Technical
Memo

ECMWF
European Centre for Medium-Range
Weather Forecasts

875

Addressing near-surface forecast biases: outcomes of the ECMWF project 'Understanding uncertainties in surface-atmosphere exchange' (USURF)

Irina Sandu, Thomas Haiden, Gianpaolo Balsamo, Polly Schmederer, Gabriele Arduini, Jonathan Day, Anton Beljaars, Zied Ben Bouallegue, Souhail Boussetta, Martin Leutbecher, Linus Magnusson and Patricia de Rosnay

November 2020

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

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

Abstract

The demand for more accurate near-surface weather forecasts is rapidly increasing, driven by various factors such as renewable energy applications, or the occurrence of more intense and frequent extreme events. In this context it is becoming increasingly important to reduce systematic biases in near-surface weather parameters, such as temperature, humidity or winds, which manifest at all forecast ranges. These biases are the result of a complex interplay between processes parametrized in the atmospheric and surface columns of the forecasting system, which can lead to locally generated errors, and advection, which constitutes a non-local source of errors. Understanding the leading causes of these systematic errors, which often have complicated geographical patterns and temporal structure, is a necessary step to enhance the near-surface forecast accuracy and improve the realism of the model. This requires disentangling the role of individual processes by using a range of diagnostics for stratifying and attributing errors. The ECMWF internal USURF project, which was initiated in 2017, has coordinated efforts in this area, and its main results as well as the necessary model developments to reduce systematic biases in forecasts of near-surface weather parameters over continental regions are summarized here. This report was a special topic paper presented to ECMWF's Scientific and Technical Advisory Committees in October 2020.

1 Motivation of USURF

The skill of ECMWF's Integrated Forecasting System (IFS) has been gradually increasing over time due to improvements in the forecast model, data assimilation, and the observing system and to advances in supercomputing that can support enhanced process fidelity through increased model resolution (Bauer et al., 2015). Verification, against both analyses and observations, shows that errors in the ECMWF medium-range forecasts keep getting smaller in the free atmosphere as well as near the surface (Haiden et al., 2019, and Section 2). However, as shown in the next section, systematic biases in near-surface weather parameters such as temperature, dew point and winds remain, hindering further progress. Systematic forecast biases in near-surface weather parameters are not only of concern for ECMWF, but are one of the major open issues in the wider Numerical Weather Prediction (NWP) and climate modelling community, as highlighted by the recent survey on systematic model errors of the Working Group for Numerical Experimentation (WGNE) of the World Meteorological Organization (Reynolds et al., 2019). Eliminating, or at least reducing, these biases is becoming increasingly important in the context of an enhanced demand for more accurate near-surface weather forecasts, driven by various factors such as renewable energy applications, or the occurrence of more intense and frequent extreme events.

To reduce biases in near-surface weather parameters in a physically consistent way it is necessary to understand their causes and their spatio-temporal patterns. In 2017, ECMWF initiated the cross-departmental project USURF ('Understanding uncertainties in surface-atmosphere exchange') aimed at identifying underlying issues in the representation of near-surface exchange processes, and how they might be improved.

While the primary focus in USURF has been on near-surface temperature over continental areas due to its importance for forecast users and applications, near-surface humidity and wind errors have been investigated as well, albeit to a lesser extent. Biases in these three variables are closely connected physically through the surface energy balance and vertical mixing processes. Precipitation has not been the subject of this project, although it is closely monitored through various scores and metrics as part of the routine verification at ECMWF. Earlier work specifically focused on monitoring improvements in precipitation forecast skill, for example by using the newly introduced Stable Equitable Error in

Probability Space (SEEPS) (Rodwell et al., 2010; Haiden et al., 2012) or by using high-density observations from Member and Cooperating States (Haiden and Duffy, 2016). Continued improvements in precipitation skill are being seen in the IFS ever since (Haiden et al., 2019).

Systematic errors in near-surface weather parameters often have complicated geographical patterns and temporal (diurnal, seasonal) structures, and disentangling their causes may be difficult due to the interplay between many processes. These include local effects from atmospheric processes and coupling to the land/snow/ocean/sea-ice, and remote effects leading to systematic errors in synoptic weather patterns. Although data assimilation of atmospheric and surface observations plays a crucial role for the quality of ECMWF forecasts, the main focus of USURF was on the sources of model error related to the representation of physical processes in the atmosphere and land-surface and not on data assimilation aspects. More specifically the focus was on the representation of fast boundary-layer processes (turbulent mixing), on land-atmosphere coupling and land processes which are among the main contributors to systematic errors in near-surface temperature, dew point and winds. Although USURF did not specifically focus on the representation of clouds in IFS, it did search to disentangle the respective roles of the representation of cloudy versus clear boundary layers to errors in near-surface temperature and dew point. This stratification was motivated by the fact that errors in radiative forcing due to errors in clouds can also significantly contribute to errors in near-surface temperature.

Due to its focus on systematic errors related to boundary-layer physics and land-surface coupling, the main emphasis of USURF was on the high-resolution deterministic ten-day forecasts (HRES) at 9 km spatial resolution. However, it was also investigated to what extent biases are different in the control (and mean) of the ensemble (ENS) forecasts (CTL, at 18 km spatial resolution) compared to the HRES, and to what extent an apparent lack of 2m temperature reliability in the ENS is due to representativeness issues related to the mismatch between model grid scale and point observations (Schmederer et al., 2019). While it is acknowledged that a large amount of satellite data supports the development of surface and near-surface forecasts (Balsamo et al., 2018), interpretation of satellite data requires complex operators. Therefore, the diagnostic work within USURF has focused on using in-situ measurements as ground-truth verification. This raises the question of representativeness, which is discussed in Section 7. Much of the analysis included in this report focuses on the period 2016-2018, when most of the project work was undertaken. Different winter and summer seasons selected from these years are used throughout the document to illustrate various points, but the systematic biases which are discussed hereafter are very similar from one year to the next.

This report, which was a special topic paper presented to ECMWF's Scientific and Technical Advisory Committees in October 2020, is structured as follows. Section 2 illustrates the continuous improvements in forecasts of near-surface weather parameters and highlights the remaining systematic biases. Section 3 contains a few methodological considerations regarding the approach followed in USURF to disentangle the causes of these remaining biases. Section 4 describes the main findings on near-surface temperature biases, while Section 5 addresses humidity biases. Issues in the forecasting of near-surface wind are discussed in Section 6. The problem of representativeness and ensemble reliability is addressed in Section 7. Ongoing attempts to make progress towards reducing some of the systematic biases discussed here are presented in Section 8. Conclusions and future steps are presented in Section 9.

2 State-of-the-art in ECMWF forecasts of near-surface weather parameters

The forecast performance of ECMWF near-surface weather-parameters is continuously monitored and is improving at a steady pace, although seasonal and interannual variability may result in alternate periods of lower and higher predictability, depending on the synoptic situation. A fair way to assess system improvements, accounting for diverse predictability conditions, is to measure the skill with respect to a fixed reference such as the latest ERA5 climate reanalysis (Hersbach et al., 2020).

Figure 1 shows that the skill (in terms of error standard deviation) of ECMWF HRES has continuously increased relative to ERA5 forecasts both for upper-air and near-surface parameters, due to a combination of increased resolution, modelling, and data assimilation upgrades. ERA5 is produced with a fixed IFS cycle (41r2 operational in 2016). All parameters show a positive skill trend in the past decade, albeit with different rates. This is partly because in Figure 1 the verification is against analyses for upper-air parameters, and against SYNOP observations for near-surface parameters, which introduces a certain amount of representativeness error.

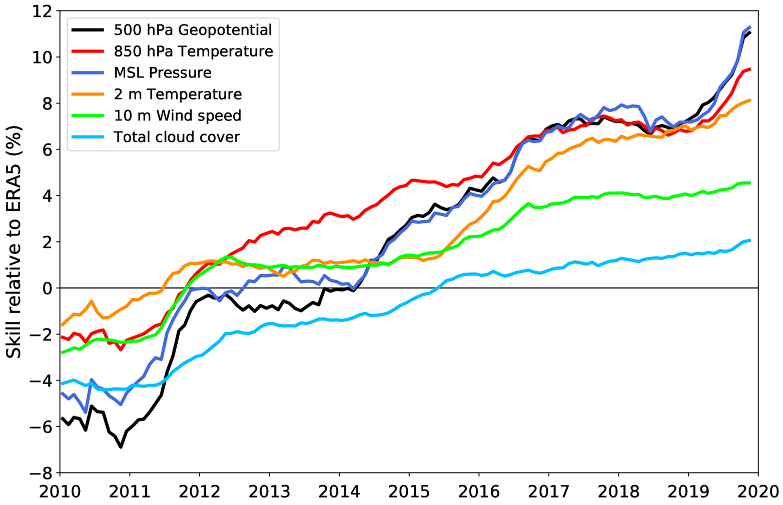

Figure 1: HRES forecast skill for different parameters at day 5 in the North Hemisphere extra-tropics, relative to ERA5 forecasts. The verification metric is error standard deviation. Upper-air parameters and mean sea level (MSL) pressure are verified against own (HRES, respectively ERA5) analysis, near-surface parameters are verified against SYNOP. Shown are 12-month running averages.

A more direct estimation of skill improvement rates for upper-air and near-surface ECMWF HRES winter forecasts, is shown in

Figure 2. Here, the increase in skill is shown as a lead time gain over the 13-yr period from 2006 to 2019 (measured in terms of RMSE against analysis). The length of this period was chosen such that geopotential forecast skill at 500 hPa has increased during the period by about one day. For other upper-air parameters such as temperature, wind components, and humidity, increases are smaller, which is due to the larger magnitude of smaller scales in these fields. This can be seen from the right panel in

Figure 2, which shows results for large scales only: except for humidity, the upper-air improvements are now very similar in the mid-troposphere. Towards the surface the curves still differ between parameters however, indicating that additional factors come into play within the atmospheric boundary-layer.

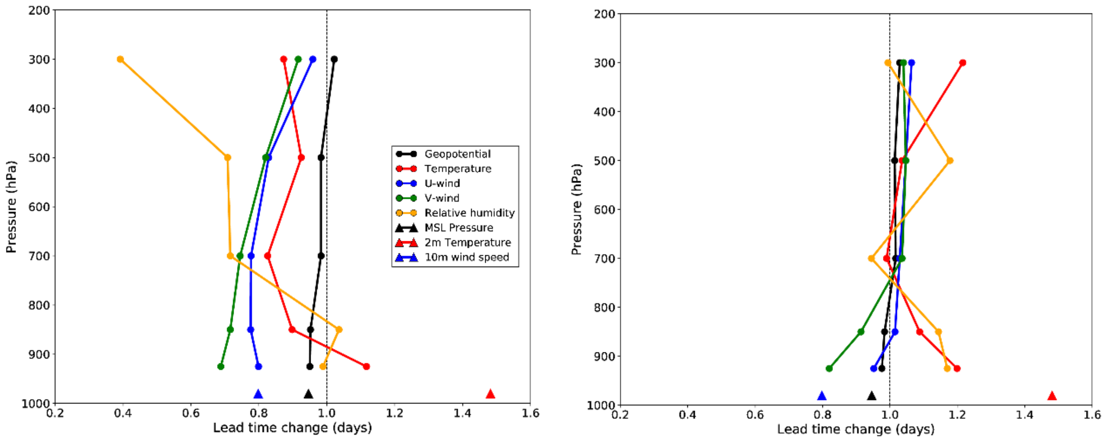

Figure 2: HRES forecast skill improvements, expressed as increases in lead time (days), of various upper-air variables (circle) and associated near-surface parameters (triangles) over the 13-yr period from NDJ 2005-06 to NDJ 2018-19. Left panel: full upper-air fields, right panel: large-scales only (up to wave number 21) for upper-air fields. Verification is against own analysis, the verification metric used is RMSE.

Over the years, non-systematic errors in 2m temperature, 2m dew point, 10m wind speed and direction have become smaller, as indicated by decreases in the error standard deviation in the ECMWF HRES (Figure 3). The ECMWF ENS system is designed to represent the random component of the forecasts, via introduced stochasticity (Leutbecher et al., 2017) and Figure 4 shows that the largest errors of the ENS have also become less frequent, as indicated by the decrease of the fraction of large errors for 2m temperature and 10m wind speed (CRPS>5K and CRPS >4ms^{-1} respectively). It is worth noting that the two most recent major decreases in the largest errors were associated with increases in the horizontal resolution of the ensemble forecasts, namely in 2010 and 2016. The same is true for the HRES forecasts.

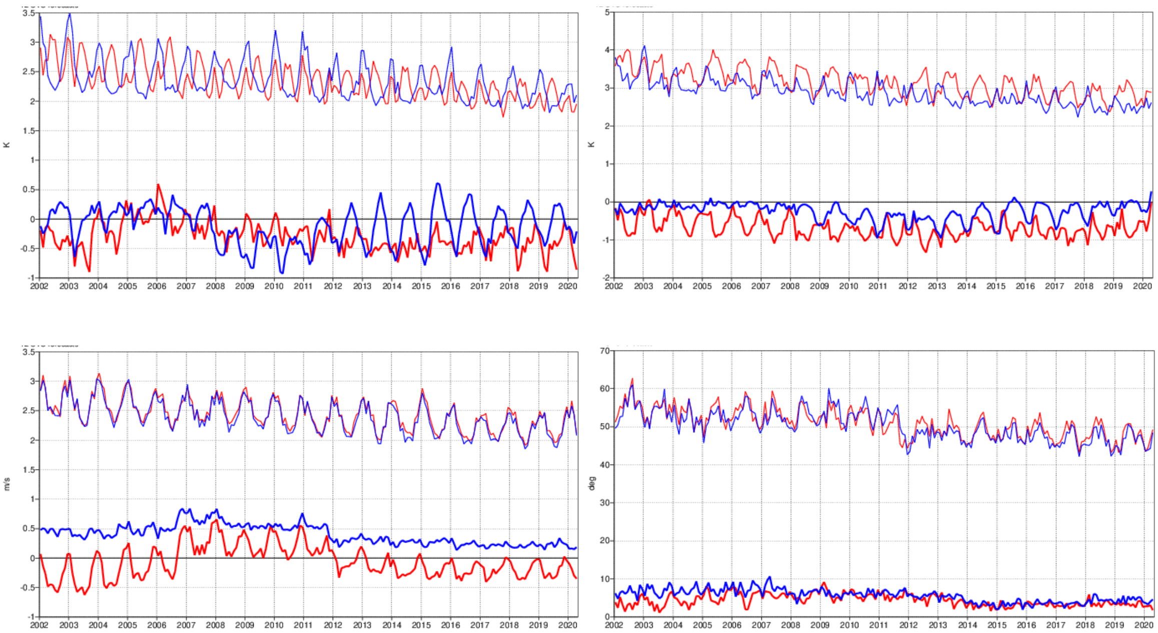

Figure 3: The HRES error standard deviation (thin curves) and mean error (bold curves) of 2m temperature, dew point, wind speed and wind direction forecasts in Europe at day 3 at night (T+60 for the 12 UTC forecasts, blue curves) and during the day (T+72 for the 12UTC forecasts red curves).

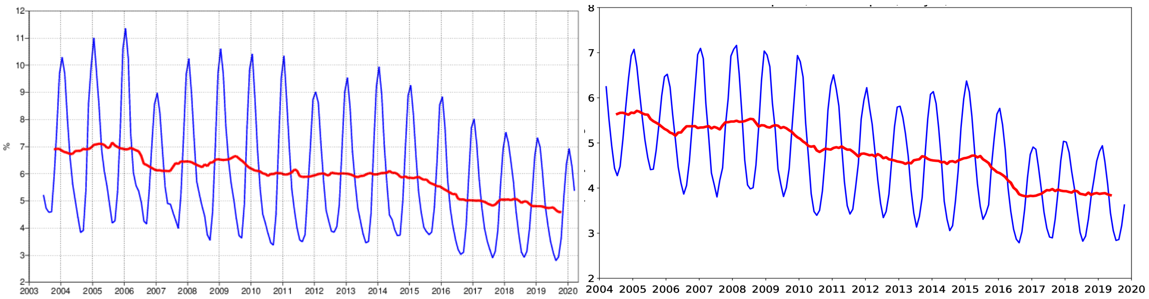

Figure 4: The fraction of large ENS errors (CRPS>5K) in day-5 forecasts of 2m temperature (left) and the fraction of large errors (CRPS>4ms^{-1}) of 10m wind speed forecasts at day 5 (right) for the extratropics. Blue curve: 3-month average values, red curve: 12-month running average. Verification is done against observations from SYNOP stations.

However, as can be seen from Figure 3, the magnitude of the systematic errors (as indicated by the mean biases) has changed little for near-surface parameters over recent years, although for certain parameters these errors have decreased due to targeted changes to the forecasting system (see for example discussion on wind errors in Section 6). Systematic biases thus persist, and they have annual and diurnal cycles, as apparent from Figure 3, and vary geographically, as illustrated for 2m temperature for the European region for a recent winter season in Figure 5. Biases typically vary somewhat with lead time but generally are fairly robust from the short into the medium range. In particular, they do not change in sign as the forecast progresses. This is not surprising given that as it will be shown hereafter these errors are due to a large extent to boundary-layer and land-atmospheric coupling processes which act on very fast (hours) timescales. The geographic and seasonal variations of the systematic forecast biases in near-surface temperature, dew point and winds will be discussed in more detail in sections 4 to 6, after some methodological considerations regarding the approach followed in USURF to disentangle their causes.

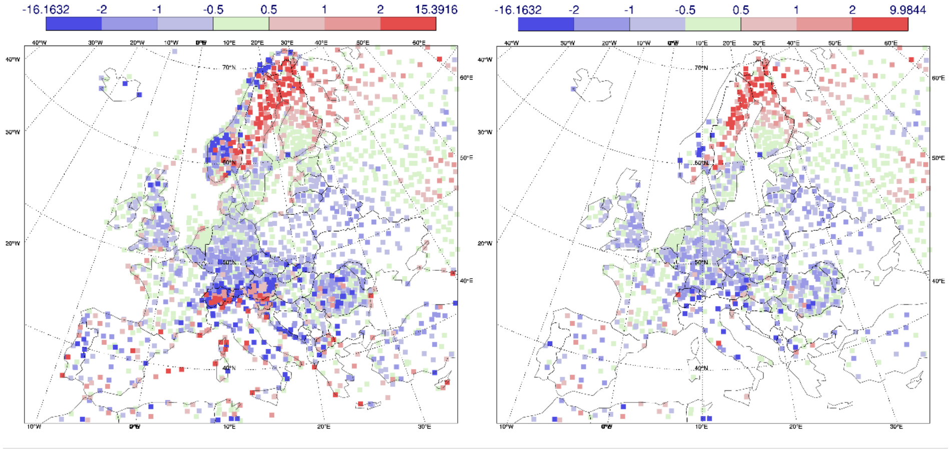

Figure 5: HRES night-time (00 UTC) 2m temperature bias during winter (DJF) 2018-19 at forecast day 3. Verification is done against observations from SYNOP stations. Left panel: all stations; right panel: stations at which the difference between model and actual station height is less than 100 m and where each of the four nearest grid points has >50% land fraction according to the land-sea mask of the IFS HRES.

