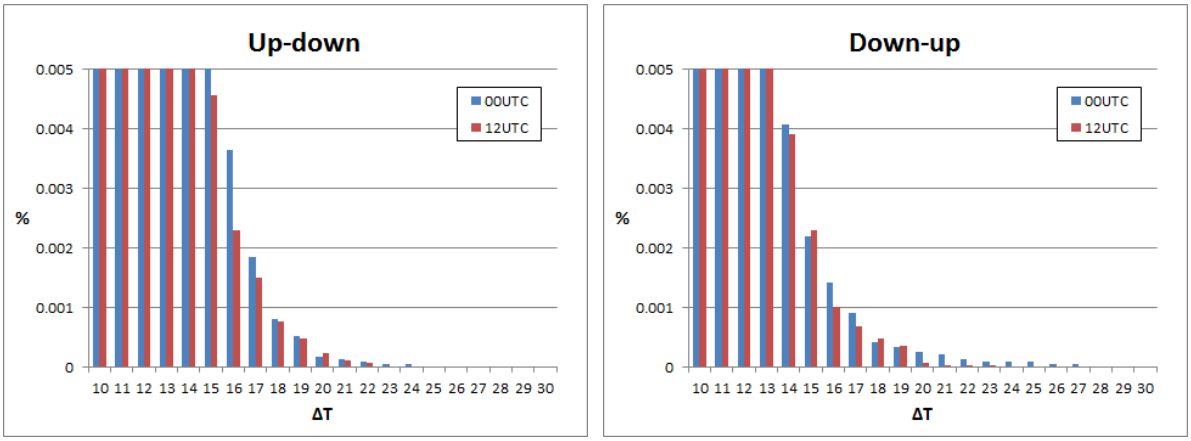

Figure 2. Distribution of 1-day transient warming and cooling episodes exceeding a given magnitude in the operational ECMWF analysis at grid-points closest to SYNOP locations in 2013 (global domain). The total number of cases in each distribution is about $2.4 \cdot 10^6$.

There are minor differences between 00 and 12 UTC, and a tendency for strong short-term warming episodes to occur more often (up to about 18 K) than cooling episodes. The cooling episodes appear to extend to a somewhat higher magnitude, however the differences are well within sampling uncertainty. These results suggest that a threshold of 30 K is unlikely to erroneously exclude more than a very small number of real cases. Keeping this threshold in mind we now look at the corresponding distributions for SYNOP observations (Figure 3).

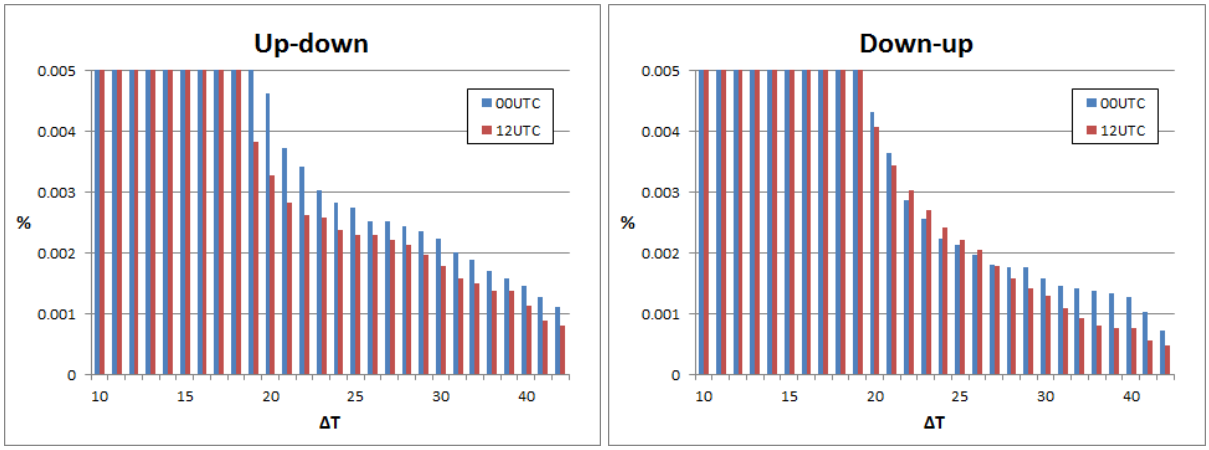

Figure 3. Same as Figure 2 but derived from SYNOP observations. Note the change of scale of the x-axis compared to Figure 2.

