Process-oriented verification

Thomas Haiden, Barbara Casati, Caio Coelho, Eric Gilleland, Raghavendra Ashrit, Manfred Dorninger, and Chiara Marsigli

Joint Working Group for Forecast Verification Research, 28 March 2019

1. What is process-oriented verification?

Verification of forecasts from numerical weather prediction (NWP) models serves a wide range of purposes, including modelling-oriented forecast evaluation, performance monitoring for documentation and management, and user-oriented evaluation of final products. The main goal of modelling-oriented verification is to better understand, and ultimately reduce, model and data assimilation issues and errors in order to improve forecast quality. To achieve this goal, the verification methodology needs to be designed such that it allows for identifying the role of specific model processes in the occurrence of forecast errors. Because this approach can include a range of methodologies we define process-oriented verification, here, not in terms of specific techniques (although examples are given below) but rather by its overall objective of improving process understanding. In this context, 'process' can mean a subset of Earth-system physics represented in models by a specific parameterisation (e.g. vertical turbulent fluxes, or the conversion of cloud water to precipitation), but it can also mean an atmospheric phenomenon that involves a wide range of sub-processes, such as a tropical cyclone.

While process-oriented verification has always been an integral part of NWP model development, some of it could be adopted to become part of operational NWP verification suites. Apart from generating verification results that can be acted upon more directly by model developers, it would provide additional insights for forecast users. Ideally, it would contribute to a more efficient research-to-operations and operations-to-research cycle in NWP, both within and between NWP centres. This report discusses methodologies that are already being used, but could perhaps be used more widely and systematically, in process-oriented verification of outputs from various NWP centres. It starts with a brief description of model intercomparison activities (Section 2) and community tools (Section 3), touches on the benefits of multiple observational datasets and the need to up-scale observations (Section 4), and continues with conditional verification including the possible role of machine learning in informing model development (Section 5). The report concludes with some remarks about the use of supersite data (Section 6) and gives a few recommendations in the conclusion (Section 7).

This report does not represent an exhaustive and objective summary but rather illustrates issues and methodologies by drawing on the subjective experience and personal involvement of its authors in relevant ongoing activities. Thus it provides a subjective view which, we hope, nevertheless contributes to further discussions within NWP about the topic of process-oriented verification.

2. Model inter-comparison

In the context of process-oriented verification, comparing forecasts from different models is an attempt to relate differences in error characteristics between forecasting systems to differences in their design. It can provide additional insights when extended beyond a mere ranking exercise, to include in-depth analyses of state-dependent score differences and statistics. In addition to a comparison of forecast performance this should include comparisons of model climatologies, energy spectra (Skamarock et al., 2014), and other diagnostics, thus constituting a natural and direct connection to verification activities in the climate modelling community.

2.1 Coupled Model Intercomparison Project (CMIP)

One of the most widely known examples of a model inter-comparison activity is the Coupled Model Intercomparison Project (CMIP) organized under the auspices of the World Climate Research Programme's (WCRP) Working Group on Coupled Modelling (WGCM). It brings together coupled climate models from institutions worldwide and constitutes an important part of the scientific basis for the Assessment Reports issued by the Intergovernmental Panel for Climate Change (IPCC). It is currently in its 6th phase (CMIP6), and the produced simulations and scientific achievements are expected to support the IPCC Sixth Assessment Report (AR6), work on which has started in 2018 and is scheduled for publication in 2021. Improved process understanding ranks high among the objectives of CMIP, and indeed several of the 21 CMIP6-Endorsed Model Intercomparison Projects (MIPs) explicitly state it as their main goal (Eyring et al., 2016).

Lessons learnt from CMIP5 include: making more use of idealized experiments to isolate processes, a more concerted effort to link model output to observations through forward operators and the creation of observation-based datasets whose structure and metadata mirror that of the model-based datasets, as well as more coordinated efforts to apply community-developed evaluation packages (Stouffer et al., 2017).