3 Methodological considerations in USURF

When in-situ measurements are used to evaluate forecasts from a numerical weather prediction (NWP) model, it is important to consider the mismatch between the model grid scale and the scale and representativeness of the (near-)point observation. In the case of 2m temperature, the measurement error is usually small (a few tenths of a degree at most), however the measurement may be of limited spatial representativeness, especially at night when turbulent mixing is reduced. As it will be shown in Section 7, representativity is even more an issue for dew point. Moreover, forecast biases in near-surface weather parameters can be to a certain extent due to discrepancies between model and reality in terms of topography, vegetation cover or soil properties. This is particularly true in regions with complex topography or in coastal areas due to limited spatial resolution. An important methodological approach in the USURF project was to therefore focus on areas away from major mountain ranges, coasts, and lakes. Focusing on ‘simpler areas’ reduces representativity issues and simplifies somewhat the difficult task of identifying the main causes of the systematic errors in near-surface weather parameters. Figure 5 shows for example that by excluding most of the coastal and mountain stations the 2m temperature bias pattern becomes more coherent with a predominantly cold bias in large parts of Europe south of 60°N and a warm bias in northern Scandinavia.

Another methodology that has proven useful in USURF is conditional verification. The stratification of biases in near-surface weather parameters according to various criteria, such as other meteorological parameters or surface characteristics, has for example provided additional insights towards some of their possible causes (Haiden et al., 2018).

One of the novel initiatives within USURF was to build tools which facilitate regular model evaluation against supersite observations. Meteorological observatories, also known as supersites, provide long-term observational records of vertical profiles in the atmosphere, soil or snow and of surface energy budget components, such as surface radiative fluxes. Evaluation against supersite observations allows to go beyond the traditional verification of near-surface weather parameters against SYNOP observations, by allowing an assessment of forecast biases not only near the surface, but also within the soil/snow, in the lower part of the atmosphere and at their interface. An example can be seen in Figure 6**,** which shows temperature profiles for forecasts and observations for a recent summer and winter period at the observational supersite Falkenberg **(**associated with Meteorologisches Observatorium Lindenberg - Richard-Aßmann-Observatorium**,** Germany**).** On the condition that biases at supersites can be regarded as representative of systematic biases over wider areas, evaluation against supersite observations can thus be very helpful for model development. The use of tower and snow/soil profile data from supersites such as Falkenberg (Germany), Cabauw (Netherlands), and Sodankylä (Finland) helped for example (i) identifying the main reasons for the underestimation of the amplitude of diurnal cycle in 2m temperature over land (Schmederer et al., 2019), and (ii) understanding how the surface energy balance, and thereby surface temperature, change in response to a change of the snow scheme in the IFS (Day et al., 2020). Within USURF, and as part of the H2020 project APPLICATE, **s**upersite datasets were thus proven to constitute a valuable resource for ECMWF’s efforts to further reduce forecast errors in near-surface weather parameters.

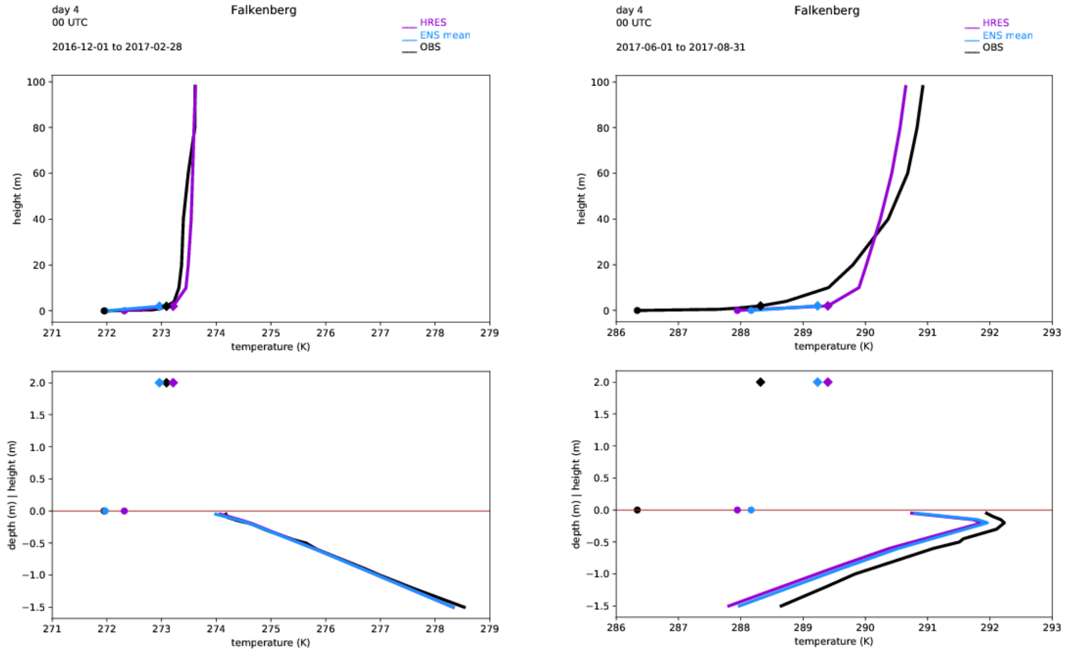

Figure 6: Night-time temperature profiles (00 UTC), in the lowest part of the atmosphere and in the soil, during winter (DJF 2016-17) and summer (JJA 2017) from observations (black), HRES (purple) and ENS mean (blue) at forecast day 4 at the Falkenberg observatory. A good agreement between the forecasts and the observations can be seen during wintertime both in the soil and lower boundary layer, while an overestimation of near-surface temperatures and a too cold soil are found during summer (see discussion in Sect. 4).

Within USURF, the various diagnostic and evaluation methods described above have been complemented by model sensitivity experiments which assess how some of the systematic near-surface biases depend on the representation of atmospheric processes such as turbulent diffusion, land processes (for example snow) or on the land-atmosphere coupling strength. A large part of this experimentation and the lessons learned are summarized in Beljaars (2020).

4 Temperature errors

4.1 Diagnosis of 2m temperature biases

Systematic biases in 2m temperature, with complicated temporal and geographical patterns, persist in the ECMWF forecasts despite consistent efforts to reduce them over time. One of the main reasons is that near-surface temperature is intricately related to, and driven by, a variety of processes: cloud cover and cloud optical properties, radiative transfer, precipitation, surface fluxes, turbulent diffusion in the atmosphere, strength of land-atmosphere coupling, soil moisture and temperature, which in turn depend on land surface characteristics (vegetation, soil type, soil texture, etc.) and processes. Moreover, the representation of these different processes can often to lead to compensating biases, so attributing what are in most regions relatively small biases (Figure 5) to one or more of these processes is therefore not straightforward.

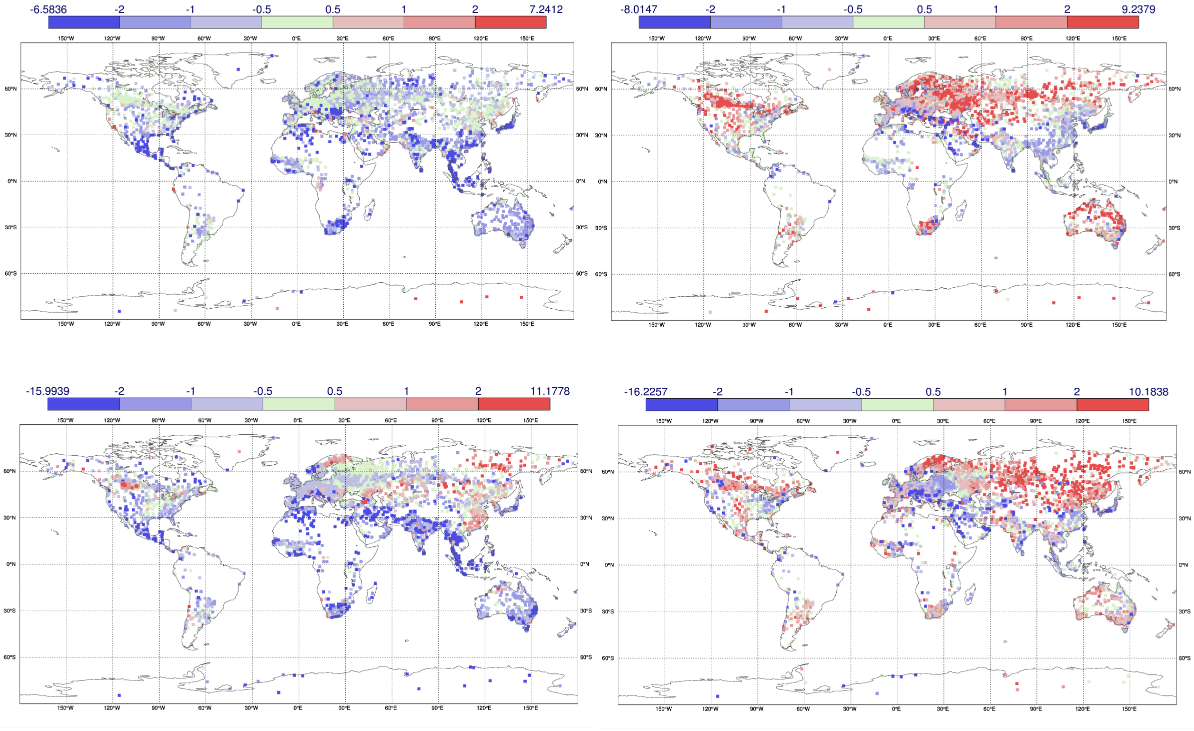

Figure 7: Bias of the 3-day HRES forecast of Tmax (left) and Tmin (right) in summer (JJA) 2018 (top) and in winter (DJF) 2018-19 (bottom). Included are only those stations at which the difference between model and actual station height is less than 100 m and where each of the four nearest grid points has >50% land fraction according to the land-sea mask of the IFS HRES.

Indeed, the geographical distribution of 2m temperature biases does not map onto patterns of surface type or vegetation in a simple way, although some patterns emerge (i.e. larger biases over orography or in snow covered areas, as discussed in Beljaars 2020). This is the case even when instead of a fixed verification time such as 00 or 12 UTC, daily maximum and minimum temperatures are considered to make results comparable across longitudes (Figure 7). 2m temperature biases differ between global models during wintertime as illustrated in Figure 8. (Here, fixed times had to be used because Tmax/Tmin was not available from other centres.) Although some features are shared between the four models shown in Figure 8, such as a cold bias in the eastern US, or a warm bias in northern Scandinavia, there are otherwise considerable differences, even with respect to the largest-scale patterns. During summertime, the patterns of the biases are again model dependent (not shown), albeit all models present the warm bias over the US Southern Great Plains investigated in the CAUSES project led by the UK MetOffice (see Morcrette et al., 2018 and references therein). This inter-model variability is a further indication that the bias patterns are not due to specific weather patterns in a given season, but rather the result of choices made in the representation of processes such as turbulent mixing in the atmospheric boundary-layer or clouds, processes in the land-surface and surface-atmosphere exchange processes.

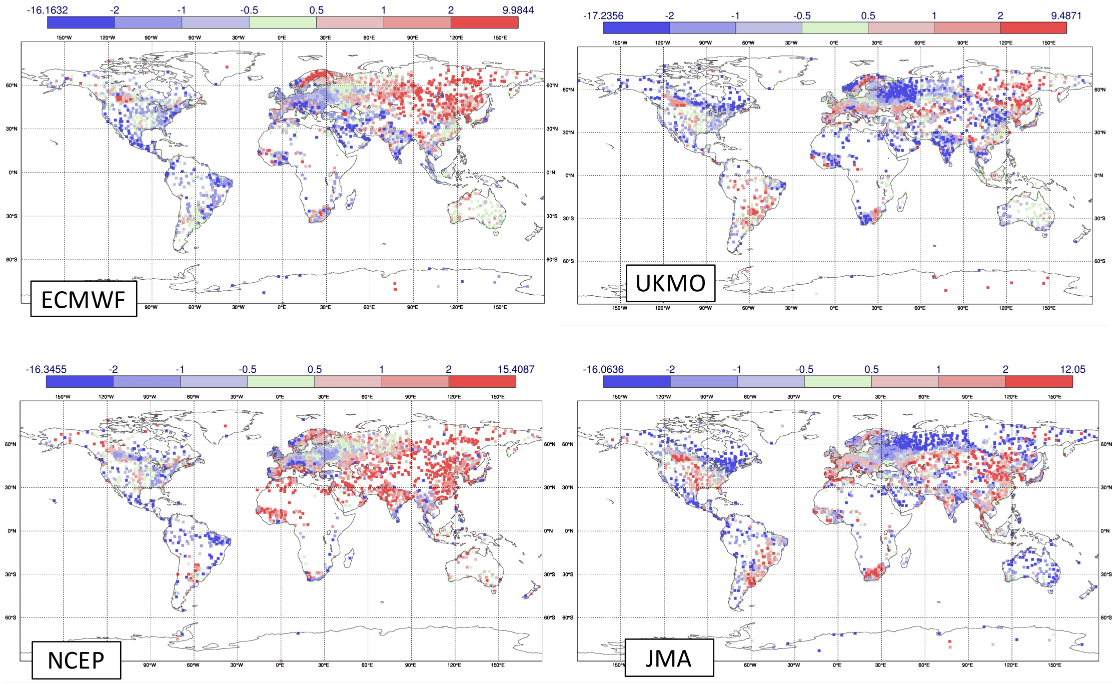

Figure 8: Bias of the 3-day HRES forecast of 2m temperature at 00 UTC in winter (DJF) 2018-19 in four global models. Included are only those stations at which the difference between the respective model's topography and the actual station height is less than 100 m and where each of the four nearest grid points has >50% land fraction according to the land-sea mask of the IFS HRES.

In USURF, the focus was on two of the most systematic near-surface temperature biases which are of importance for ECMWF's Member States: the opposing cold/warm biases over central Europe/Scandinavia in winter (Figure 5), and the underestimation of the diurnal cycle over land during summer. As can be seen in Figure 7, but also from Figure 3, the amplitude of the diurnal cycle is generally underestimated over land in the IFS. In Europe, this is the case especially during summer, when this underestimation reaches ~2 K across large areas (Figure 9). Near-surface temperatures are generally too warm during nighttime and slightly too cold during the day, although the degree to which the amplitude of the diurnal cycle is underestimated depends on region and season.

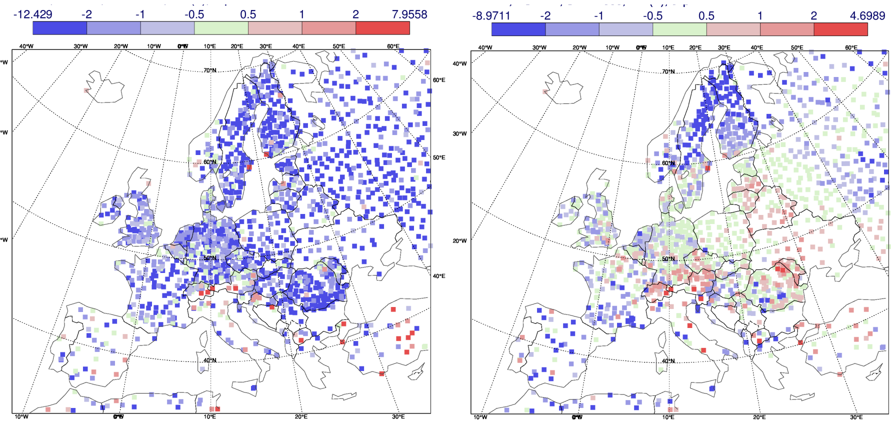

Figure 9: Bias of the 3-day HRES forecast of the diurnal cycle amplitude Tmax-Tmin in summer (JJA) 2018 (left panel) and in winter (DJF) 2018-19 (right panel). Included are only those stations at which the difference between model and actual station height is less than 100 m and where each of the four nearest grid points has >50% land fraction according to the land-sea mask of the IFS HRES.

4.2 Causes of wintertime 2m temperature biases

Diagnostic work and model sensitivity experiments within the USURF project have revealed some of the drivers of the wintertime, nighttime 2m temperature bias pattern in Europe. Apart from coastal and mountain effects, this pattern is characterized by a cold bias in most of Europe south of 60°N, and a warm bias north of this latitude in northern Scandinavia and Russia (right panel in Figure 5). During the day the cold bias changes to neutral/mixed in most parts (not shown), but the warm bias in northern Scandinavia persists.

Using conditional verification (e.g. a stratification of the forecast error in 2m temperature against the forecast errors in cloud cover with respect to SYNOP observations) it was found that part of the negative nighttime 2m temperature bias south of 60°N is associated with an underestimation of (low) cloudiness. As discussed more in detail in Haiden et al. (2018), the wintertime night-time T2m bias in central Europe is smaller for days which are (nearly) clear-sky, both in the forecast and SYNOP observations. This suggests that cloudiness plays a role in this bias. In addition to stratifying T2m forecasts according to a quantity like cloudiness, one can also stratify T2m errors according to the forecast error for cloudiness. The left-hand panel of Figure 10 shows that the night-time negative temperature bias in the IFS in central Europe in winter increases roughly linearly with the amount by which total cloud cover is underestimated (against SYNOP observations). Indeed, even if the prediction of wintertime low stratus in the IFS has noticeably improved over the last decade, some underestimation in cloud cover remains, especially in parts of central and eastern Europe (on the order of 10% against SYNOP observations, not shown). However, when the temperature errors are weighted by the frequency distribution of the cloud cover errors (shown as green bars in the plot), it turns out that cases where the total cloud cover is underestimated and cases where it is nearly correct contribute about equally to the negative T2m bias (Figure 10 right). This indicates that the wintertime negative total cloud cover bias in the IFS over central

Europe does not fully explain the negative night-time T2m bias in this region. In cases when the total cloud cover is correctly predicted, the negative T2m bias could be due to other cloud errors, e.g. an underestimation of cloud optical depth, erroneous cloud type or erroneous cloud base height. It could also be due to errors in processes not directly related to clouds, such as vertical mixing or coupling with the surface.