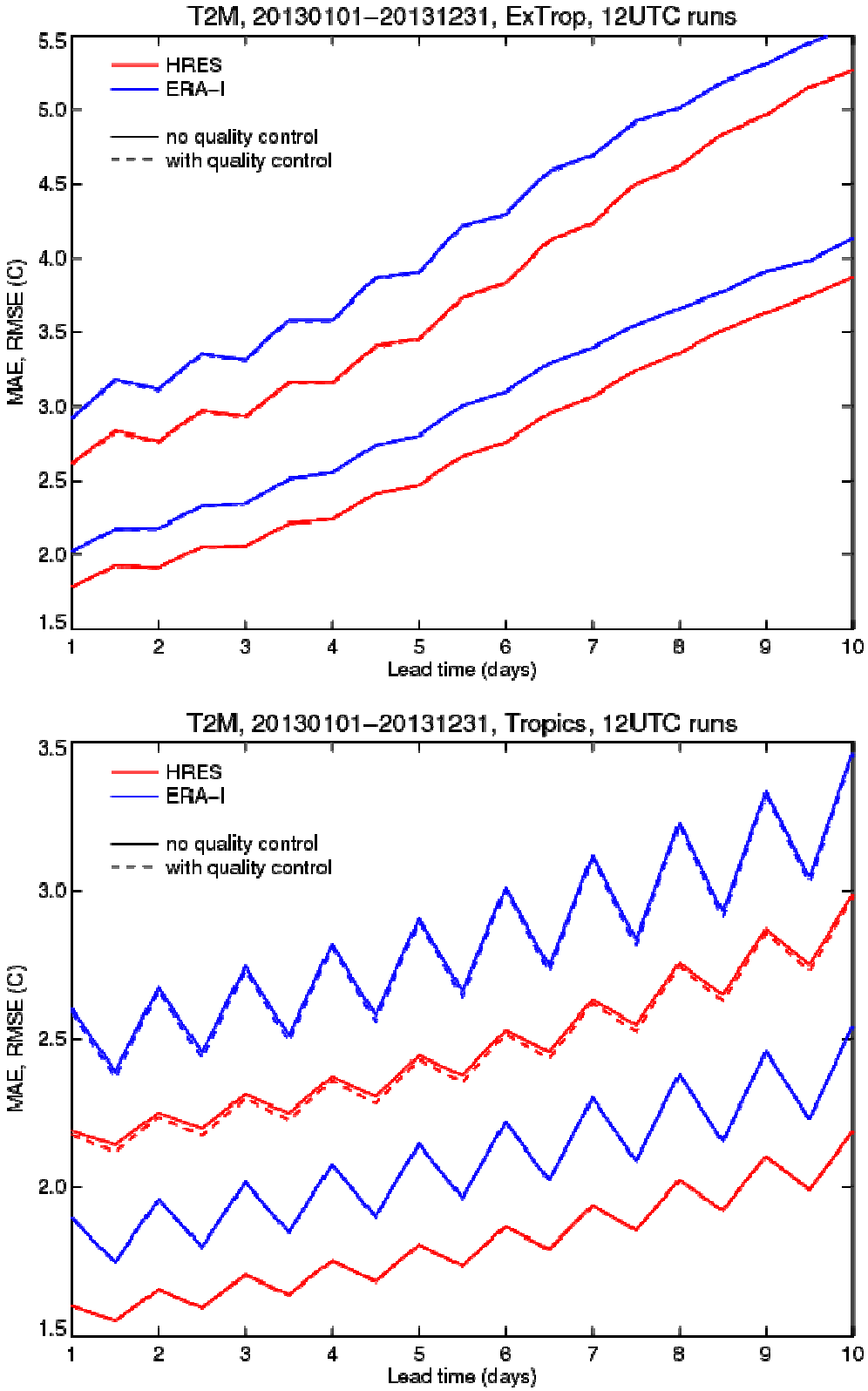

Figure 3. Mean absolute error (MAE) and root mean square error (RMSE) for 2-m temperature from the 12 UTC runs of the operational forecast (red) and ERA-Interim (blue), verified using observational quality control (dashed), and without quality control (continuous). Verification domains are the extra-tropics and the tropics. The verification period is 1 Jan – 31 Dec 2013.