2.2 WMO Commission for Basic Systems (CBS) exchange of NWP scores

In global NWP, the World Meteorological Organization (WMO) Commission for Basic Systems (CBS) encourages, and has set guidelines for, the exchange of upper-air and surface scores between global operational forecasting centres. The collection and display of scores is managed by the Lead Centre for Deterministic NWP Verification (LC-DNV) at the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Lead Centre for Verification of Ensemble Prediction Systems (LC-EPV) at the Japanese Meteorological Administration (JMA).

Comparison of verification results from different models may already provide some useful information about the characteristics of a problem. One example is forecast 'busts' (usually defined as drops in anomaly correlation below a defined threshold, Rodwell et al., 2013), which in some cases may occur for all available models, for a subset of models, or a single model (Figure 1). Differences in bust characteristics across models may result from differences in the data assimilation and modelling systems, while busts which occur in all

models suggest a more fundamental issue with the observational data in this case and/or genuinely reduced atmospheric predictability. However, broad-scale measures, such as the 500 hPa geopotential anomaly correlation over a large area, can give some first indications only, and further investigations are needed to gain a deeper understanding of the role of specific processes in the occurrence of forecast busts. Rodwell et al. (2013) employed the method of composites, which can be considered a variant of conditional verification, to identify features common to certain types of forecast busts. Furthermore, by swapping initial conditions between models and by assessing state-dependent predictability using the ensemble forecast they showed that springtime busts in Europe are sensitive to the initial state around the Rocky Mountains and the assimilation and forecasting of mesoscale convective systems (MCSs) over North America.

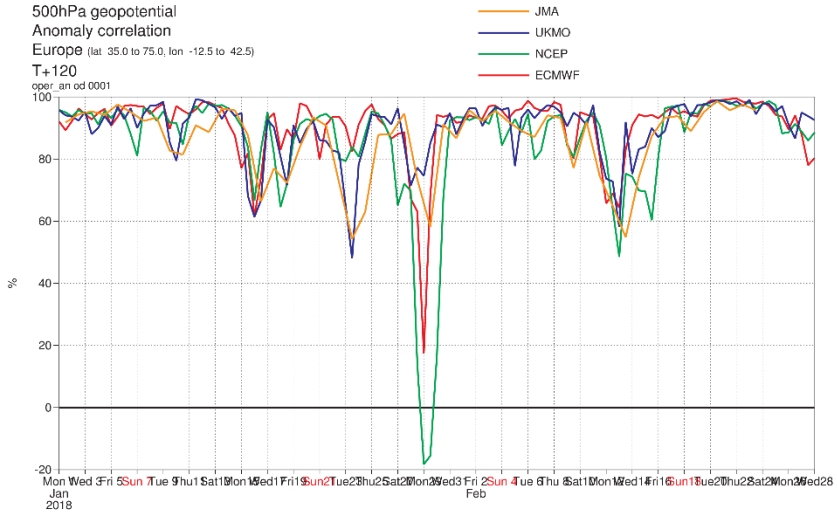

Figure 1: Time-series of 500 hPa anomaly correlation over Europe at forecast day 5 from four global NWP models, showing periods with substantial drops in skill shared by different numbers of models.

A model intercomparison can also contribute to the understanding of systematic errors, especially for surface parameters. However, to be useful, more detailed information is required than just areal means of scores (on which the example in Figure 1 is based). The recently added surface component of the WMO CBS exchange of scores therefore asks participating NWP centres to provide scores for individual stations, which also avoids the need for coordinating and updating station lists between centres. Having station-wise scores allows to condition the verification depending on orography, distance from the coast, surface type, etc. In order to draw conclusions about processes, however, up-to-date knowledge about the different parameterizations used in the models of different centres is required.