Some of the warm bias in northern Scandinavia (Figure 5) has been related to the modelling of snow in the IFS (Haiden et al., 2018; Arduini et al., 2019; Day et al., 2020; Beljaars, 2020). In clear-sky nights in the real atmosphere the top snow layer cools rapidly due to longwave radiative heat loss at the surface and reduced heat input from the ground underneath because of the snow insulation properties. The correct representation of these processes is challenging for NWP systems using a single-layer snow scheme and initialisation such as that used currently in the IFS (Dutra et al., 2010; de Rosnay et al., 2014, 2015). The large thermal inertia associated with a deep single layer of snow does not allow the snowpack to cool rapidly enough. A multi-layer snow scheme would help to address this issue by allowing a representation of a thin top snow layer with a lower thermal inertia which can more rapidly respond to changes in the radiative forcing. Work undertaken at ECMWF in the framework of the H2020 APPLICATE project to develop and evaluate the benefits of a multi-layer snow scheme for IFS forecasts has indeed demonstrated the potential for further reducing forecast biases in snow covered areas (Arduini et al., 2019; Day et al., 2020). The benefits of a multi-layer snow scheme, and further work needed to bring this development to an operational implementation will be discussed in Section 8.

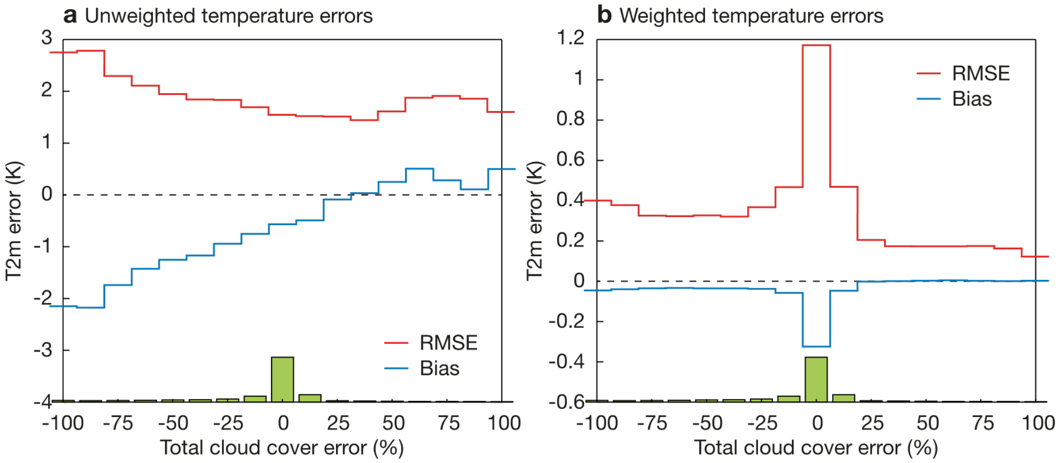

Figure 10: Root mean square error (RMSE) and mean error (bias) for T2m forecasts valid at 00 UTC as a function of the total cloud cover (TCC) error for December–January–February 2016/17 in a central European domain (48–55°N, 0–15°E) at a lead time of 12 hours (a) averaged for each TCC error bin and (b) averaged for each TCC error bin and weighted by the TCC error relative frequency of occurrence. Green bars show the TCC error frequency distribution (arbitrary vertical scale). (from Haiden et al., 2018)

Biases in near-surface temperatures during winter conditions are also very sensitive to the representation of turbulent mixing in stable boundary layers (Sandu et al., 2013). A discussion of the impact of choices made in the parametrization of turbulent mixing in stable conditions on near surface forecasts is also included in Section 8.

4.3 Diurnal cycle of 2m temperature during summertime

As shown in Figure 9, there is a pronounced underestimation of the diurnal cycle amplitude in Europe in summer, which mainly results from too high nighttime temperatures (Figure 7). The use of observations from two European supersites (Falkenberg, Germany and Cabauw, Netherlands) has been key to understanding the reasons of this systematic error.

As described in detail in Schmederer et al. (2019), it was found that the HRES, the ENS mean, but also the ICON model of the German weather service (DWD), underestimate the diurnal cycle of temperature, with larger biases closer to the surface (Figure 11). During night, in all forecasts the temperatures are about 1–2 K too warm at 2 m and about 2 K too warm at the surface. HRES and the ENS mean slightly overestimate the diurnal cycle of soil temperature in the first soil layer, being up to 2 K too cold at night. In all other soil layers, the HRES and the ENS mean are always too cold. ICON is warmer than the IFS in all soil layers during the day, and slightly colder during the night, which leads to a slightly stronger overestimation of the diurnal cycle. These results suggest that too much energy is exchanged between the atmosphere and the land, especially for the IFS. This means, for example, that during the night too much energy is extracted from the soil and transferred to the atmosphere. This results in soil temperatures that are too cold and skin temperatures and 2m temperatures that are too warm. The same qualitative behaviour can be observed at Cabauw (not shown). However, as will be discussed in Section 8, the influence of other factors, such as compensating errors resulting from the representation of vegetation in semi-arid areas and from small-scale variations in vegetation and soil type near measurement stations, mean that it is difficult to adjust the energy exchange in a way which leads to an

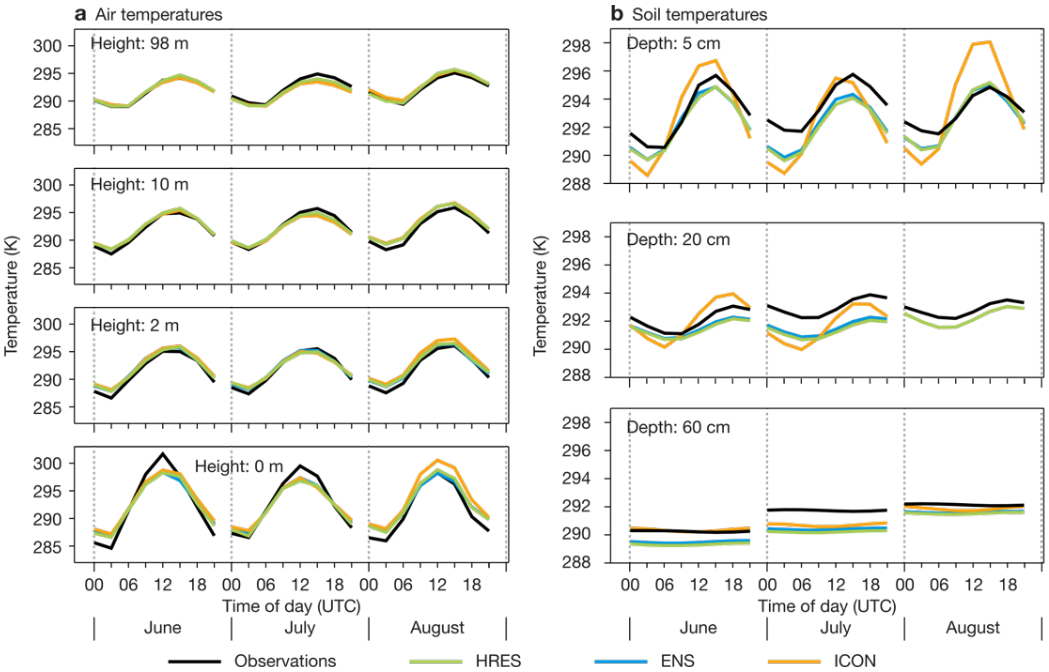

Figure 11: Monthly averaged diurnal cycles of temperatures at (a) different heights in the air and (b) different depths in the soil at Falkenberg for observations, HRES, ENS mean and ICON at forecast day 4 for the months of June, July and August 2017 (from Schmederer et al., 2019).

overall error reduction on the European scale. This was demonstrated through sensitivity experiments in which the strength of the atmosphere-land coupling coefficient was modified, as discussed in Schmederer et al. (2019) and Beljaars (2020).

5 Humidity errors

As for 2m temperature, 2m dew point biases vary geographically, as well as with season and time of day. In summer (Figure 12, top panels), there is a dry bias in Europe, especially during daytime. While the nighttime dry bias over Europe has varied in magnitude over the years, the daytime dry bias has been a robust feature of ECMWF forecasts since at least 2002 (Figure 3). Over Canada, the US (outside the Rockies), and Russia there is however a neutral to moist bias both at 00 and 12 UTC. In winter (Figure 12, bottom panels), the dominant pattern in the northern hemisphere is a neutral to moist bias in Canada, Scandinavia, and Russia, and a generally dry bias elsewhere. In the tropics, India and sub-Saharan West Africa stand out with a dry bias that varies in magnitude but is present throughout the year, both during day and night. To a lesser extent this is also true for Central America and parts of Brazil.

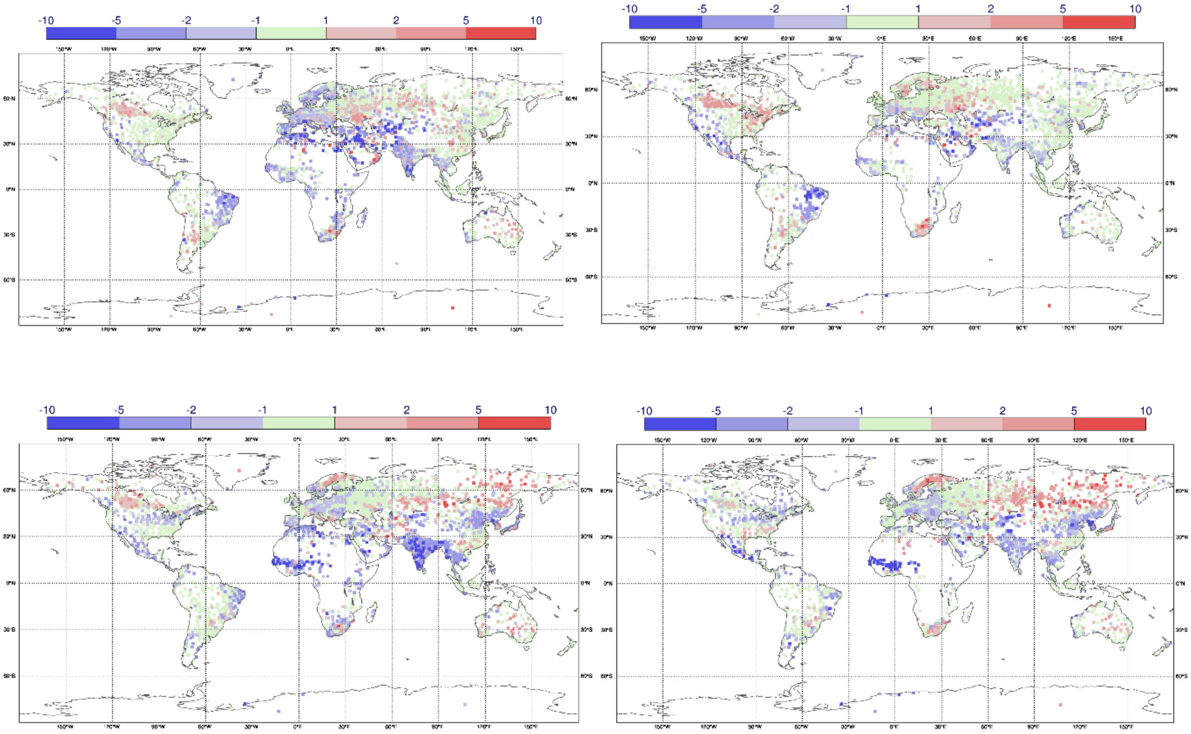

Figure 12: Bias of the 3-day HRES forecast of 2m dewpoint at 12 UTC (left) and at 00 UTC (right) in summer (JJA) 2018 (top panels) and in winter (DJF) 2018-19 (bottom panels). Included are only those stations at which the difference between model and actual station height is less than 100 m and where each of the four nearest grid points has >50% land fraction according to the land-sea mask of the model.

It should be noted that 2m dew point verification against surface observations is even more affected by representativeness issues than 2m temperature. Microclimatic effects, such as the type and amount of high vegetation around a site, the fraction of paved surfaces, and the presence of irrigation in the area

(Yang et al., 2020), can all influence humidity close to the surface. A more detailed analysis of representativeness is included in Section 7.

The dry bias in Europe (most noticeable during daytime in summer) is a large-scale feature (Figure 12) and evaluation against supersite observations suggests it extends into the lower part of the boundary layer (Figure 13). Comparison with radiosondes suggests the model underestimates the gradient in temperature and especially humidity across the lowest 200m of the atmosphere (Haiden et al., 2018; Beljaars, 2020). This underestimation is particularly pronounced at lower latitudes and contributes to the negative biases there. It means that part of the daytime cool/low humidity bias in summer is likely due to the surface layer in the model being too strongly mixed.

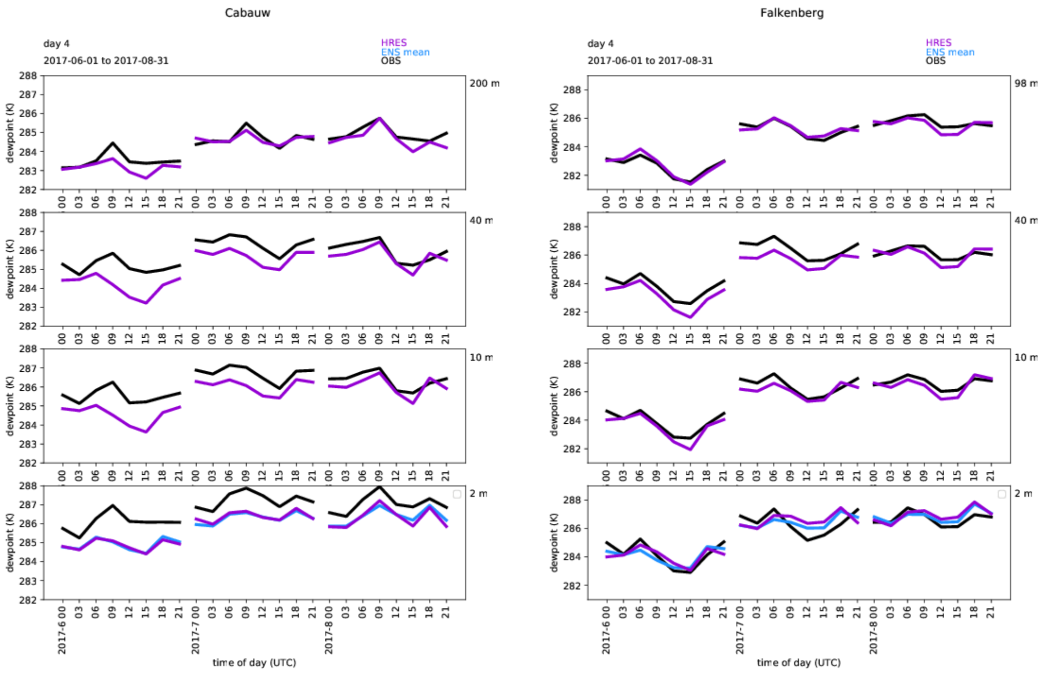

Figure 13: Monthly averaged diurnal cycles of dew point at different heights in the air for observations (black) and HRES (purple)_and ENS mean (blue) at forecast day 4 for the months of June, July and August 2017, at Cabauw and Falkenberg.

Sensitivity experiments discussed in Haiden et al. (2018) demonstrated that while 2m temperature is basically insensitive to choices made in the parametrization of turbulent mixing for a summer period, 2m dew point and the humidity amount in the boundary layer change significantly during daytime depending on the strength of the turbulent or convective mixing. A stratification of the 2m dew point biases according to cloud cover (both observed and forecasted) suggests that while in clear-sky conditions the IFS forecasts predominantly have little bias during the day, and a moist bias in the evening, in cloudy conditions the daytime bias is dry (Figure 14). This suggests that the mixing may be too strong in cloudy boundary layers.

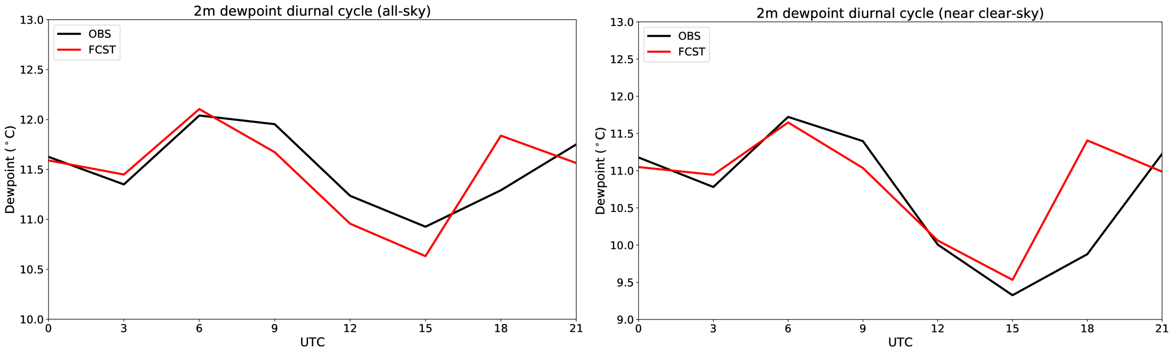

Figure 14: Diurnal cycle of 2m dewpoint in an area in Central Europe (48-53N, 6-14E) for all conditions (left panel), and for near clear-sky conditions (right panel) in JJA 2018. Near clear-sky conditions are defined by both the forecast and observations having a total cloud cover not exceeding one okta. Verification is against SYNOP stations.

Errors in the representation of evaporation, which is poorly constrained by observations at global, and even regional scale, as well as errors in soil moisture can also impact forecasts of near-surface humidity. In particular, it is known that spring evaporation is too high in the model, and in summer the vegetation gets into stress conditions too quickly. This was found to be partly due to the canopy resistance formulation and the linear treatment of soil moisture stress function as depicted in the comparison with the A-gs canopy resistance formulation in CTESSEL (Boussetta et al., 2013). Future work will also consider exploring better soil moisture stress formulation as tested in Ronda et al. (2001) and Verhoef et al. (2014).

Evaporation over bare soil is also problematic (Beljaars, 2020). Moreover, the combination of dry and cold biases during daytime summer (e.g. over Europe) suggests that these are driven by different, compensating, processes, as a cold bias should go hand in hand with a moist bias if they are simply due to an error in the turbulent energy flux partitioning. As discussed in Section 8, a revision of the vegetation classification in the model could help reduce errors in spring evaporation, while an increased discretization in the soil can increase the amplitude of the diurnal cycle in 2m temperature and partially eliminate the cold daytime bias in summer. However, as suggested previously, the dry bias may be partially associated with the representation of turbulent and convective mixing in cloudy situation.