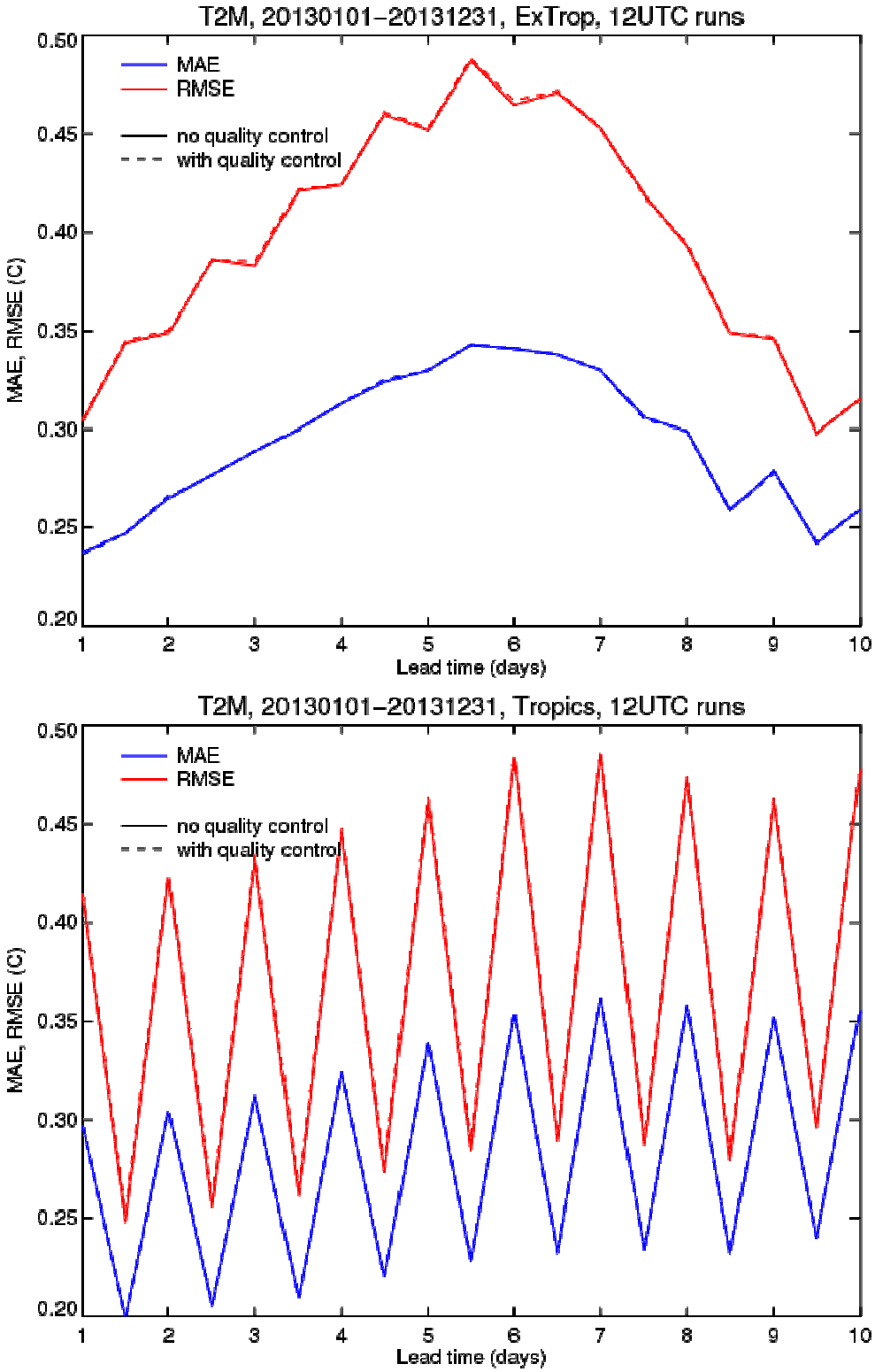

Figure 4. Differences between ERA-Interim and the operational forecast of mean absolute error (MAE, blue) and root mean square error (RMSE, red) for 2-m temperature from the 12 UTC runs verified using observational quality control (dashed), and without quality control (continuous). Verification domains are the extra-tropics and the tropics. The verification period is 1 Jan – 31 Dec 2013.

In the observations the distribution extends to much higher values, with 0.001-0.002 % of cases in the meteorologically unlikely range >30 K. Thus a filter threshold of 30 K would indeed identify a noticeable number of likely erroneous SYNOP observations. In the up-down plot of Figure 3 the

distribution is relatively flat around 25-27 K which suggests that this range may separate the real from the erroneous cases. It also coincides with the range where the highest values are found in Figure 2, adding confidence to this assumption. In the corresponding down-up plot of Figure 3 an indication of flatness is only seen at higher values, in the range 35-40 K. This could indicate that there are indeed real cases of transient cooling of that magnitude which are not captured by the ECMWF analysis. Thus, for the following QC sensitivity experiment we conservatively choose 30 K as a limit for 1-day transient warming and 40 K for 1-day transient cooling.

The effect of such QC on 2-m temperature verification is shown in Figure 3. The mean absolute error (MAE) is not noticeably affected. As expected the root mean square error (RMSE) is more sensitive due to its quadratic nature, however in the extra-tropics the effect is practically negligible. In the tropics, where the number of SYNOP stations is smaller by about an order of magnitude, the effect on the RMSE is slightly larger but still only about 1% at shorter lead times, and less than that at longer ranges.

To what extent these differences could potentially affect model inter-comparison is shown in Figure 4. It can be seen that the effect of QC on the difference in scores between the two forecasts is hardly visible both in the extra-tropics and the tropics. From these results it could be concluded that QC has little effect on scores and scores inter-comparison. However, it should be noted that the QC applied here is quite simple and filters out only part of the erroneous observations. A more sophisticated QC should be able to identify more observations as erroneous and consequently have a somewhat larger effect on scores.

3. Nearest grid-point v nearest land grid-point

Because of the finite resolution of NWP models, the nearest grid-point for a coastal station may be an ocean (or lake) point. If the station is actually located a few kilometres inland, the forecast at that grid-point will not be the most suitable one for verification. Another one of the four surrounding grid-points may represent conditions at the stations much better. However, if the station is located close to the coast of a small island not resolved by the model, local conditions will be closer to those at an ocean point, and simply using the nearest grid-point will be more appropriate. The magnitude of the land-sea contrast in parameters such as 2-m temperature or 10-m wind will furthermore depend on the time of day and season. Contrasts are especially enhanced at higher latitudes during clear, calm nights when there is snow on the ground and the ocean is free of sea-ice.

In order to see how strongly the choice of nearest land grid-point v nearest grid-point affects verification results for a larger domain, we compare both methods for 2-m temperature and 10-m wind speed in Europe (including North Africa), which has a large number of stations near coasts. Since the temperature contrast undergoes strong diurnal and seasonal cycles, we verify 00 and 12 UTC, and winter and summer seasons separately. The nearest land grid-point is chosen out of the four grid-points surrounding the station. If none of them is a land point, then the nearest grid-point is used. Thus, no station is dropped, and the set of stations verified is the same with both methods. Note that all results given refer to the average over all stations within the domain, not just the coastal ones.