These compensating biases need to be urgently addressed as they can affect the quality of the forecasts both directly through the model integration, and indirectly by interfering with the land data assimilation. Indeed, at present, the land data assimilation adds or removes soil moisture to minimize errors in near-surface temperature (de Rosnay et al., 2013; Fairbairn et al., 2019; Munoz Sabater et al., 2019). But removing soil moisture for example to address a cold bias in summer (through reduced evaporation), necessarily enhances the existing dry bias. A more in-depth discussion of the role of the soil moisture data assimilation and the interactions with these compensating model errors can be found in Beljaars (2020).

6 Wind errors

Both the mean and standard deviation of 10m wind speed error over Europe have decreased over time (Figure 3). In particular, a reduction of the 10m wind speed biases was achieved in 2011 through a retuning of the roughness lengths for momentum (Sandu et al., 2012). The representation of the 10m wind speed (during daytime) is controlled to a large extent by the (constant) values of the momentum roughness lengths associated with the (20) vegetation types considered in IFS. Given that it is difficult to determine the grid-scale roughness length from observations, the revision of the values used in the IFS was based on theoretical considerations and SYNOP observations of wind speed at 10m. The basic idea was to search, for each vegetation type, for a new value of the momentum roughness length for which the mean forecast error in 10 m wind speed (during daytime) with respect to SYNOP observations drops to zero. At the time, this calibration showed that the momentum roughness length values should be increased for nine and decreased for one of the eighteen vegetation types characterizing land areas.

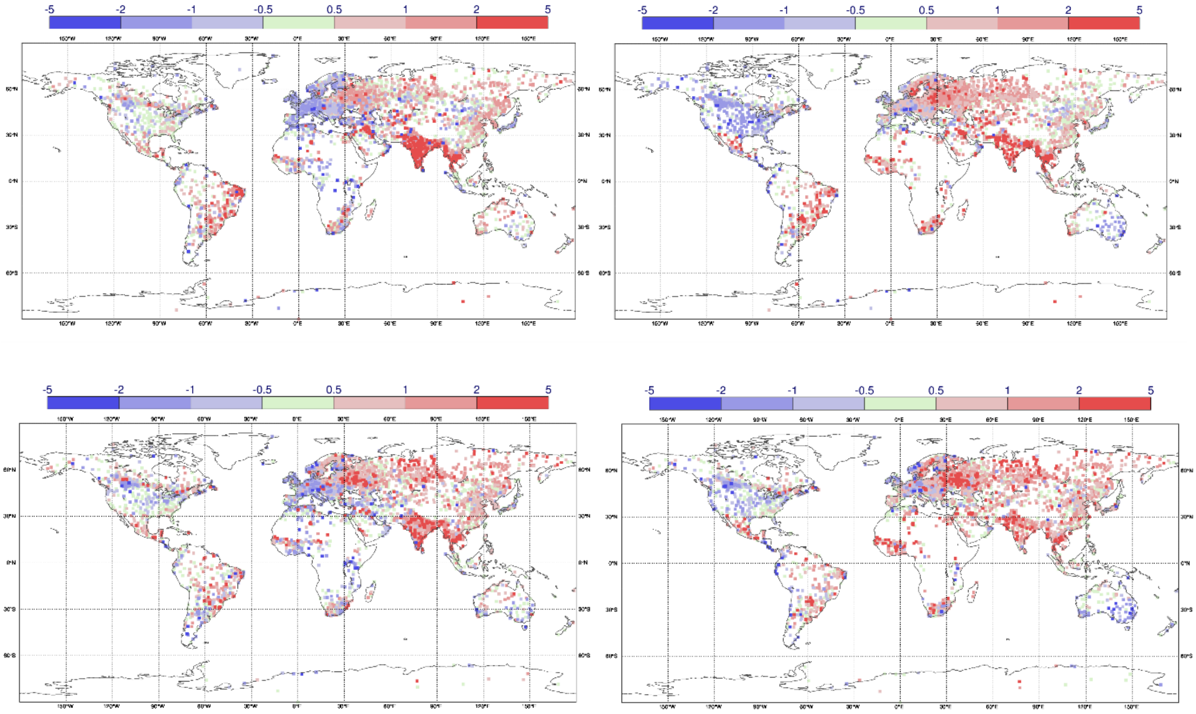

Figure 15: Bias of the 3-day HRES forecast of 10m wind speed at 12 UTC (left) and at 00 UTC (right) in summer (JJA) 2018 (top panels) and in winter 2018-19 (bottom panels). Included are only those stations at which the difference between model and actual station height is less than 100 m and where each of the four nearest grid points has >50% land fraction according to the land-sea mask of the model.

As the roughness lengths values for the different vegetation types are constant, calibrating them can only shift the daily mean value of the near-surface wind speed but cannot help eliminating issues in the diurnal cycle as can be seen from Figure 3. Furthermore, the extent to which such a calibration works depends on the quality of the underlying vegetation maps. Moreover, other processes such as the strength of turbulent or convective mixing of momentum may play a role as well in the near-surface wind speed biases. Figure 15 shows that indeed, despite the calibration of the roughness length made in 2011, 10m wind speed biases exhibit large geographical variations. In Europe there is a slow bias during the day, especially in summer, while some wind biases are present throughout the year during day and at night, such as the fast bias over Russia and, especially, India. The fact that Europe and India have mostly the same vegetation types (crops and interrupted forest, see Figure 11.8-Figure 11.11 IFS documentation

Part IV), but opposite 10m wind speed biases suggests that either the vegetation maps and more generally the representation of vegetation seasonality, are not accurate (see discussion in Section 8) or that other processes (e.g. turbulent or convective mixing, and the partition between the two, Sandu et al., 2020) play a role. It is possible that the wind speed biases over India may also be partly due to observing practices, and these are seen in other global models as well (Ingleby, 2015).

One effort within USURF consisted in constructing a more modern and automatic tool that allows to redo a calibration of roughness lengths for momentum for recent ECMWF forecasts, and which could be easily reapplied each time the vegetations maps change, for example. Applying this tool to operational HRES and ENS forecasts of 10m wind speed for one recent year (June 2018 to June 2019) showed that mean errors during daytime for the most frequent vegetation types are already quite small (Figure 16). The mean errors of 10m wind speed for each vegetation type are virtually identical for the ENS CTL and the HRES, suggesting that at least for the ‘simple’ conditions considered (no orography, no snow, rather uniform grid boxes with one vegetation type covering more than 75% of the grid box) there is no undesired resolution dependency. The fact that the mean errors for the most frequent vegetation types are small means that not much skill can be gained by retuning the roughness length for the current vegetation maps (except over India and South East Asia). This was demonstrated through sensitivity experiments with new values for the roughness length for momentum, derived from the analysis of 10m wind speed biases over the period June 2018 to May 2019 (not shown).

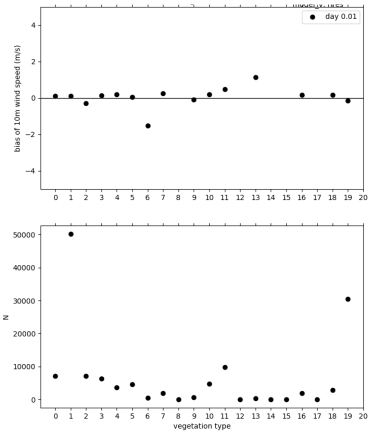

Figure 16: Stratification of 10m wind speed biases against SYNOP observations (top), as a function of the vegetation type globally for day 2 of HRES forecasts from June 2018 to May 2019. Only stations for which neighbouring model grid point is at altitudes lower than 500m, in snow free, daytime conditions, and the dominant vegetation type covers 75% of the grid box, are included (bottom). Vegetation types are as follows: 0 - bare soil; 1 - crops; 2 - short grass; 3 - evergreen needle-leaf trees ; 4 - deciduous needle-leaf trees ; 5 - deciduous broad-leaf trees; 6 - evergreen broad-leaf trees; 7 -tall grass; 8 - desert; 9 - tundra; 10 - irrigated crops ; 11 - semi-desert ; 12 - ice caps and glaciers ; 13 - bogs and marshes; 14 - inland water; 15 - ocean; 16 - - evergreen shrubs; 17 - deciduous shrubs; 18 - mixed forest/ woodland; 19 - interrupted forest; 20 - water and land mixtures. The plots look nearly identical for the ENS control.

As discussed in Section 7, in terms of representativeness, verifying 10m wind against SYNOP is probably even more problematic than 2m humidity. It is therefore very helpful to have tower observations like those from the Falkenberg or Cabauw supersites that extend up to 200 m and ~100 m respectively. Figure 17 suggests that during summer daytime, the wind speed is not only underestimated at 10m at these two sites, but throughout the lowest part of the boundary layer. Although the amount by which the wind speed is underestimated during daytime in summer varies from month to month (and with the considered lead time), the picture is roughly the same on average at the two towers for the summers 2017, 2018, 2019. During nighttime, there is a much better match between forecasts and observations at these two supersites, corroborating the analysis done in Sandu et al. (2014) after the revision of the diffusion in stable boundary layers. Indeed, decreasing the turbulent mixing in stable conditions in cycle 40R1 (Nov 2013) has led to an improvement of the representation of the wind speed at the low-level jet height during summer nights. During winter, differences in wind speed (profile) biases at Falkenberg and Cabauw between nighttime and daytime conditions are less pronounced (not shown), suggesting that the summer daytime bias is related to non-local mixing of momentum in convective boundary layers.

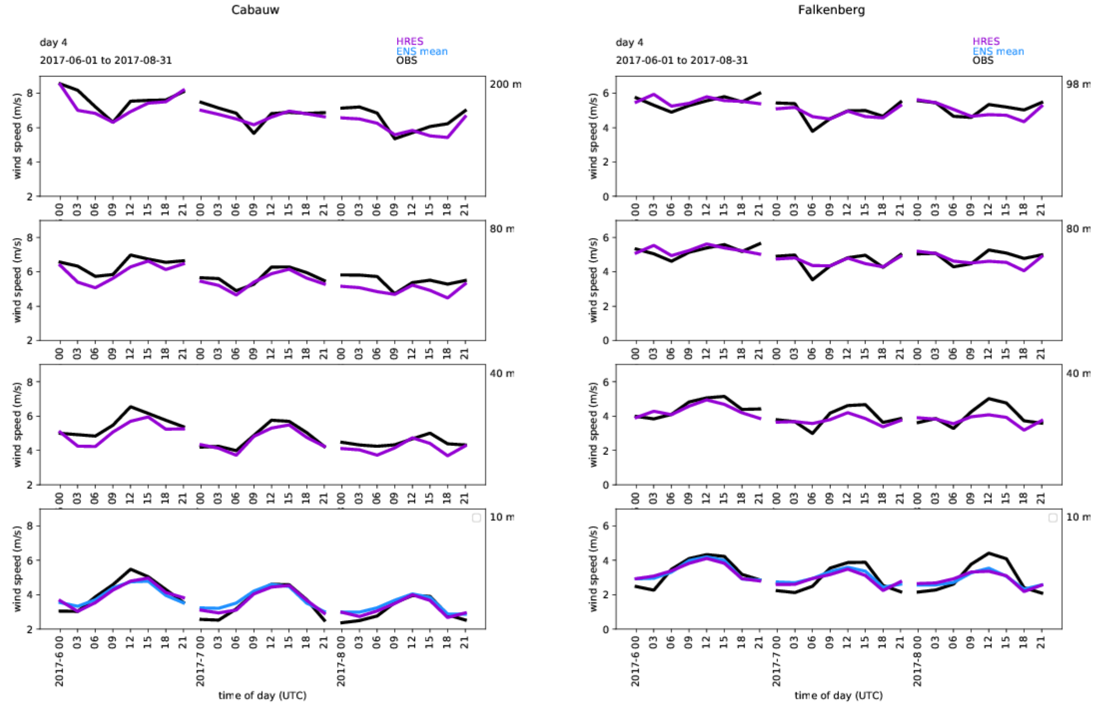

Figure 17: Monthly averaged diurnal cycles of wind speed at different heights in the air for observations (black) and HRES (purple) and ENS mean (blue) at forecast day 4 for the months of June, July and August 2017, at Cabauw and Falkenberg.

The reduction in turbulent mixing in stable conditions in November 2013 had also led to a reduction of wind direction biases at the surface, which can be seen in Figure 3 (bottom). Maintaining too strong mixing in stable boundary layers, which is typically done in global NWP models, is indeed known to lead to an underestimation of the wind turning throughout the boundary layer (Svensson et al., 2009), and consequently to errors in wind direction at the surface. However, biases in wind direction both over land (Figure 3, bottom) and over oceans (not shown) remain. As discussed in Sandu et al. (2020), the remaining biases over ocean (~3-5 deg) are mainly associated with unstable boundary layers and are very sensitive to the momentum transport by shallow convection and to the partitioning between turbulent and convective mixing in the convective boundary layer. This partitioning, and more generally certain aspects of momentum transport by shallow convection (Sagioratto et al., 2020; Schlemmer et al., 2016), are poorly constrained and could explain both biases in wind speed and wind direction in unstable boundary layers.

7 Representativeness and reliability

Verification of near-surface parameters from an NWP model against in-situ observations is affected by a varying amount of representativeness error (sometimes also called representativeness mismatch). Although supersites provide a larger range of co-located observations of different parameters than SYNOP stations, studies based on their data suffer from this issue as well, and an effort needs to be made to understand to what extent systematic errors at the supersite are representative of biases in a wider region. This has been done for example for Sodankyla (Finland) and Summit (Greenland) in Day et al. (2020). Representativeness issues need to be taken into account when drawing conclusions from verification results. They can for example affect the interpretation of the spread-skill relationship in the ENS for near-surface weather parameters (Saetra et al., 2004). Indeed, a major contribution to the apparent lack of spread at shorter forecast ranges can be due to representativeness. Figure 18 and Figure 19 show the spread-skill relationships at forecast day 4 for 2m temperature and dew point for a summer and winter period at Falkenberg, Cabauw and Sodankyla, extending the analysis done in Schmederer et al. (2019) for 2m temperature at Falkenberg. The raw model output (blue curves) is generally under-spread, and a simple bias correction tends to improve the spread-skill relationship somewhat (yellow). However, taking into account the representativeness error has a much larger effect towards increased reliability (purple blue curves). Figure 18 and Figure 19 also show that the spread-skill relationship is generally better in summer (panels in left column), and that the model is very far from capturing conditions in Sodankyla in winter, even when correcting for bias and representativeness. This corroborates previous findings that the ECMWF ensemble forecasts and data assimilations are generally underspread in polar regions, particularly in the lower troposphere and stratosphere (Bauer, et al., 2016; Lawrence et al., 2019). As discussed in Section 8.1 below, the multi-layer snow scheme considerably increases the spread and improves the reliability of the ensemble at high latitudes. It should also be noted, however, that the magnitude of the representativeness correction applied here is suitable as an average across Europe but does not represent the more extreme conditions present in northern Scandinavia in winter.

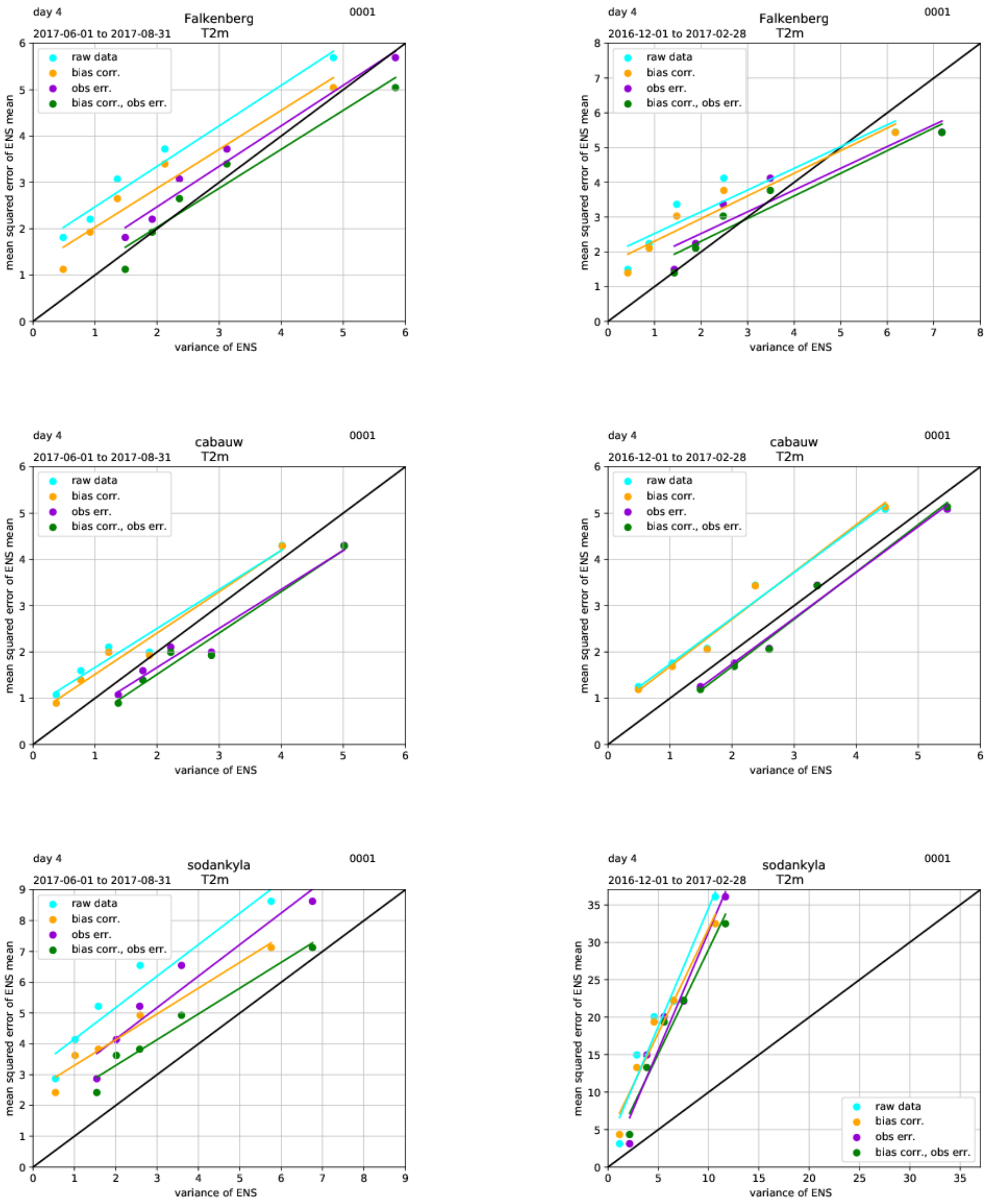

Figure 18: Reliability diagrams for ENS 2-metre temperature forecasts for various supersites at forecast day 4 in June, July and August 2017 (left) and December, January, February 2016/2017 (right). To create the charts, three-hourly data were grouped into five equally populated classes of increasing ensemble variance. The mean ensemble variance and the mean squared ensemble mean error were then computed for each class (i) with the raw data; (ii) with raw ensemble data but accounting for observation errors; (iii) with bias-corrected ensemble data but raw observations; and (iv) with bias-corrected ensemble data and accounting for observation errors.

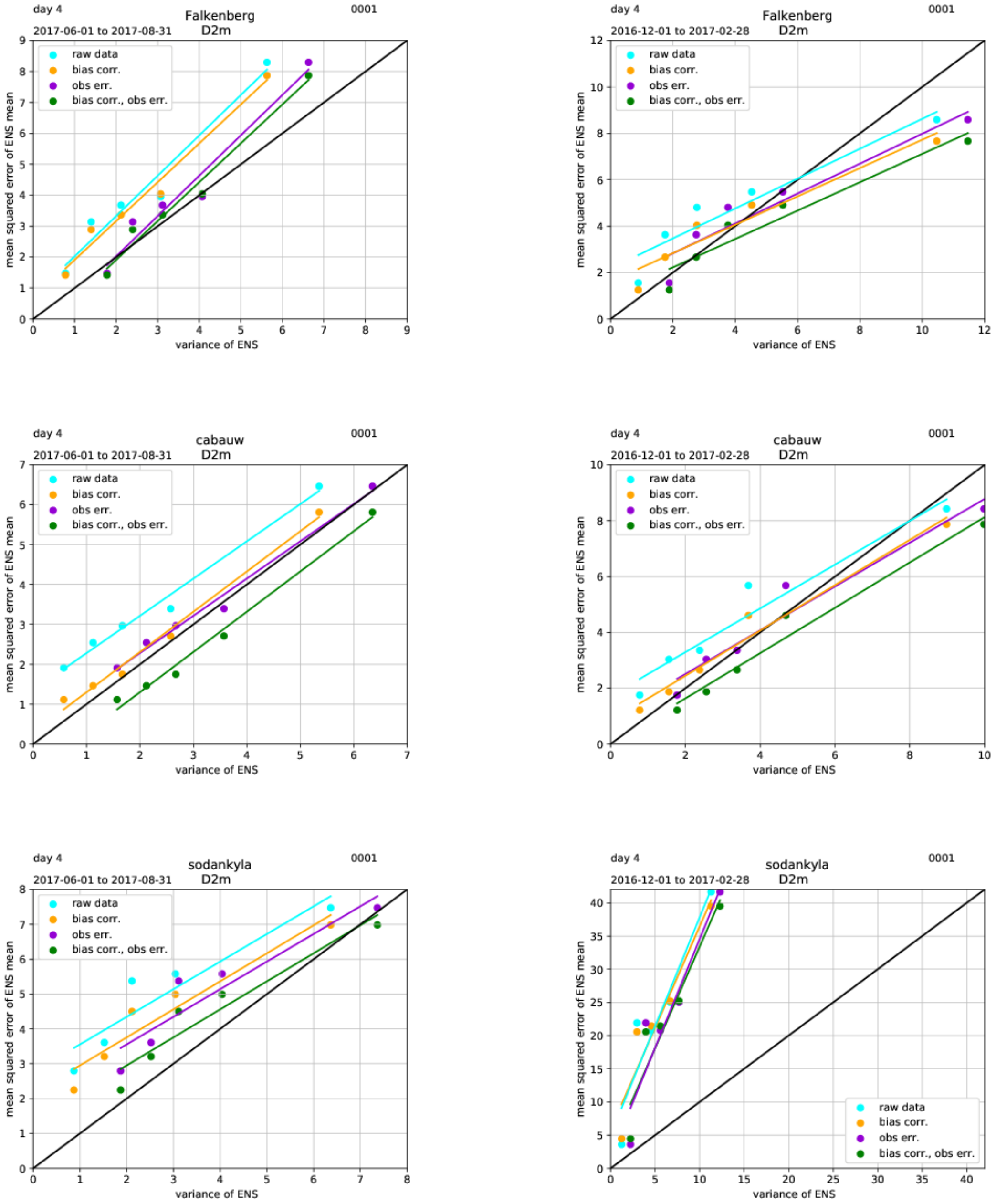

Figure 19: Reliability diagrams for ENS 2-metre dew point forecasts for various supersites at forecast day 4 in June, July and August 2017 (left) and December, January, February 2016/2017 (right). To create the charts, three-hourly data were grouped into five equally populated classes of increasing ensemble variance. The mean ensemble variance and the mean squared ensemble mean error were then computed for each class (i) with the raw data; (ii) with raw ensemble data but accounting for observation errors; (iii) with bias-corrected ensemble data but raw observations; and (iv) with bias-corrected ensemble data and accounting for observation errors.

A quantitative assessment of the T2m representativeness of locations like Lindenberg was attempted by Schmederer et al. (2019) using SYNOP stations. They computed the RMS difference between observed T2m values and averages of these observations over a radius of 20 km around each station for an area in Central Europe. The resulting RMS difference was on the order of 0.5-1 K, depending on season and time of day (Figure 20 left). A bit more than half of it was due to random errors, while the remaining part was due to bias. Representativeness explains nearly a third of the RMSE of the HRES against synop observations at forecast day 3 (which is about 2 K on average over Europe) and this estimate gives an indication of the magnitude of minimum error below which even a nearly perfect grid-box average forecast (at the given resolution) cannot get. This is also the case for 2m dew point (Figure 20 right). Interestingly, for dew point the representativeness is more of an issue during the day than at night, while for temperature it is rather the opposite. This points to the important role of sub-grid scale variations of evapotranspiration (e.g. due to vegetation cover inhomogeneities) in driving small-scale spatial variability of surface-layer humidity.

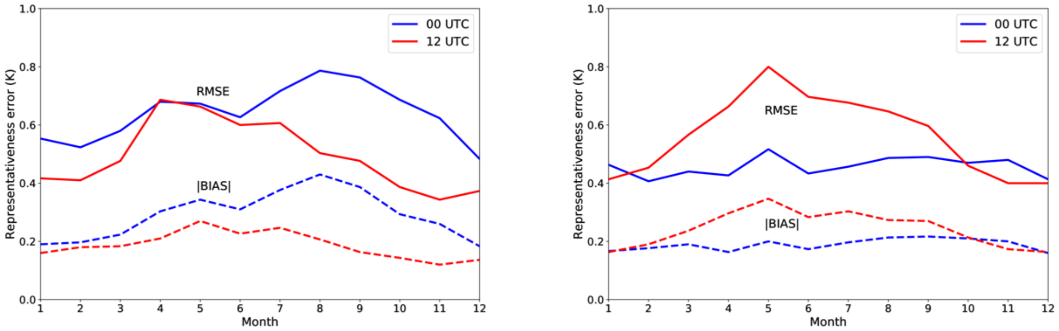

Figure 20: Estimation of 2m temperature (left) and dew point (right) representativeness error at 00 UTC and 12 UTC based on SYNOP observations in central Europe (48–55°N, 6–15°E) in the period 2016–2018. The chart shows the absolute value of the bias (|BIAS|) and the root-mean-square difference (RMSD) between the point observations and the mean observed value within 20x20 km boxes.

A more systematic investigation of the representativeness of surface parameters (temperature, wind speed, precipitation) has been carried out recently at ECMWF by Ben Bouallegue (2020). Using high-density surface observations which ECMWF collects from its Member and Cooperating States (Haiden and Duffy, 2016), parametric distributions were derived which relate the probability of values at a small scale (e.g. point observations) to the value at a larger scale (such as the model grid scale). For temperature, a normal distribution was applied, for wind speed a truncated normal distribution. Using a perturbed ensemble approach, noise sampled from these parametric distributions is added to the forecast. Ben Bouallegue (2020) was able to show that the apparent increase of 10m wind speed ENS forecast skill with increasing lead time in the short range was due to representativeness not being accounted for and disappears when it is taken into account (Figure 21). In the case of 2m temperature, the representativeness mismatch is less than for wind, but especially in the short range the resulting ENS verification gives significantly different values when the perturbations are added.

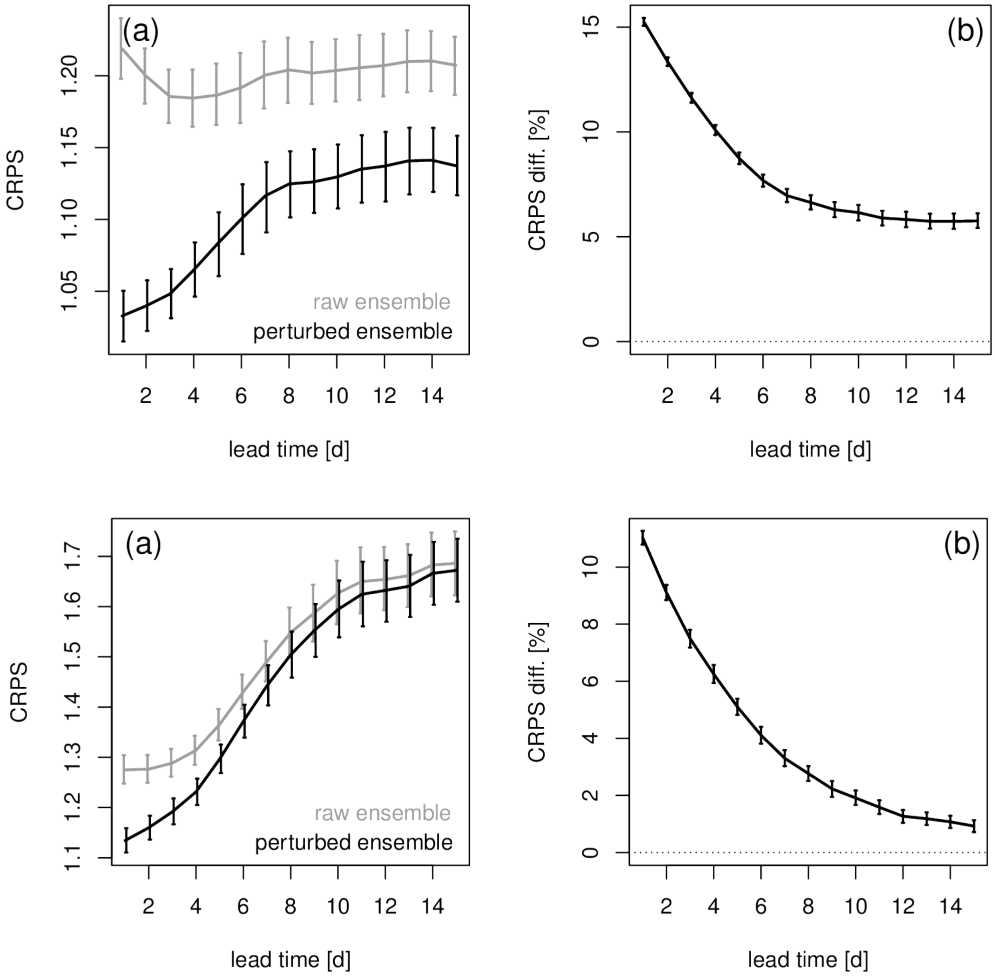

Figure 21: (a) CRPS computed with (black) and without (grey) accounting for representativeness uncertainty and (b) the corresponding CRPS relative difference (in %) as a function of the forecast lead time. Vertical bars indicate 95% confidence intervals. Results valid for 10 m wind speed ensemble forecasts (upper panels, CRPS units are ms^{-1}) and for 2m temperature ensemble forecasts (bottom panels, CRPS units are K), Europe, Summer 2018.

Even the parametric approach can only give a rough approximation to the true representativeness error, which varies with location, season, time of day, atmospheric stability, etc. However, for the purpose of verification it is a good approximation that captures much of the representativeness at stations which are located in relatively homogeneous terrain, i.e. away from mountains and coasts, and which are not affected too much by special microclimatological effects (e.g. cold air pooling due to the site being in a shallow basin). In April 2020, the methodology was fully implemented into ECMWF's operational verification system (quaver), which allows to routinely apply it for evaluation, e.g. in scorecards. A more comprehensive evaluation of flow-dependent sub-grid variability, representativeness and model bias is being explored as part of the ecPoint post-processing initiative (Pillosu and Hewson, 2017). Initially this is being applied to precipitation, but the methodology could be applied to other weather parameters in future.

8 Ongoing efforts to address systematic biases in near-surface weather parameters

The investigations made during USURF, and knowledge gathered through other research and evaluation work at ECMWF in recent years, has highlighted a number of aspects on which work is needed in order to improve forecasts of near-surface weather parameters.

For cold processes, a multi-layer snow scheme combined with a revision of snow albedo could reduce temperature biases in snow covered areas. A further reduction of the diffusion in stable boundary layers, which is still too strong in the IFS, could also help improve 2m temperatures in particular under very cold conditions at low wind speeds (Day et al., 2020; Koltzow et al., 2019). Finally, there is still room for improvement in low cloud forecasts and their radiative impacts in polar regions (Forbes and Ahlgrimm, 2014; Sotiropoulou et al., 2016) and improvements in the parametrization of mixed phase cloud processes could help to further reduce errors in cloud, radiation and near-surface temperatures.

For warm processes, a long due revision of the vegetation maps, for which new datasets have become available in recent years, and a revision of the vegetation seasonality could help address biases in all near-surface variables, through effects on evaporation, and land-atmospheric coupling strength for both momentum and heat. Furthermore, a better soil discretization (at present the soil is represented by 4 layers, the topmost one being 7cm deep), could help increase and thereby improve the diurnal cycle in temperature. A recent revision to the land surface albedo in IFS Cycle 47r1 involved implementing the solar zenith angle dependence of the MODIS albedo climatology described by Schaaf et al. (2002), and redefining the wavelength splitting the two albedo bands from 0.625 to 0.7 microns. This has had the effect of warming most land surfaces by up to 0.25 K, a slight improvement particularly in Africa and other tropical land areas where the IFS is systematically too cold.

Finally, work is also needed on the turbulent mixing in unstable layers and in particular on the partitioning between mixing by convective versus turbulent transport (cloudy versus clear boundary layers).

In parallel with, and partially motivated by, the investigations within USURF, efforts have been recently made to advance on, or investigate the potential benefits of, some of these aspects. In this section, we outline these efforts and the prospects of recent developments, before summarizing the key results from USURF and outlining the necessary steps for addressing the remaining biases in near-surface weather parameters in Section 9.

8.1 Development and evaluation of a multi-layer snow scheme

In the framework of the H2020 projects EartH2Observe and APPLICATE, Arduini et al. (2019) have developed a multi-layer snow scheme in the IFS, building on the work of Dutra et al. (2012), and evaluated its benefits in both offline and coupled (forecast only)[1] experiments. The offline evaluation at well- instrumented field sites from the Earth System Model Snow Intercomparison Project (ESM-SnowMIP, Krinner et al., 2018) demonstrated that the new scheme largely improves the representation of snow depth for most of the sites considered, reducing the root- mean- square error averaged over all

[1] experiments in which the forecasts are started from the operational analysis, that means experiments with no data assimilation cycling

sites by more than 30%. The improvements are due to a better description of snow density in thick and cold snowpacks, but also due to an improved representation of sporadic melting episodes. The evaluation of coupled 10- day weather forecasts against SYNOP observations shows an improved representation of snow depth at all lead times, demonstrating a positive impact at the global scale (Figure 7 of Arduini et al., 2019).

As expected, the multi-layer snow scheme also leads to a more realistic representation of the snow-atmosphere coupling. Day et al. (2020) used observations from the Sodankyla (Finland) and Summit (Greenland) supersites and process-based diagnostics building on the ideas of Miller et al. (2018) to examine how the surface energy budget terms change when the single-layer snow scheme of IFS is replaced with the multi-layer scheme. The response of the modeled ground and surface sensible heat fluxes to the radiative forcing (defined as the downwelling longwave plus the net shortwave radiative fluxes) becomes closer to that found for observations in the forecasts with the multi-layer snow scheme (Figures 5, 6 of Day et al, 2020), at both the Sodankyla and Summit supersites. Indeed, when using the single-layer snow scheme, the total fraction of the radiative forcing balanced by the turbulent fluxes and ground heat flux is too high at both sites. As a result, the fraction balanced by the upwelling longwave flux is too low, which is equivalent to the surface temperature response to radiative forcing being too low compared to observations. This is partly due to the coupling strength of the atmosphere to the land-surface being too strong at both sites (i.e. the fraction of the radiative forcing going into heating the land surface is too large) due to the large-thermal inertia of the deep single-layer snowpack. When the multi-layer snow scheme is used, the thin top snow layer effectively decouples the rest of the snowpack from the atmosphere. The fraction of the radiative forcing going into heating the ground decreases, and the surface and near-surface temperatures become more responsive to variations in the radiative forcing, as is the case in observations. This generally results in a lowering of the 2m temperature during nighttime (particularly during very cold conditions) and a warming during daytime, leading to an enhanced and thereby improved diurnal cycle, particularly during clear-sky conditions (Figure 11 of Arduini et al., 2019). The multi-layer snow scheme thus reduces, although it does not eliminate, the positive bias in the simulated minimum 2m temperature over parts of Scandinavia and most of Russia (Figure 22). It also reduces the cold daytime bias in most regions, except northern parts of Siberia which are characterized by a warm daytime bias (Figure 22 d, e).

Moreover, as showed in Day et al (2020) the multi-layer snow scheme is also reducing the Continuous Ranked Probability Score (CRPS) in ENS forecasts [2](see their Figure 2b) at all lead-times for the Arctic winter. The spread of the ENS increases considerably in the Arctic region (by up to 25% with respect to observations) and the spread-skill relationship improves both at Sodankyla (Figure 23 top), and over the Arctic region as a whole (Figure 23 bottom). Moreover, including the ML snow results in a ~10% reduction in the number of cases with CRPS>5K in the Arctic (from 23.5 to 21.3 %, not shown), which is a large improvement in skill. One can note that the fraction of cases in the extra-tropics with values of the CRPS>5K for 2m temperature at a lead time of 5 days is one of ECMWF's headline scores, as it is of major interest to forecast users.