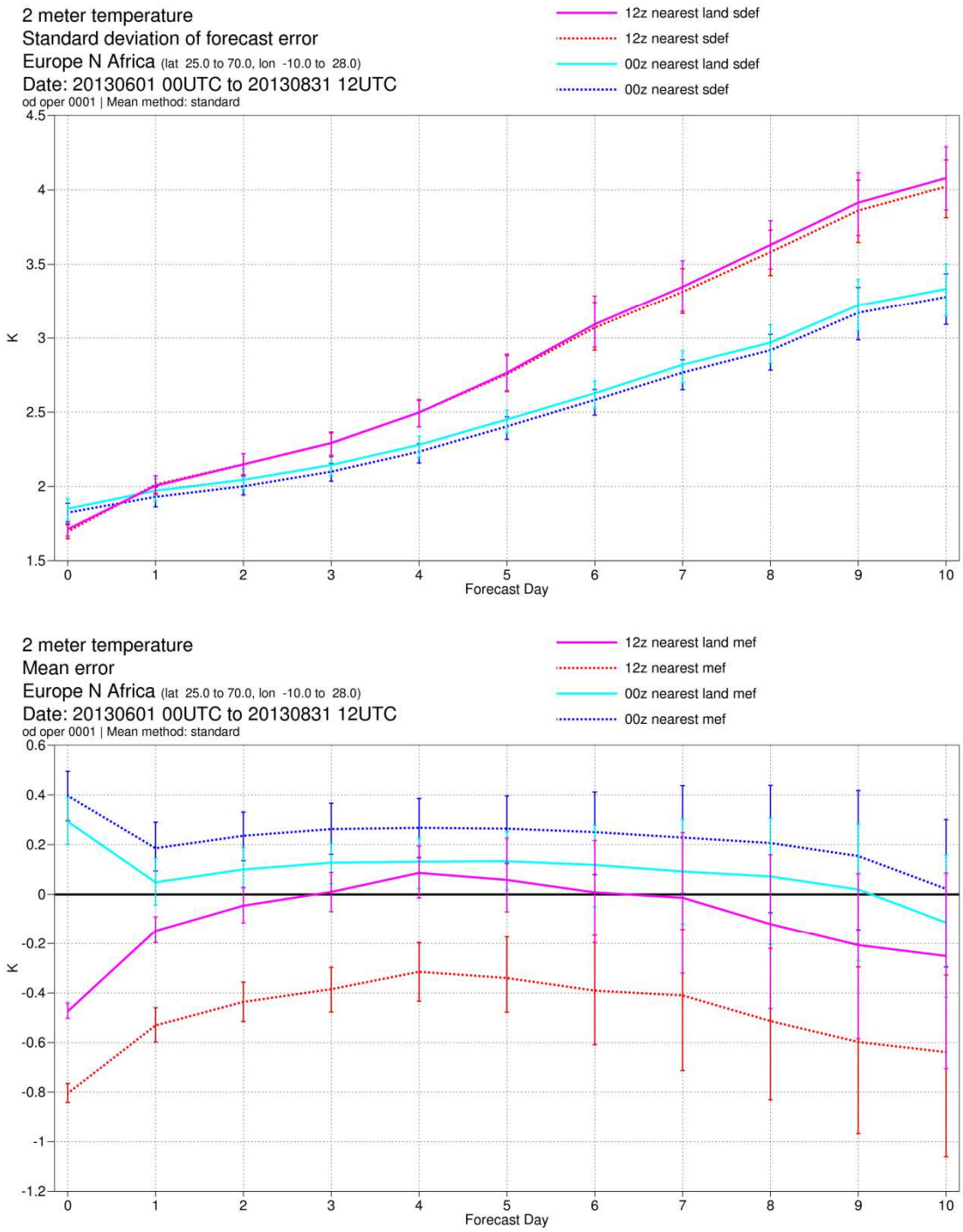

Figure 5. Error standard deviation (top) and mean error (bottom) of the ECMWF HRES forecast (T1279, 16 km) of 2-m temperature in Europe (including North Africa). Shown are results for 00 and 12 UTC using the nearest land grid-point (continuous curves) and the nearest grid-point (dotted). Verification period is JJA 2013. Uncertainty bars indicate 95% confidence and are based on a bootstrapping method.

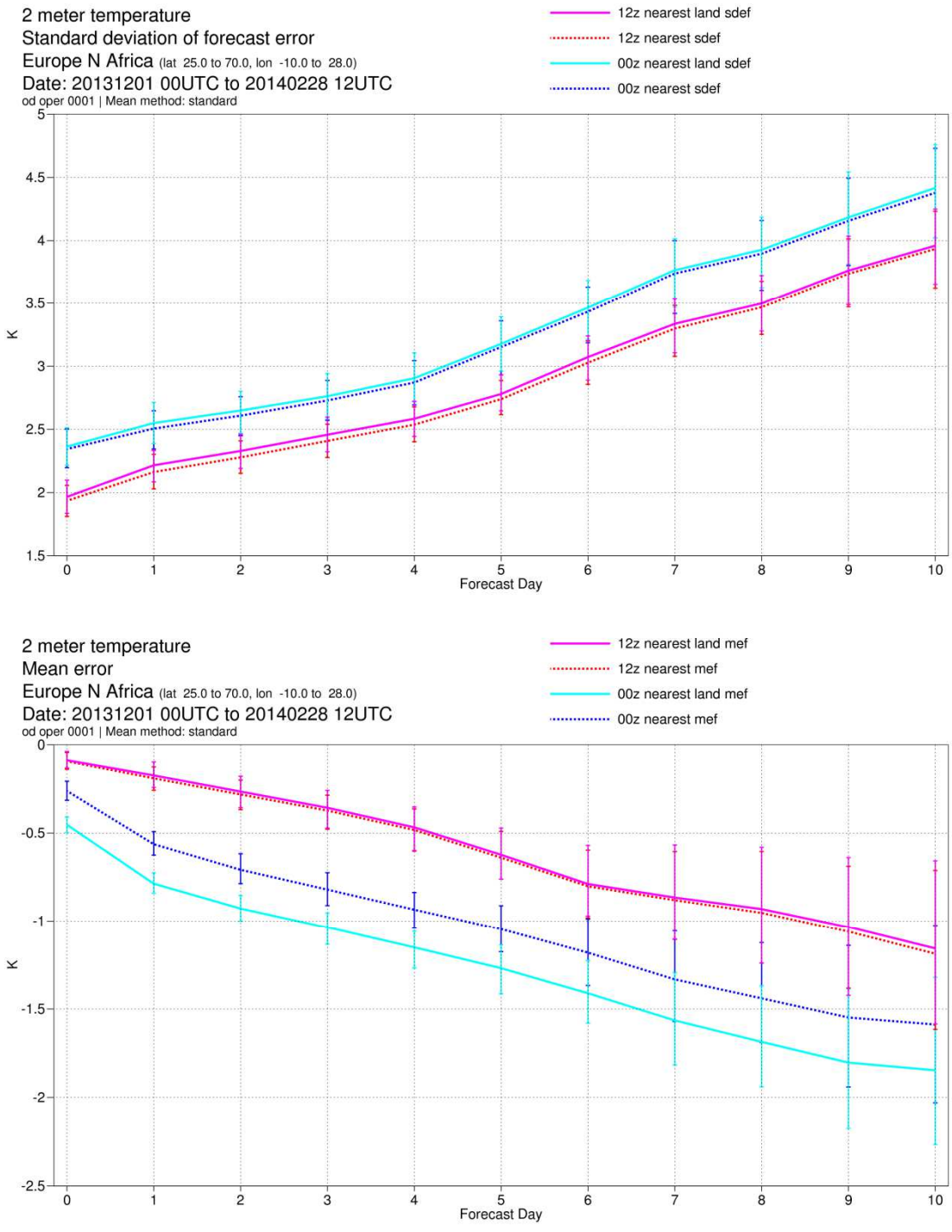

Figure 6. Same as Figure 5 but for the period DJF 2013/14.

(a) 2-m temperature

The non-systematic error of 2-m temperature is slightly increased when the nearest land point is used (Figure 5, top panel). The 2-m temperature forecast at land points undergoes a stronger diurnal cycle, which increases the likelihood of larger errors. This effect apparently slightly outweighs the more suitable choice of grid-point. The difference is smaller than 0.1 K and amounts to about 3-4% of the error. The systematic error (Figure 5, bottom panel), in contrast, is noticeably reduced by using the nearest land point. The positive bias of 0.2-0.3 K at night is reduced by about 0.1 K, and the negative bias of 0.3-0.7 K during the day is reduced by about 0.4 K.

In the winter season (Figure 6), differences in error standard deviation are very similar to those in summer, however the changes in mean error are different. AT 12 UTC there is hardly any effect, apparently due to the weak daytime warming in this season. At 00 UTC the model has a negative bias of the order of 0.5 to 1.5 K which increases with lead time, and which is increased further by about 0.2 K when the nearest land point is used.