[2] All the coupled forecast and ENS experiments described in Arduini et al. (2019) and Day et al. (2020) were performed at a lower resolution and with a reduced ENS configuration compared to the operational HRES and ENS forecasts (32 km and 8 ensemble members).

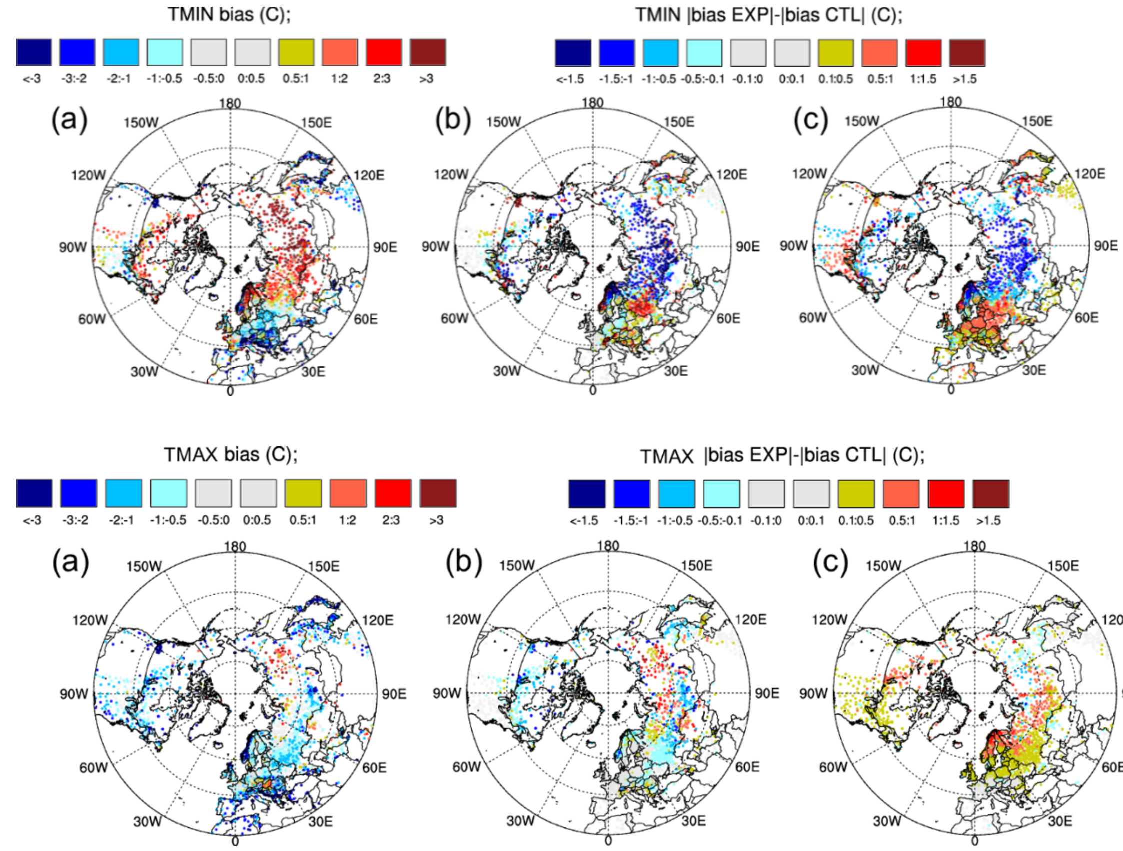

Figure 22: Top row: bias of the daily minimum 2- m temperature of the coupled forecasts performed for DJF 2016/2017 using the (a) control IFS which uses a single- layer snow scheme (CTL) and the change in the absolute value of this bias when using instead (b) a multi-layer snow scheme, (c) short instead of long tails in stable boundary layers. Bottom row: same as the top row, but for the daily maximum 2- m temperature. All plots are shown at a lead time of 3 days, warm colours in panels b-c, e-f indicating a deterioration of the forecast skill, while cold colours indicate an improvement.

However, in regions characterized by errors in the daily mean temperature or in the maximum daily temperature, caused by errors in cloud processes or surface albedo, a more ‘responsive’ snow scheme (to the radiative forcing) can lead to larger errors in 2m temperature. This is the case for example over Norway and east Russia for which the biases in the minimum and maximum temperature, respectively, increase when using the multi-layer snow scheme (see Figure 22 and the more in-depth discussion in Arduini et al., 2019). Another illustrative example is given in Figure 24, which shows the evolution of 2m temperature, snow cover and cloud cover at an alpine site (Payerne) for two sets of forecasts performed using the single-layer and the multi-layer snow schemes. This example shows the challenge associated with the simulation of the correct snowfall amount and low stratocumulus clouds in a wide Alpine valley. The snow amount is largely overestimated over this period, and between 19 and 25 January the model simulates mostly clear sky, while in reality clouds were present, resulting in a much larger diurnal cycle of 2m temperature compared to observations during this period. This is due to too much solar radiation reaching the surface during daytime, and an enhanced radiative cooling of the surface during night-time (with a lack of low-level clouds the downwelling longwave flux is smaller, and hence the net longwave loss at the surface is larger). The multi-layer snow forecasts show larger errors because the reduced thermal inertia of the thin top snow layer enables a larger response (and hence variation) of the surface temperature to the radiative forcing both during day and night-time. This results in colder temperatures at night and warmer temperatures at day, and thus in an even stronger diurnal cycle. This example illustrates that improving the physical realism of a particular process in the model (e.g. allowing stronger and more realistic night-time cooling) can enhance errors due to other processes (e.g. lack of low cloud and/or excessive amount of snowfall).

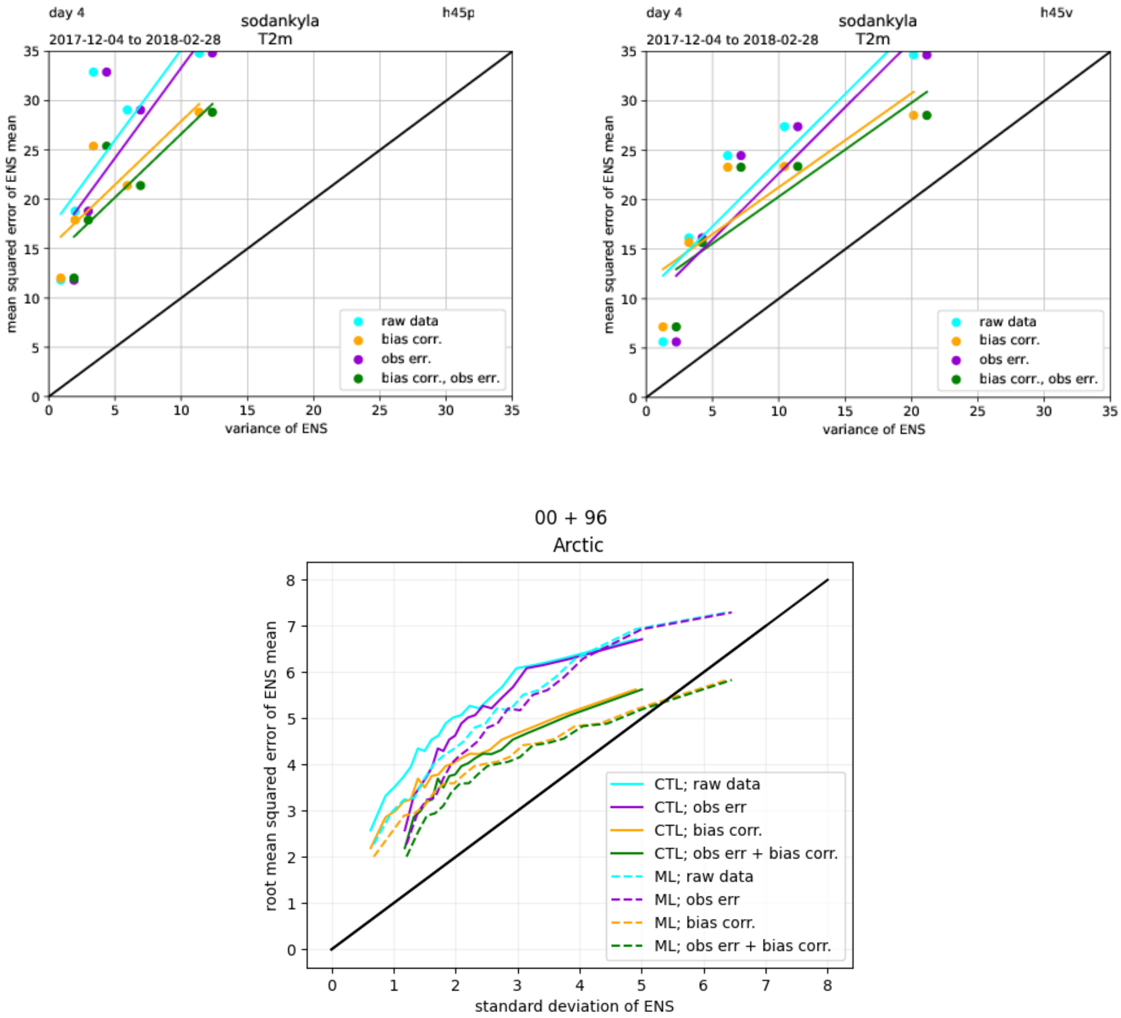

Figure 23: Top: Reliability diagrams for ENS 2-metre temperature against observations at Sodankyla, as in Fig. 18 but for the single-layer (left) and multi-layer snow (right) ENS experiments discussed in Day et al.,2020. Bottom: reliability diagrams for ENS 2-metre temperature against Synops for the entire Arctic region, at a lead time of 96 hours, for the control single-layer (full) and multi-layer snow (dashed) ENS experiments discussed in Day et al.,2020. The mean ensemble variance and the mean squared ensemble mean error were then computed for each class (i) with the raw data; (ii) with raw ensemble data but accounting for observation errors; (iii) with bias-corrected ensemble data but raw observations; and (iv) with bias-corrected ensemble data and accounting for observation errors.

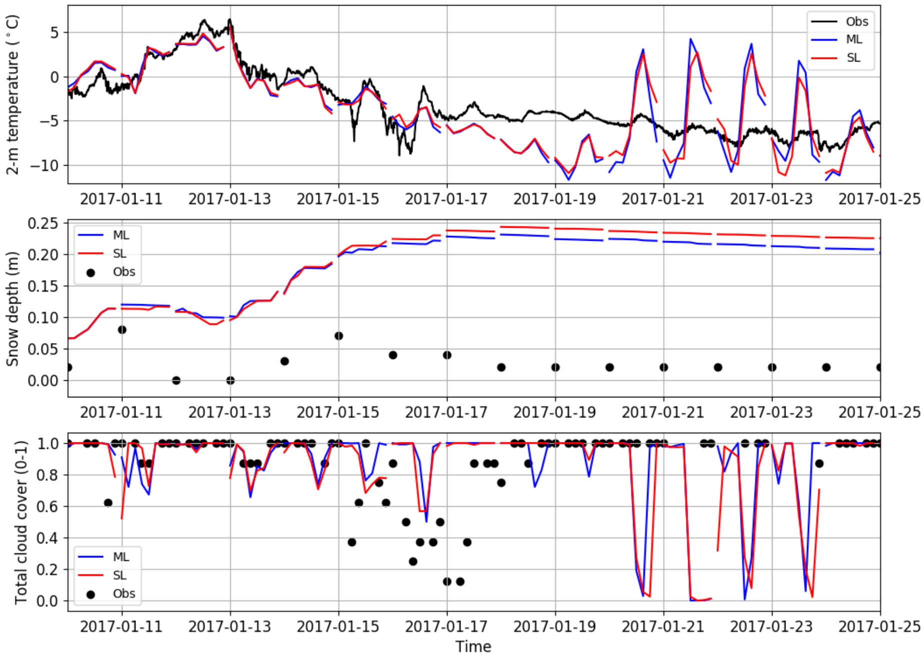

Figure 24: Evolution of 2m temperature, snow depth and total cloud cover at Payerne from observations (black), and from forecasts with a single layer (red) and multi-layer snow scheme (blue). Day 2 forecasts are concatenated for the illustrated period.

The developments of the multi-layer snow scheme have now reached a stage where modifications to the forecasting system are made so that forecasts can be run in the data assimilation configuration. Data assimilation experiments and their evaluation are underway. The careful evaluation in operational configuration for both the HRES and ENS is a pre-requisite prior to operational implementation.

The multi-layer snow model also opens possibilities to improve radiative transfer modelling approaches over snow covered surfaces. Preliminary investigations of multi-layer microwave emission modelling have been conducted in the IFS, showing promising performances of the multi-layer approach at low frequencies (Hirahara, 2020). This will be further developed in the next four years to enable coupled assimilation of satellite radiances over snow covered surfaces and extension of the all sky approach to all sky and all surfaces approach.

Further enhancement of the snow processes will include the representation of snow sublimation, the interaction with the forests affecting the overall albedo. These modifications are foreseen upgrades that will be examined in conjunction with the treatment of vegetation (see section 8.3) that is anticipated to change the distribution of forested areas, particularly over Europe.

8.2 Evaluation of the impact of reduced turbulent mixing in stable conditions

During USURF we also reassessed the relative contribution to the biases in near-surface temperature of other processes known to play a role during wintertime, e.g. the magnitude of turbulent mixing in stable conditions. It is widely acknowledged that global numerical weather prediction models, including the IFS, generally maintain too strong diffusion in stable conditions (Svensson et al., 2009; Sandu et al., 2013; Holtslag et al., 2013). The stability functions used to compute the turbulent exchange coefficients in stable conditions in IFS are the so-called 'long-tail' functions, which prescribe a much weaker decrease of the exchange coefficients with the Richardson number than the Monin-Obukhov stability functions (also known as 'short-tail' functions).

Sensitivity experiments were performed for several winter periods in which the turbulent mixing in stable boundary layers was reduced by using short-tail instead of long-tail stability functions. Day et al. (2020) have shown that such a model change would improve the response of the surface energy budget terms to radiative forcing at Sodankyla by reducing the coupling of the surface to the atmosphere which is too strong in such conditions. However, using short-tail instead of long-tail stability functions in stable boundary layers has mixed effects on the biases in the minimum and maximum 2m temperature (Figure 22c, f). Reducing the turbulent mixing in stable boundary layers results in a widespread cooling both during night and daytime (not shown), corroborating previous experimentation by Sandu et al. (2013). During night-time, this cooling, and the impact on 2m temperature biases is of the same order of magnitude as that obtained when replacing the single-layer with the multi-layer snow scheme (Figure 22b, c). An improvement can be seen in most regions, except east Europe/west Russia, where the 2m temperature was already too cold during nighttime. During day however, the existing cold bias is amplified, and the short-tail formulation leads to a widespread deterioration of the 2m temperature forecasts.

As discussed earlier, over flat terrain (e.g. over Central Europe) the existing cold bias is associated to a certain extent with cloud biases (Haiden et al., 2018), so further improvements to cloud fraction and cloud optical properties could perhaps allow in the future to revisit the choice of the stability functions and further reduce the amount of diffusion maintained in stable boundary layers. In mountainous areas, work would be needed to consider the effect of subgrid orography on the heat budget. At present, the effect of subgrid orography is only accounted for in the momentum budget through the orographic drag parametrizations. However, a higher resolution orography was also shown to affect the near-surface temperature by maintaining more low-level wind shear, and thereby more mixing, and warmer temperatures near the surface. As found in a previous investigation led by ECMWF (Sandu et al., 2014b), and more recently shown by Beljaars (2020), the lack of a parametrization for these effects is at the heart of the resolution dependency of near-surface temperatures over mountainous areas in IFS forecasts (e.g. between the HRES and ENS forecasts). Research on this subject in the community is at its infancy, but finding a solution will remain relevant, in particular at monthly and seasonal timescales.

Reducing the turbulent mixing in stable boundary layers is also detrimental for large-scale forecast skill, leading to a deterioration of about 3-4 % in the RMSE of geopotential height at 500hPa at a lead-time of 5 days (not shown). This corroborates previous results by Sandu et al. (2013) which suggested that such a change could not be introduced operationally, unless an overhaul of the partitioning between the different terms contributing to surface stress is made. This partitioning remains poorly constrained at the moment and it is one of the largest sources of uncertainty in global models (Sandu et al, 2019). However, efforts are underway in the framework of the WGNE/GASS project COORDE to constrain orographic drag using km-scale simulations (Van Niekerk et al., 2020). This should lead to constraints on the partitioning between the different drag processes, and hopefully allow us in the future to use either short tail stability functions in stable boundary layers which are theoretically consistent with Monin-Obukov theory, or a prognostic turbulent kinetic energy equation. The effect of either of these changes on ENS forecasts, in particular on the spread-skill relationship has not yet been assessed and will be the subject of future work.

8.3 Vegetation developments

Vegetation related improvements are being explored to address some of the systematic biases and representativeness issues highlighted in the framework of USURF. These developments focus on the use of more accurate and up-to-date vegetation Land Use/Land Cover (LU/LC) maps. These maps are based on the ESA-CCI/C3S LU/LC and allow removing “hybrid” vegetation types such as the interrupted forest (which covers about 25% of the land points in the current ECMWF HRES). It is important to note that having realistic vegetation types allows a better characterization of the model parameters, especially for parameters based on observed quantities. Preliminary results using these new vegetation maps in offline surface simulations show substantial impacts on the energy fluxes partition mostly driven by the increase of low vegetation cover at the expense of the high vegetation cover. Figure 25 illustrates the impact of the new ESA-CCI LU/LC maps on the sensible and latent heat fluxes for the month of July 2017. The sensible/latent heat fluxes become substantially slower/larger in the areas in which the low vegetation cover increases and the high vegetation decreases (not shown) compared to the operational vegetation maps. This obviously has an impact on the surface skin temperature and subsequently the near surface atmosphere. In a recent collaborative study with the University of Lisbon, it was demonstrated that using these new vegetation maps improves the model skin temperature and its diurnal cycle over Iberia, with respect to satellite observations of skin temperature provided by the EUMETSAT-Land Surface Analysis-Satellite Application Facility (Nogueira et al., 2020).

Additional developments are focused on improving the model phenology and its seasonality. A revision of the Leaf Area Index (LAI) observation operator was evaluated within USURF. The current LAI operator tends to overestimate the LAI during the transition periods (spring and autumn). The revised LAI operator disaggregates the observed LAI into high and low vegetation components, and thereby allows to derive a LAI climatology which is closer to the observed LAI climatology, in particular in the transition seasons. The new operator also paves the way for an operational assimilation of the LAI which would allow better monitoring of the surface conditions inter-annual variability and representation of extremes as demonstrated in Boussetta et al. (2015).