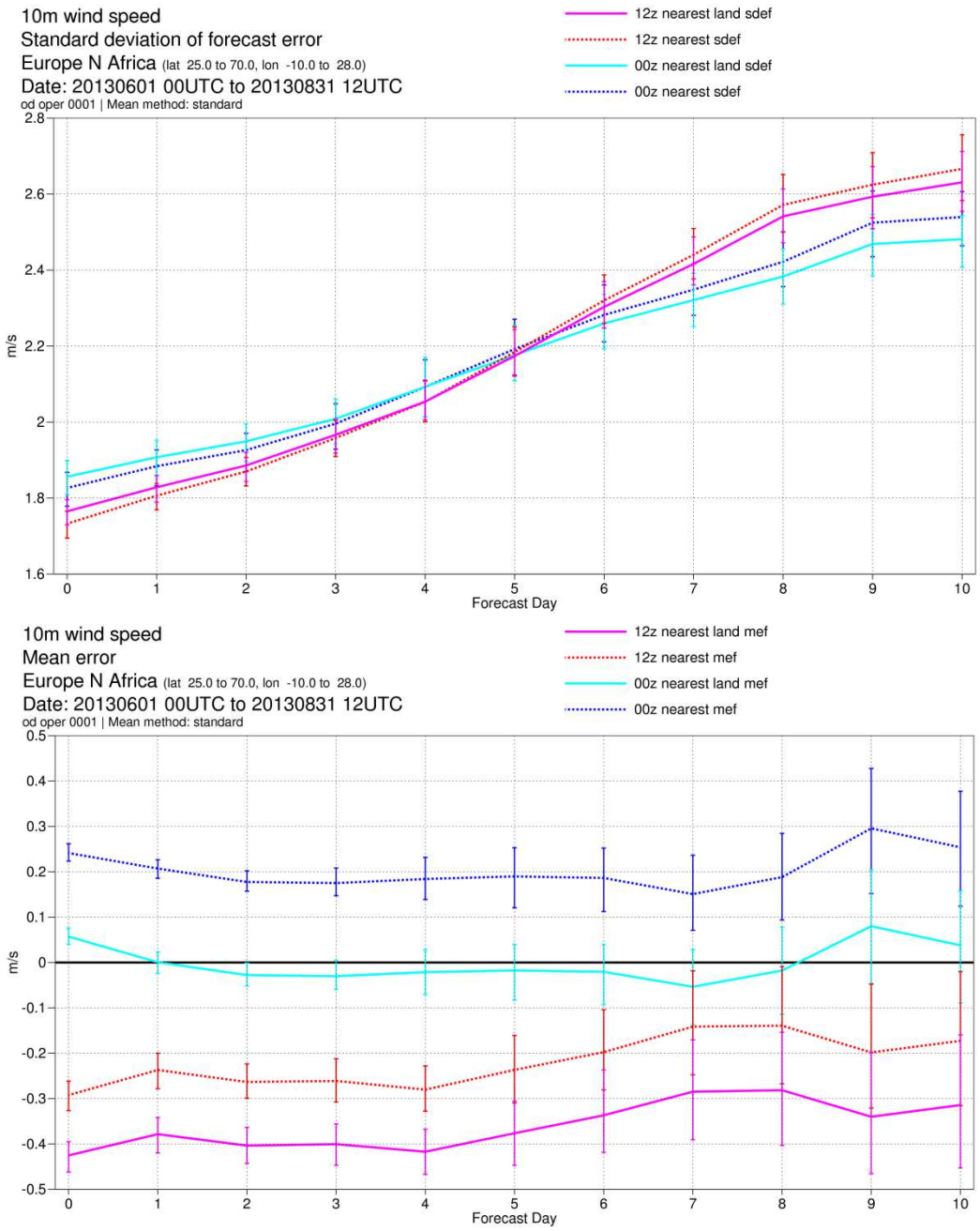

Figure 7. Same as Figure 5 but for 10-m wind speed.

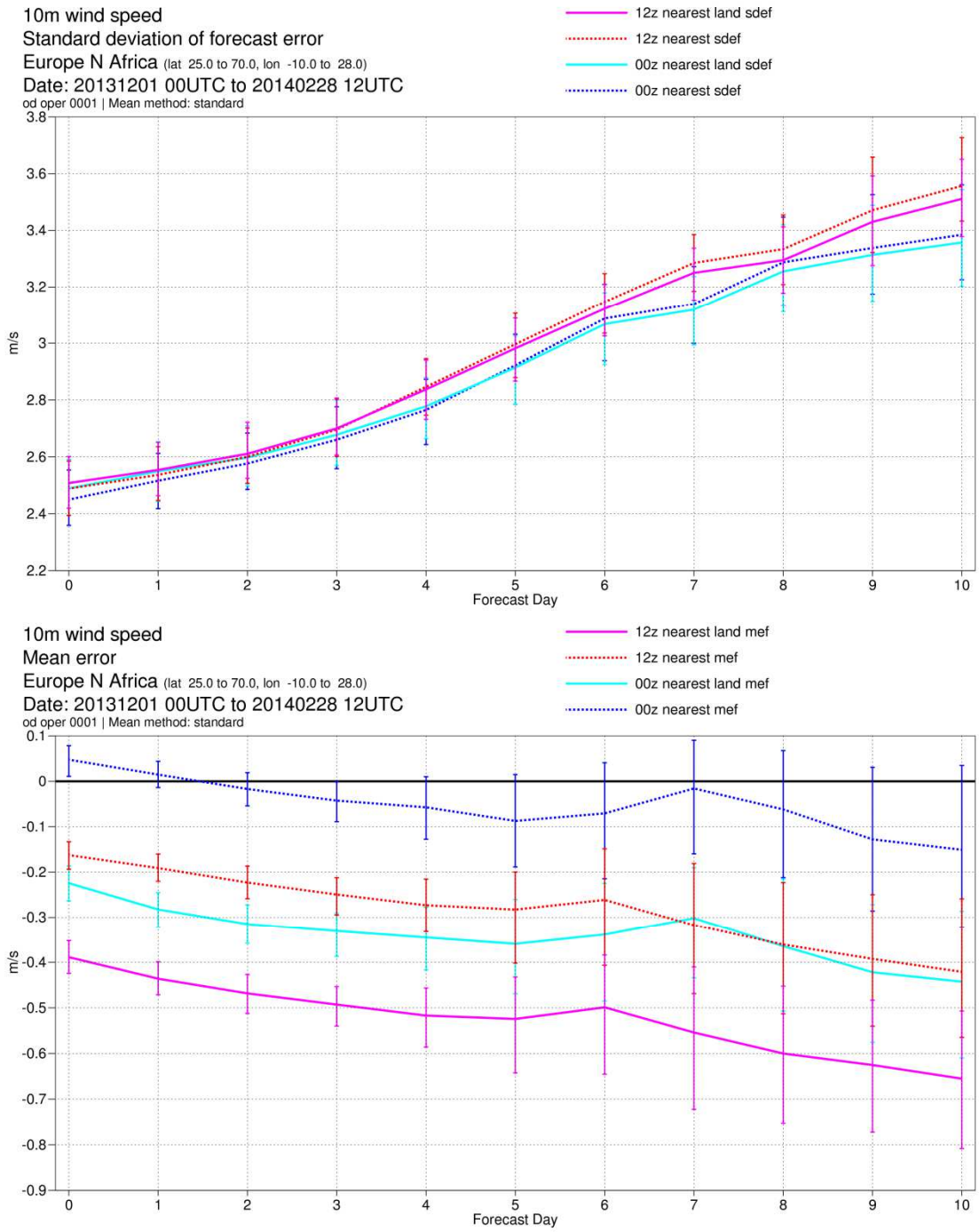

Figure 8. Same as Figure 6 but for 10-m wind speed.

(b) 10-m wind speed

Both in summer (Figure 7) and winter (Figure 8) and both at 00 and 12 UTC the non-systematic error of 10-m wind speed increases in the short range and decreases in the longer range if the nearest land point is used. It is unclear what causes this behaviour. The crossing-over happens in the range between 3 and 5 forecast days. In relative terms, differences are of the order of 1-2% and thus somewhat smaller than for 2-m temperature.

As expected, the mean error of 10-m wind speed is shifted to less positive (or more negative) values when the nearest land point is used. Whether this is beneficial depends on the sign and magnitude of the mean error the model has in the domain considered. In the ECMWF model, this

is the case only in summer at night, where the positive bias is reduced. The magnitude of the changes in mean error is generally in the range 0.1-0.3 m/s.

4. Height correction in 2-m temperature verification

When 2-m temperature is verified, it is common practice to correct for the difference in elevation between the model topography and the station by applying a standard-atmosphere gradient of -0.0065 K/m. While this value is a reasonable approximation of the annual mean temperature profile in mountain areas, it tends to underestimate the gradient in summer when the atmosphere is well mixed, and (even more strongly) overestimate it in winter, when inversions and cold air pools form in valleys and basins. In such cases it may happen that applying no height correction at all puts the forecasted temperature closer to the observed one.

In order to see the effect of the standard height correction on verification results, 2-m temperature verification has been performed for the DJF 2012/13 and JJA 2013 periods with and without height correction. Results are shown in Figures 9 and 10 for the ECMWF high-resolution forecast and ERA-Interim for different large domains. The ME shows that including height correction generally leads to warming (because there are more stations in unresolved valleys than on unresolved peaks), that it is beneficial in terms of MAE and RMSE, and that the effect is stronger for the lower resolution forecast (ERA-Interim), as would be expected. For model inter-comparisons this means that by using a height correction, part of the benefit of having higher resolution (i.e. the fact that station locations are represented closer to their true height) gets masked. On the other hand, one may actually want to exclude this aspect of improvement since it can be easily achieved by a simple height correction. Comparing Figures 9 and 10 we see that the correction is generally more successful in the extra-tropical domains during summer.

5. Summary

It has been shown that a simple quality control for SYNOP 2-m temperature based on temporal consistency has a negligible effect on scores and score comparison between different forecasts. Using the nearest land grid-point instead of the nearest grid-point leads to differences in domain-averaged errors of the order of a few %. It has a large but necessarily beneficial effect on 2-m temperature and 10-m wind speed biases. Applying a height correction to 2-m temperature has the largest effect and it is generally beneficial. It reduces the score differences between forecasts with different resolutions.

References

WMO, 2012: Final report of the CBS Coordination Group on Forecast Verification (CG-FV), Reading, 28p.