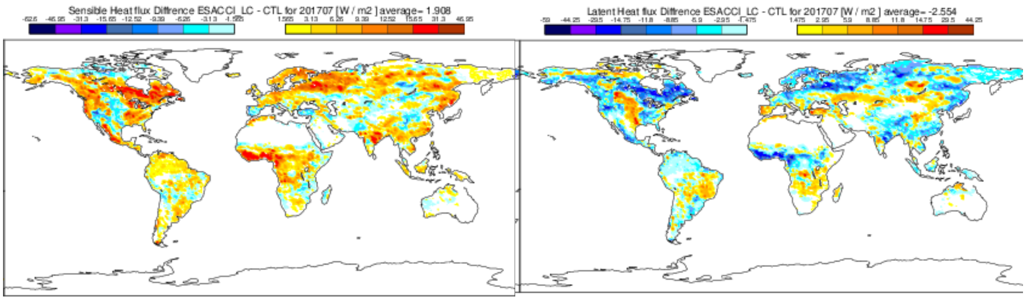

Figure 25: Differences in sensible (left) and latent (right) heat flux (W/m2) for July 2017 between offline simulations using ESA-CCI LU/LC and the current operational setup. The IFS flux convention is used here, where downward/upward fluxes are negatively/positively defined.

Figure 26 shows that the operational LAI overestimates the LAI by 50 to 200% over SW Russia in April 2019, compared to the revised LAI derived with the new disaggregation operator. Using the revised LAI (cyan line, Figure 27) allows to overcome most of the 2m temperature cold bias and 2m dew point wet bias, which is typically found with respect to synop observations (black line, Figure 27) for ECMWF forecasts over SW Russia during spring. The 2m temperature and humidity errors are further reduced when the moisture deficit stress function used at present for high vegetation in the Jarvis formulation of the canopy resistance is also activated for low vegetation (yellow line Figure 27). These results corroborate the findings of other recent studies (Liu et al., 2020; Lansu et al., 2020), but a more thorough evaluation is required to assess the impacts of this change at global scale.

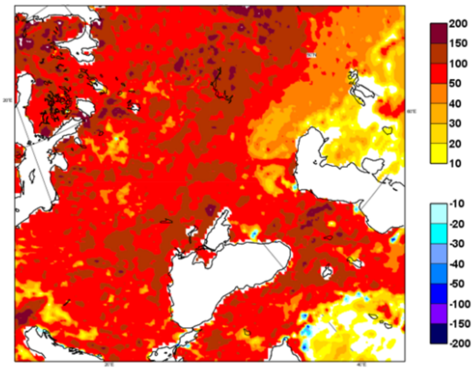

Figure 26: Relative differences (%) between the operational Leaf-Area-Index (LAI) and the LAI obtained with a conservative disaggregation.

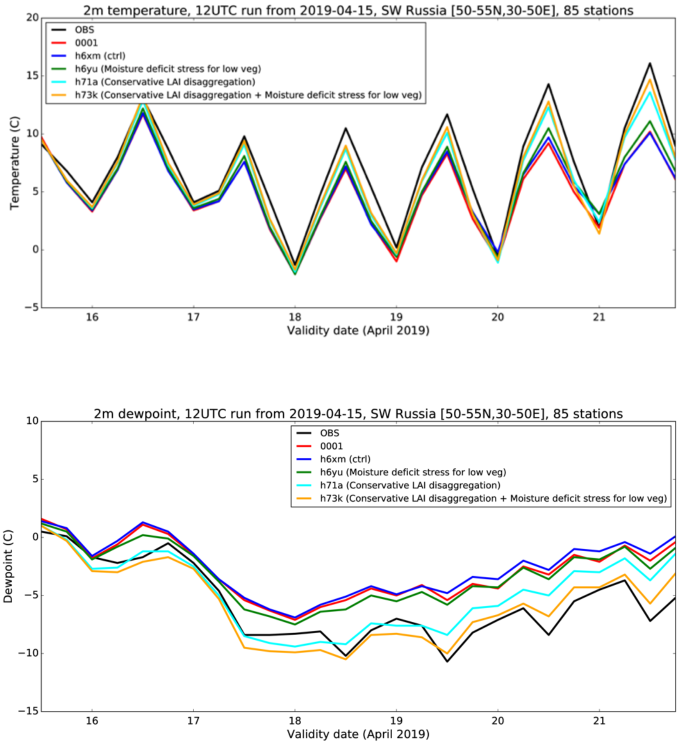

Figure 27: Impact of conservative LAI climatology on 2m temperature (upper) and 2m dewpoint (lower): black: observation, red: operational ECMWF HRES, blue: control experiment - HRES like configuration but at lower resolution (approximately 30km), green: moisture deficit stress for low vegetation; cyan: conservative LAI and yellow Conservative LAI + moisture deficit stress for low vegetation.

In terms of improved surface condition seasonality, a vegetation cover seasonality parametrization is also being investigated. This parametrization links the vegetation cover with the LAI through the clumping index. This change goes in the right physical direction; however, it has a substantial atmospheric impact given its link to both energy and momentum fluxes as it changes the tile fractions and would also add a layer of the LAI uncertainties to the wind errors.

Preliminary coupled experiments show that at the global scale, these changes to the vegetation, maps or seasonality do not always have positive impacts and would need to be accompanied by re-tuning/optimization of some parameters, and more specifically minimum stomatal resistance, roughness lengths for heat and momentum and skin conductivity. Given the complexity of the tuning/optimization procedure, one efficient way is to group these changes and stratify the errors by vegetation types and climate regions (Boussetta et al., 2012) and then use in-situ observations where available such as the FLUXNET energy and carbon fluxes network and the ISMN International soil moisture network, as recently used to evaluate ERA5 and ERA5-Land (Beck et al., 2020). It is worth noting that addressing the land surface processes and its coupling with the atmosphere is tightly dependent on a proper identification of the atmospheric related compensating errors as was depicted earlier. Vilà- Guerau et al. (2020) also suggest that a proper behavior in the IFS of the CTESSEL canopy closure over the Amazon region was affected by a misrepresentation of the heat and moisture transport from the sub-cloud into the cloud layer as compared to local observations and LES simulations.

8.4 Moist physics revision

Work on a major revision of the moist physics has been ongoing over the past 5 years. This has now reached a stage where the revised moist physics package is planned for operational implementation in IFS Cy48R1. This package comprises a large number of changes to the turbulence, convection and cloud schemes, aiming to: (i) address a number of long-standing issues in the formulation and interaction of parametrized convection, turbulent mixing and cloud-related processes; (ii) improve the physical representation of the convective boundary layer, deep convection, cloud and precipitation in the forecast and (iii) reduce large-scale systematic errors in cloud and radiation. This revision of the moist physics is also an important and vital step towards ECMWF strategic target of moving towards high resolution (~5km) ensemble forecasts, down from 18 km today. The planned moist physics changes and their main impacts on cloud, radiation and precipitation processes are described in Bechtold et al. (2020). Final revisions and comprehensive testing of this upgrade across resolutions and timescales, from analysis increments to medium-range, extended-range and seasonal forecasts, will continue, in readiness for operational implementation.

9 Conclusions and way forward

Systematic errors in forecasts of near-surface weather parameters are due to a multitude of processes and their interactions. Disentangling error sources involves (sometimes painstaking and time-consuming) detective work. An attempt to characterize the biases in temperature, humidity and winds and to identify their possible causes was made in the USURF project, which lasted from 2017 to 2020. It was shown that, although near-surface forecasts have gradually improved in the past decades, systematic biases with often complicated spatio-temporal patterns remain. These biases were found to be to a large extent consistent between the HRES, ENS control and ENS mean, and USURF brought answers to some longstanding questions regarding their causes. We were able to demonstrate that:

- a cold bias over southern Europe in winter is mostly related to an underestimation of the low cloud cover amount (Haiden et al., 2018).
- a warm bias in north Europe, and more generally at high latitudes in the north hemisphere during winter is related to the representation of snow (Haiden et al., 2018, Arduini et al., 2019; Beljaars, 2020) and diffusion in stable boundary layers (Day et al., 2020).
- the underestimation of the amplitude of the diurnal cycle in temperature during summer is in part associated with an overestimation of the land-atmosphere coupling strength (Schmederer et al., 2019).
- temperature and dew point biases (in particular during transition seasons) are in part related to the representation of vegetation (in terms of cover and seasonality, Boussetta et al., 2015) and evaporation over bare soil (Beljaars, 2020).
- the dry summer bias is also in part related to the representation of turbulent mixing in particular in cloudy convective cases. It could also be related to some extent to remaining biases in the diurnal cycle of convection (Bechtold et al., 2013).
- the remaining biases in wind profiles in the boundary layer and wind direction at the surface are in part related to the representation of mixing in convective boundary layers, and in particular with the partition of momentum transport between dry and moist updrafts (Sandu et al., 2020).
- it is important to explicitly take into account observation uncertainty in the verification process in order to be able to truly assess the reliability of the ENS using near-surface point observations, which are not necessarily representative of the grid-box mean (Ben Bouallegue, 2020).

Improved interface modelling (e.g. atmosphere–land, atmosphere–wave–ocean–ice) to improve the skill of predictions of near-surface weather parameters especially during the spring and autumn transition seasons, is one of the foci of the recently approved ECMWF 2021-2030 strategy. Although, as highlighted above, efforts have been made in recent years to address some of the systematic biases in near-surface forecasts (also in the context of USURF), a much more concerted effort is necessary in order to address the remaining issues and significantly improve the skill of near-surface forecasts. Hereafter we highlight some of the avenues and next steps that should be taken to achieve this.

A number of issues with the representation of different processes have been identified, as well as potential solutions and, as described above, work has started on several aspects. A non-exhaustive list of these includes:

a) Finalize the testing and implementation of the multi-layer snow scheme developed in APPLICATE (EU H2020 project) in the first cycle in Bologna, which would help address both snow and temperature biases in snow covered regions, in particular at high latitudes.
b) Upgrade the Land-Sea-Mask to use the latest Global Surface Water Explorer dataset (Pekel et al., 2016) and associated water body bathymetry (Choulga et al., 2019), since water-land contrast is the largest contributor to surface energy balance variations.
c) Upgrade the Land-Use (LU) and Land-Cover (LC) with the latest available datasets. Current research focuses on the ESA-CCI dataset available also in the Copernicus Climate Change Service’s Climate Data Store, covering several decades (enabling land-use-change in upcoming reanalysis, such as ERA6) and contextually recalibrate the roughness length for momentum and heat wind speed on new LU/LC data sets to minimize wind-bias. This

development is also crucial to the Carbon cycle modelling (see SAC paper on CO_2) and other Environmental applications of the IFS.

d) Revise the vegetation-state observation operator to disaggregate conservatively the Leaf-Area-Index, retuning the canopy resistance and the rooting depth as function of the upgraded LU/LC. This will particularly help improving temperature and dew point representation during transition seasons.

e) Revise the thermal land-atmosphere coupling to address the diurnal cycle biases and temporal shifts (Albergel et al., 2015) in the 2m temperature, skin and soil temperature that are currently due to a too strong coupling.

f) Revise the 2m temperature and relative humidity post-processing, replacing dominant low vegetation tile by averaged/dominant tile post-processing, avoiding grid-average aggregation of roughness lengths and remove the stability limiters (e.g. Richardson-number limit) in the surface transfer coefficients.

g) Enhance the vertical discretization of the soil to improve the temperature gradients and infiltration in case of intense precipitation (Mueller et al., 2016), with an overall increase of the soil depth to better satisfy the no-flux bottom boundary condition.

h) Enhance the vertical discretization in land ice to better represent temperature gradients within ice and utilize the ice parametrization on partial and resolved glacier points (to avoid unphysical snow-depth settings over glaciers and better take into account sub-grid glaciers) to activate the use of multi-layer snow on glaciers.

i) Revise the snow-forest albedo to utilize a weighted contribution of the multi-spectral snow-free albedo and the open-snow albedo to increase the realism of the seasonal cycle (bypassing use of fixed land-use dependent look-up tables, as set in Dutra et al., 2010)

j) Further assess the partition of momentum mixing between cloudy and clear conditions, using Large-Eddy Simulations and observations together using TU Delft, in order to improve the wind representation in convective boundary layers.

In order to implement some of the major developments listed above (i.e. the revised land-use, land-water and vegetation maps and vegetation seasonality) a tuning/calibration exercise for all poorly constrained parameters (e.g. roughness lengths, land-atmosphere coupling strength, canopy properties, hydrological properties) is necessary. Using the new maps with current parameter values (tuned for the present maps) would prevent extracting the maximum advantage from these developments in terms of forecast skill particularly in the transition seasons. A way forward is to aim towards the implementation of combinations of changes (e.g. b) to f) above in a first step, and g) to j) in the longer term), as is currently being done for the forthcoming moist physics upgrade, rather than for individual changes. This approach is motivated by the fact that parametrization improvements need to be tested together as processes affecting near-surface weather parameters have compensating effects. For example, the momentum and thermodynamics biases are intrinsically connected, e.g. when wind-mixing is larger in the model than reality in calm situations which results in a smaller temperature diurnal cycle than observed. Changing land-use and retuning the roughness parameters can thus both enhance wind-prediction capability, with foreseen benefits for wind-energy users, and lead to better near-surface temperature forecasts. Similarly, processes affecting the water cycle such as land evaporation, are tightly connected with vertical transport and cloudiness. At present excessive evaporation is needed to compensate for too strong atmospheric mixing, particularly in cloudy conditions.

Increasing realism in the representation of some processes (i.e. snow, soil) can potentially lead to increased random forecast errors, by exposing biases due to other processes (e.g. clouds). However, this is not necessarily an issue in the ensemble forecast and associated probabilistic verification, as was shown for the multi-layer snow scheme development. Benefits in terms of CRPS skill can counterbalance increases in RMSE. For example, increasing the diurnal cycle amplitude or getting closer to representing cold/hot extremes, will inevitably enhance the effects of radiative-forcing errors on temperature errors at the surface (thus increasing RMSE), but will also lead to an increased spread, and in some cases can thereby lead to an improved CRPS and improved spread-skill relationship and thus be beneficial for the reliability of the forecasts.

The skill of near-surface weather parameters forecasts can also be increased through land data assimilation developments, as it has been demonstrated by Fairbairn et al. (2019), Munoz-Sabater et al. (2019) and de Rosnay et al. (2014, 2015). Recent results, which will be further investigated in the future, have for example shown that the assimilation of snow over mountainous regions (and in particular Himalayas) can have an effect not only at local scale but also on circulation (de Rosnay et al., 2020; Orsolini et al., 2019). Current developments of land data assimilation capabilities in the offline model suite will enable further investigations of the relative roles of model and assimilation coupling on model biases. They will support consistent land surface modelling and assimilation research developments.

A testing strategy will involve a complementary use of well-instrumented sites (e.g. supersites/observatories, FLUXNET sites, ESM-SnowMIP sites) and observing networks (e.g. SYNOP, ISMN, the International Soil Moisture Network; GHCN, the Global Historic Climatology Network), which proved very valuable in the past for example for assessing the quality of soil moisture forecasts (Albergel et al., 2012; Fairbairn et al., 2019). Additionally, a greater use will be made of satellite observations informative of the surface layer (see Balsamo et al., 2018 for a review), as well as a hierarchy of tests from single-column site simulations to nudged and fully coupled experiments. Moreover, an increased coupling of land surface and river-basin hydrological modelling will allow to validate model developments, within an Earth system benchmarking (as illustrated in Clark et al, 2015) by using water cycle information via the runoff (river discharge). The interplay between surface-hydrology and river-discharge at ECMWF has been shown for snow processes in Zsoter et al. (2019), for which a systematic error in snow-depth/snow density leads to erroneous river discharges in several northern latitude river-basins. Improved handling of snow density in Arduini et al. (2019) is foreseen to result in improved river-discharge accuracy.

Advances in the surface-atmosphere interactions and hydrometeorological scheme at increasingly high resolutions towards kilometric scales are not only relevant to ECMWF but equally to all Member and Cooperating States, and the list above offers a number of topics for collaboration.

Finally, it should be stressed that USURF has provided further evidence that increases in near-surface forecast skill in a complex Earth System model with its interrelated physical parametrizations critically depend on the availability of comprehensive observations and in-depth studies using process-based diagnostics that can correctly attribute model error. For example, for near-surface weather forecasts it will be very valuable to consolidate and extend the supersite evaluation methodology, as well as the tools for conditional verification. A more comprehensive conditional assessment of systematic temperature errors as a function of cloud properties at observational supersites, particularly mixed-phase boundary layer cloud, will help to further identify deficiencies in cloud processes which also impact the quality of near-surface weather forecasts. Ongoing improvements to the diagnostic and verification tools at ECMWF are therefore an important contribution towards further enhancements of forecast skill.

Acknowledgements

Supersite observations from Lindenberg (Germany, DWD), Cabauw (Netherlands, KNMI), Sodankyla (Finland, FMI), as well as support from H2020 projects such as Earth2Observe, APPLICATE and Copernicus (EFAS/GloFAS) has been crucial for the work undertaken in USURF and more generally for developments on the ECMWF forecasting system. This work also greatly benefited from collaborations with many colleagues from our Member States and beyond, and in particular Emanuel Dutra (IPMA/U. Lisbon), Clement Albergel (CNRS/Météo-France), Louise Nuijens (TU Delft), Gunilla Svensson and Michael Tjernström (MISU), Frank Beyrich and Claudia Becker (DWD), Marion Mittermaier, Annelize Van Niekerk and Simon Vosper (UKMO) and Barbara Casati (EC). We would also like to thank David Richardson, Patricia de Rosnay, Nils Wedi, Richard Forbes, Peter Bechtold, Andy Brown, Florian Pappenberger and Florence Rabier, for highlighting open points and making useful suggestions in the review of this manuscript.

Key ECMWF publications triggered by, or relevant to, USURF

Arduini, G., Balsamo, G., Dutra, E., Day, J.J., Sandu, I., Boussetta, S. and Haiden, T., 2019: Impact of a multi- layer snow scheme on near- surface weather forecasts, *Journal of Advances in Modeling Earth Systems*, 11, 4687– 4710. https://doi.org/10.1029/2019MS001725.

Beljaars, A., 2020, Near-surface temperature and humidity biases, ECMWF Tech. Memo. 870, 37 pp.

Ben Bouallegue, Z., 2020: Accounting for representativeness in the verification of ensemble forecasts. ECMWF Tech. Memo. 865, 26pp.

Day, J. J. et al, 2020: Measuring the impact of a new snow model using surface energy budget process relationships, *Journal of Advances in Modeling Earth Systems,* in review. Pre-print at: https://doi.org/10.1002/essoar.10502951.1.