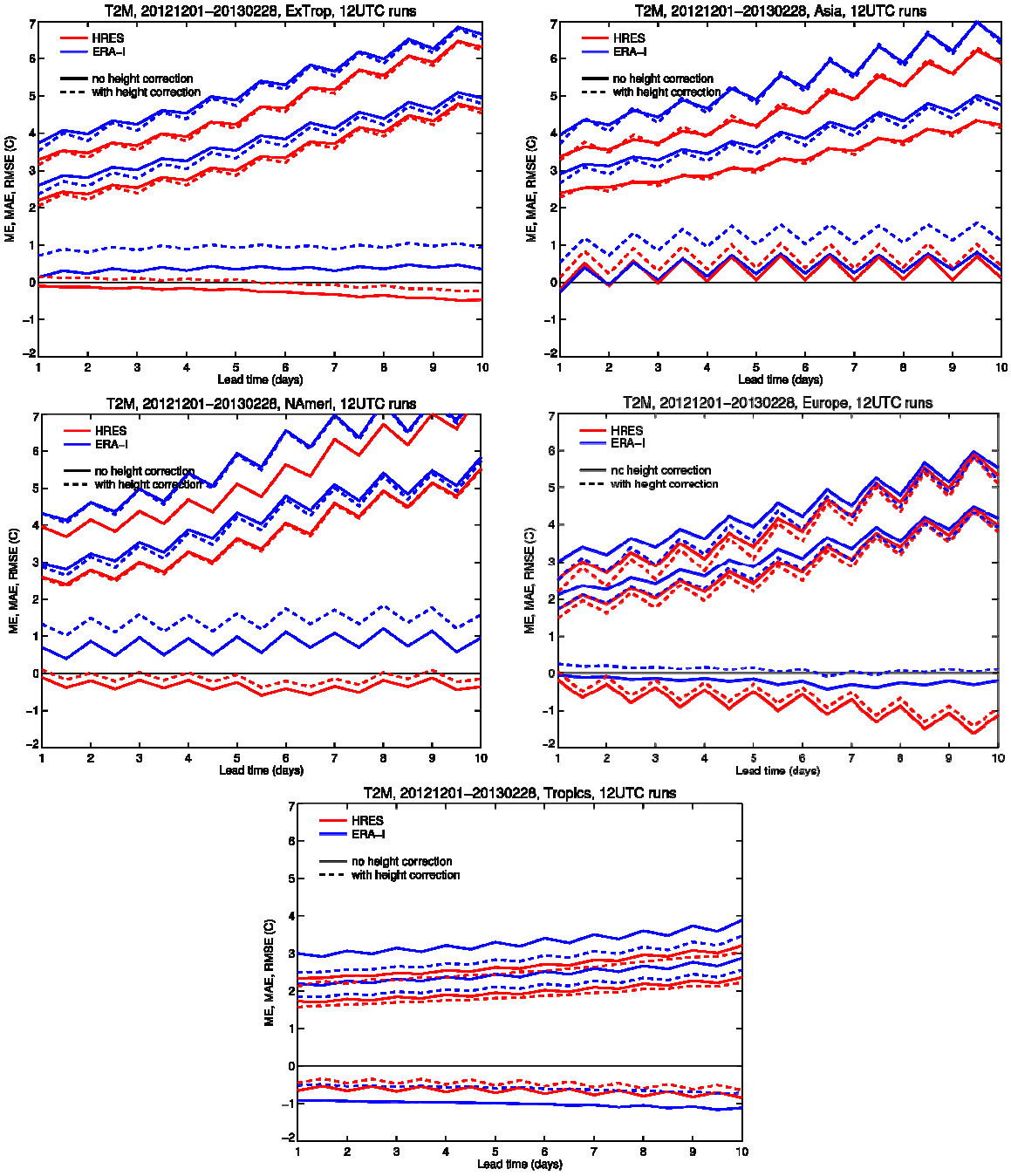

Figure 9. Mean error (ME), mean absolute error (MAE), and root mean square error (RMSE) for 2-m temperature from the 12 UTC runs of the operational forecast (red) and ERA-Interim (blue), verified with a height correction of -0.0065 K m^{-1} (dashed), and without height correction (continuous). Verification domains are the extra-tropics, Asia, North America, Europe, and the tropics. The verification period is DJF 2012/13.

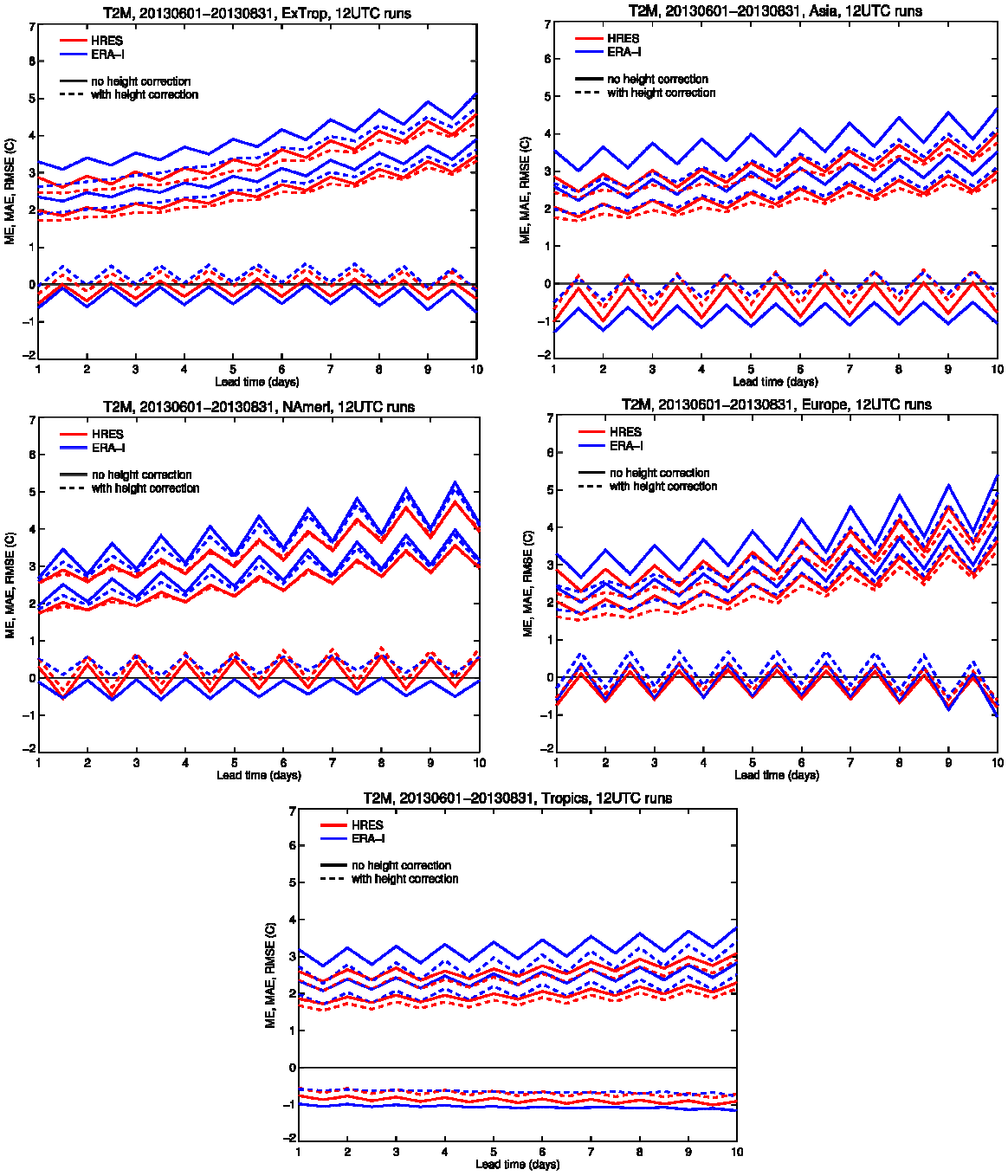

Figure 10. Same as Figure 9 but for JJA 2013.