Haiden, T., I. Sandu, G. Balsamo, G. Arduini and A. Beljaars, 2018: Addressing biases in near-surface forecasts. ECMWF Newsletter No. 157, 20-25.

Sandu, I., Bechtold, P., Nuijens, L., Beljaars, A. and Brown, A., 2020: Systematic biases in surface wind direction in ECMWF Integrated Forecasting System, ECMWF Tech. Memo. 866, 21 pp.

Schmederer, P., I. Sandu, T. Haiden, A. Beljaars, M. Leutbecher and C. Becker, 2019: Use of super-site observations to evaluate near-surface temperature forecasts. ECMWF Newsletter No. 161, 32-38.

References

Albergel C., P. de Rosnay, G. Balsamo, L. Isaksen and J. Muñoz Sabater, Soil moisture analyses at ECMWF: evaluation using global ground-based in situ observations, Journal of Hydrometeorology, 13, 1442-1460, 2012 http://dx.doi.org/10.1175/JHM-D-11-0107.1

Albergel, C., Dutra, E., Muñoz- Sabater, J., Haiden, T., Balsamo, G., Beljaars, A., Isaksen, L., de Rosnay, P., Sandu, I. and Wedi, N., 2015: Soil temperature at ECMWF: An assessment using ground- based observations. J. Geophys. Res. Atmos., 120: 1361– 1373. doi: 10.1002/2014JD022505

Arduini, G., Balsamo, G., Dutra, E., Day, J. J., Sandu, I., Boussetta, S. and Haiden, T., 2019: Impact of a multi- layer snow scheme on near- surface weather forecasts, Journal of Advances in Modeling Earth Systems, 11, 4687– 4710. https://doi.org/10.1029/2019MS001725.

Bauer, P., A. Thorpe and G. Brunet, 2015: The quiet revolution of numerical weather prediction. Nature, 525, 47-55.

Bauer, P., Magnusson, L., Thépaut, J.- N. and Hamill, T.M., 2016: Aspects of ECMWF model performance in polar areas. Q.J.R. Meteorol. Soc., 142: 583-596. doi:10.1002/qj.2449.

Balsamo, G.; Agusti-Panareda, A.; Albergel, C.; Arduini, G.; Beljaars, A.; Bidlot, J.; Blyth, E.; Bousserez, N.; Boussetta, S.; Brown, A.; Buizza, R.; Buontempo, C.; Chevallier, F.; Choulga, M.; Cloke, H.; Cronin, M.F.; Dahoui, M.; De Rosnay, P.; Dirmeyer, P.A.; Drusch, M.; Dutra, E.; Ek, M.B.; Gentine, P.; Hewitt, H.; Keeley, S.P.; Kerr, Y.; Kumar, S.; Lupu, C.; Mahfouf, J.-F.; McNorton, J.; Mecklenburg, S.; Mogensen, K.; Muñoz-Sabater, J.; Orth, R.; Rabier, F.; Reichle, R.; Ruston, B.; Pappenberger, F.; Sandu, I.; Seneviratne, S.I.; Tietsche, S.; Trigo, I.F.; Uijlenhoet, R.; Wedi, N.; Woolway, R.I. and Zeng, X., 2018: Satellite and In Situ Observations for Advancing Global Earth Surface Modelling: A Review. Remote Sens. 2018, 10, 2038.

H. E. Beck, M. Pan, D. G. Miralles, R. H. Reichle, W. A. Dorigo, S. Hahn, W. Wagner, J. Sheffield, L. Karthikeyan, G. Balsamo, R.M. Parinussa, N. Vergopolan and E.F. Wood, 2020: Evaluation of 18 satellite- and model-based soil moisture products using in situ measurements, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-184, in review.

Bechtold, P, R. Forbes, I. Sandu, S. Lang, M. Ahlgrimm, A major moist physics upgrade for the IFS, ECMWF Newsletter, Summer 2020.

Beljaars, A.C. and F.C. Bosveld, 1997: Cabauw Data for the Validation of Land Surface Parametrization Schemes. J. Climate, 10, 1172–1193, https://doi.org/10.1175/1520-0442(1997)010.

Beljaars, A., 2020: Near-surface temperature and humidity biases, ECMWF Tech. Memo. 870, 37pp.

Ben Bouallegue, Z., 2020: Accounting for representativeness in the verification of ensemble forecasts. ECMWF Tech. Memo. 865, 26 pp.

Boussetta, S., G. Balsamo, E. Dutra, A. Beljaars and C. Albergel, 2015: Assimilation of surface albedo and vegetation states from satellite observations and their impact on numerical weather prediction. Rem. Sens. Env.,163, 15, 111–126."http://dx.doi.org/10.1016/j.rse.2015.03.009".

Boussetta, S., G. Balsamo, A. Beljaars, A. Agusti-Panareda, J.-C. Calvet, C. Jacobs, B. van den Hurk, P. Viterbo, S. Lafont, E. Dutra, L. Jarlan, M. Balzarolo, D. Papale and G. van der Werf, 2013: Natural land carbon dioxide exchanges in the ECMWF Integrated Forecasting System: Implementation and offline validation, J. Geophys. Res., 118, 5923–5946, DOI: 10.1002/jgrd.50488.

Boussetta, S., G. Balsamo, A. Beljaars, T. Kral, L. Jarlan, 2012: Impact of a satellite-derived leaf area index monthly climatology in a global numerical weather prediction model. Int. J. of Rem. Sens., Vol. 34, Iss. 9-10, 3520-3542. DOI:10.1080/01431161.2012.716543.

Clark, M.P., Fan, Y., Lawrence, D.M., Adam, J.C., Bolster, D., Gochis, D.J., Hooper, R.P., Kumar, M., Leung, L.R., Mackay, D.S., Maxwell, R.M., Shen, C., Swenson, S.C. and Zeng, X., 2015:

Improving the representation of hydrologic processes in Earth System Models, Water Resour. Res., 51, 5929– 5956, doi:10.1002/2015WR017096.

Choulga, M., Kourzeneva, E., Balsamo, G., Boussetta, S. and Wedi, N., 2019: Upgraded global mapping information for earth system modelling: an application to surface water depth at the ECMWF, Hydrol. Earth Syst. Sci., 23, 4051–4076, https://doi.org/10.5194/hess-23-4051-2019.

Day, J.J. et al, 2020: Measuring the impact of a new snow model using surface energy budget process relationships, Journal of Advances in Modeling Earth Systems, 12, e2020MS002144. https://doi.org/10.1029/2020MS002144

Dutra, E., Balsamo, G., Viterbo, P., Miranda, P.M., Beljaars, A., Schär, C. and Elder, K., 2010: An improved snow scheme for the ECMWF land surface model: Description and offline validation. Journal of Hydrometeorology, 11(4), 899– 916.

Dutra, E., Viterbo, P., Miranda, P.M. and Balsamo, G., 2012: Complexity of snow schemes in a climate model and its impact on surface energy and hydrology. Journal of Hydrometeorology, 13(2), 521–538.

Fairbairn, D., P. de Rosnay and P. Browne: "The new stand-alone surface analysis at ECMWF: Implications for land-atmosphere DA coupling", Journal of Hydrometeorology 20, 2023-2042, 2019 https://doi.org/10.1175/JHM-D-19-0074.1

Forbes, R.M. and Ahlgrimm, M., 2014: On the representation of high- latitude boundary layer mixed-phase cloud in the ECMWF global model. Mon. Wea. Rev. 142: 3425–3445, doi: 10.1175/MWR-D-13-00325.1.

Haiden, T., M.J. Rodwell, D.S. Richardson, A. Okagaki, T. Robinson and T. Hewson, 2012: Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. Mon. Wea. Rev., 140, 2720-2733.

Haiden, T. and S. Duffy, 2016: Use of high-density observations in precipitation verification. ECMWF Newsletter No. 147, 20-25.

Haiden, T., I. Sandu, G. Balsamo, G. Arduini and A. Beljaars, 2018: Addressing biases in near-surface forecasts. ECMWF Newsletter No. 157, 20-25.

Haiden, T., M. Janousek, F. Vitart, L. Ferranti and F. Prates, 2019: Evaluation of ECMWF forecasts, including the 2019 upgrade. ECMWF Tech. Memo., 853, 54p.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz- Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.- N., 2020: The ERA5 Global Reanalysis. Q J R Meteorol Soc, 146: 1999– 2049. doi:10.1002/qj.3803.

Hirahara, Y., P. de Rosnay and G. Arduini: Evaluation of a Microwave Emissivity Module for Snow Covered Area with CMEM in the ECMWF Integrated Forecasting System, to be submitted in June 2020, Remote Sensing 2020

Holtslag, A.A.M., Svensson, G., Baas, P., Basu, S., Beare, B., Beljaars, A.C.M. et al, 2013: Stable atmospheric boundary layers and diurnal cycles: challenges for weather and climate models, Bull. Am. Meteorol. Soc. 94, 1691–1706. doi: 10.1175/BAMS-D-11-00187.1

Køltzow, M., Andreas Ø., Casati, B., Haiden, T., Bazile, E. and Valkonen, T.M., 2019: An NWP Model Intercomparison of Surface Weather Parameters in the European Arctic during the Year of Polar Prediction Special Observing Period Northern Hemisphere 1, Weather and forecasting,Volume 34.(4) s. 959-983, DOI: 10.1175/WAF-D-19-0003.

Ingleby, B., 2015: Global assimilation of air temperature, humidity, wind, and pressure from surface stations. Q.J.R. Meteorol. Soc., 141, 504-517.

Lansu, E.M., van Heerwaarden, C.C., Stegehuis, A.I. and Teuling, A.J., 2020: Atmospheric aridity and apparent soil moisture drought in European forest during heat waves. Geophysical Research Letters, 47, e2020GL087091. https://doi.org/10.1029/2020GL087091

Lawrence, H, Goddard, J, Sandu, I, Bormann, N, Bauer, P and Magnusson, L., 2019: An Assessment of the use of observations in the Arctic at ECMWF, ECMWF Tech. Memo., 845

Leutbecher, M., Lock, S.- J., Ollinaho, P., Lang, S.T.K., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H.M., Diamantakis, M., Dutra, E., English, S., Fisher, M., Forbes, R.M., Goddard, J., Haiden, T., Hogan, R.J., Juricke, S., Lawrence, H., MacLeod, D., Magnusson, L., Malardel, S., Massart, S., Sandu, I., Smolarkiewicz, P.K., Subramanian, A., Vitart, F., Wedi, N. and Weisheimer, A., 2017: Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. Q.J.R. Meteorol. Soc, 143: 2315-2339. doi:10.1002/qj.3094.

Liu, Y., Kumar, M., Katul, G.G. et al. Plant hydraulics accentuates the effect of atmospheric moisture stress on transpiration. Nat. Clim. Chang, 2020: https://doi.org/10.1038/s41558-020-0781-5.

Morcrette, C. J., Van Weverberg, K., Ma, H.- Y., Ahlgrimm, M., Bazile, E., Berg, L. K. et al., 2018: Introduction to CAUSES: Description of weather and climate models and their near- surface temperature errors in 5 day hindcasts near the Southern Great Plains. Journal of Geophysical Research: Atmospheres, 123, 2655– 2683. https://doi.org/10.1002/2017JD027199.

Muñoz Sabater J., H. Lawrence, C. Albergel, P. de Rosnay, L. Isaksen, S. Mecklenburg, Y. Kerr and M. Drusch. 2019: "Assimilation of SMOS Brightness Temperatures in the ECMWF Integrated Forecasting System", 145, issue 723, pp. 2524-2548, QJRMS, https://doi.org/10.1002/qj.3577

Mueller, A., Dutra, E., Cloke, H., Verhoef, A., Balsamo, G. and Pappenberger, F., 2016: Water infiltration and redistribution in Land Surface Models. ECMWF Tech. Memo. 791. doi: 10.21957/ppksejqu9.

Nogueira, M., Albergel, C., Boussetta, S., Johannsen, F., Trigo, I. F., Ermida, S.L., Martins, J.P.A. and Dutra, E.: Role of vegetation in representing land surface temperature in the CHTESSEL (CY45R1) and SURFEX-ISBA (v8.1) land surface models: a case study over Iberia, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-49https://doi.org/10.5194/gmd-2020-49.

Orsolini Y., M. Wegmann, E. Dutra, B. Liu, G. Balsamo, K. Yang, P. de Rosnay, C. Zhu, W. Wang and R. Senan: "Evaluation of snow depth and snow-cover over the Tibetan Plateau in global reanalyses using in-situ and satellite remote sensing observations", 13, 2221–2239, The Cryosphere, 2019, https://doi.org/10.5194/tc-13-2221-2019

Pekel, J.-F., Cottam, A., Gorelick, N. et al, 2016: High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422 https://doi.org/10.1038/nature20584.

Pillosu, F and T Hewson, 2017: New point-rainfall products for flash-flood prediction, ECMWF Newsltter, 153.

Reynolds, C., Williams, K. and Zadra A., 2019: WGNE Systematic Error Survey Results Summary, https://www.wcrp-limate.org/JSC40/12.7b.%20WGNE_Systematic_Error_Survey_Results_20190211.pdf

Rodwell, M. J., D. S. Richardson, T. D. Hewson and T. Haiden, 2010: A new equitable score suitable for verifying precipitation in NWP. Q. J. R. Meteorol. Soc., 136, 1344-1363.

Ronda, R. J., H.A.R. de Bruin and A.A.M. Holtslag, 2001: Representation of the canopy conductance in modeling the surface energy budget for low vegetation. J. Appl. Meteor., 40, 1431–1444.

de Rosnay P., G. Balsamo, C. Albergel J. Muñoz-Sabater and L. Isaksen: Initialisation of land surface variables for Numerical Weather Prediction, Surveys in Geophysics, 35(3), pp 607-621, 2014http://dx.doi.org/10.1007/s10712-012-9207-x

de Rosnay P., Isaksen L. and Dahoui M.: Snow data assimilation at ECMWF, ECMWF Newsletter no 143, article pp 26-31, Spring 2015, doi: 10.21957/lkpxq6x5

de Rosnay P., G. Balsamo, Y. Orsolini, E. Dutra, B. Liu, R. Senan, W. Wang, M. Wegmann, K. Yang and C. Zhu, invited keynote presentation: "Impact Of Snow Cover Data Assimilation Over The Tibetan Plateau On Medium Range Numerical Weather Prediction", EARSeL, Bern, 3-5 February 2020

Saetra, Ø., H. Hersbach, J. Bidlot and D.S. Richardson, 2004: Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability. Mon. Wea. Rev., 132, 1487–1501, https://doi.org/10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Saggiorato, B., Nuijens, L., Siebesma, A.P., de Roode, S., Sandu, I. and Papritz, L., 2020: The influence of convective momentum transport and vertical wind shear on the evolution of a cold air outbreak. Journal of Advances in Modeling Earth Systems, 12, e2019MS001991. https://doi.org/10.1029/2019MS001991

Sandu, I., Beljaars, A., Balsamo, G. and A. Ghelli, 2012: Revision of the surface roughness length table, ECMWF newsletter, No. 130, 8-9

Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T. and Balsamo, G., 2013: Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models? J. Adv. Model. Earth Syst., 5, 117– 133, doi:10.1002/jame.20013.

Sandu, I., Beljaars, A., Bechtold, P. and Balsamo, G., 2014: Improving the representation of stable boundary layers, ECMWF Newsletter, No.138, 24-29.

Sandu, I, Wedi,N. , Bozzo, A., Bechtold, P, Beljaars, A. and Leutbecher, M., 2014b, On the near surface temperature differences between HRES and CTL ENS forecasts, ECWMF Technical Report, RD14-315

Sandu, I., van Niekerk, A., Shepherd, T.G. et al, 2019: Impacts of orography on large-scale atmospheric circulation, Perpective, npj Clim Atmos Sci 2, 10, https://doi.org/10.1038/s41612-019-0065-9

Sandu, I., Bechtold, P., Nuijens, L., Beljaars, A. and Brown, A., 2020: Systematic biases in surface wind direction in ECMWF Integrated Forecasting System, ECMWF Tech. Memo. 866, 21pp.

Schaaf, C.B., Gao, F., Strahler, A.H., Lucht, W., Li, X., Tsang, T., Strugnell, N.C., Zhang, X., Jin, Y., Muller, J.P. and Lewis, P., 2002: First operational BRDF, albedo nadir reflectance products from MODIS. Remote sensing of Environment, 83(1-2), pp.135-148.

Schlemmer, L., Bechtold, P., Sandu, I. and Ahlgrimm, M., 2017: Uncertainties related to the representation of momentum transport in shallow convection, J. Adv. Model. Earth Syst., 9, 1269–1291, doi:10.1002/2017MS000915.

Svensson, G. and A. Holtslag, 2009: Analysis of model results for the turning of the wind and the related momentum fluxes and depth of the stable boundary layer. Boundary-Layer Meteorol.,132, 261-277.

Schmederer, P., I. Sandu, T. Haiden, A. Beljaars, M. Leutbecher and C. Becker, 2019: Use of super-site observations to evaluate near-surface temperature forecasts. ECMWF Newsletter No. 161, 32-38.

Sotiropoulou, G., Sedlar, J., Forbes, R. and Tjernström, M., 2016: Summer Arctic clouds in the ECMWF forecast model: an evaluation of cloud parametrization schemes. Q.J.R. Meteorol. Soc., 142, 387-400, doi:10.1002/qj.2658

Van Niekerk, A., Sandu, I., Zadra, A. et al, 2020: COnstraining ORographic Drag Effects (COORDE): a model comparison of resolved and parametrized orographic drag, J. Adv. Model. Earth Syst., in review.

Verhoef A. and Egea, G., 2014: Modeling plant transpiration under limited soil water: comparison of different plant and soil hydraulic parametrizations and preliminary implications for their use in land surface models. Agricultural and Forest Meteorology, 191. pp. 22-32. ISSN 0168-1923. Agricultural and Forest Meteorology, 191. pp. 22-32. ISSN 0168-1923.

Vilà- Guerau de Arellano, J., Wang, X., Pedruzo- Bagazgoitia, X., Sikma, M., Agusti- Panareda, A., Boussetta, S. et al, 2020: Interactions between the Amazonian rainforest and cumuli clouds: A large- eddy simulation, high- resolution ECMWF and observational intercomparison study. Journal of Advances in Modeling Earth Systems, 12, e2019MS001828. https://doi.org/10.1029/2019MS001828.

Yang Q., X. Huang and Q. Tang, 2020: Irrigation cooling effect on land surface temperature across China based on satellite observations, Science of The Total Environment, Volume 705,135984, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2019.135984