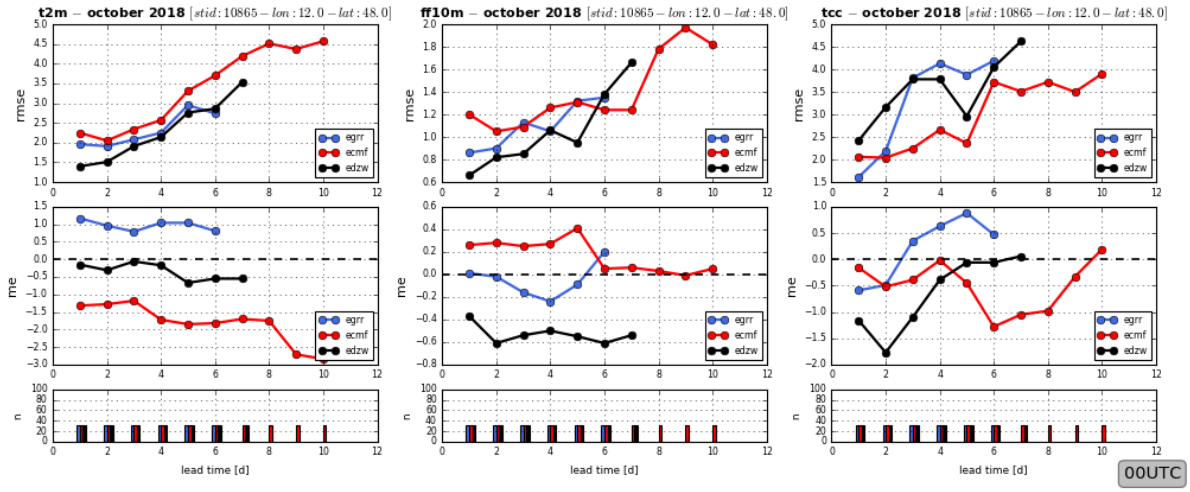

Figure 2: Example of the CBS surface score model intercomparison for the SYNOP station of Munich, Germany. Currently DWD (black), UKMO (blue), and ECMWF (red) provide station-wise surface scores on a regular basis. Shown are RMSE (top) and ME (centre) of forecasts for 00 UTC in October 2018 for 2m temperature (left column), 10m wind speed (middle column), and total cloud cover (right column). Bottom panels show number of values which are actually available from each model over the 1-month period.

While the CBS exchange of upper-air scores has been well established for many years, the CBS surface score exchange is still at the developing stage. At the time of this writing, DWD, UKMO, and ECMWF are providing surface scores on a regular basis (Figure 2), and JMA has sent test data. Ideally all major global forecasting centres would contribute, allowing a comprehensive evaluation of near-surface forecast quality and systematic errors in different geographical settings across a range of models.

Figure 2 illustrates that models can have similar RMSEs and at the same time rather different biases. This behaviour can be seen for 2m temperature and 10m wind speed in the above example. The above type of plot can be produced interactively for thousands of locations using a clickable map interface on the web page of the WMO LC-DNV and is accessible by participating centres. Areal averages (means over the set of common stations in predefined areas) will also be computed by the LC-DNV in the near future. This requires the setting up and maintenance of a CBS surface scores database, which is currently being developed. Once this is achieved (second half of 2019), LC-DNV will be able to provide area-averaged surface scores similar to CBS upper-air verification.

2.3 The Common Verification of the COSMO consortium

The Consortium for Small-scale Modelling (COSMO), one of the European consortia for Limited-Area Modelling, has established as permanent activity the Common Verification. The aim is to verify all operational configurations of the COSMO model run by members of the Consortium. This activity permits to identify problems in one or more of the national implementations, as well as to detect systematic model deficiencies.

Verification is performed by each member using the common verification software VERSUS, which was developed as part of the consortium activities, over the same area, given by the

intersection of model domains (Figure 3), using the same set of stations. The common software allows for a homogeneous, standardized and objective way to apply, calculate and present the verification scores.

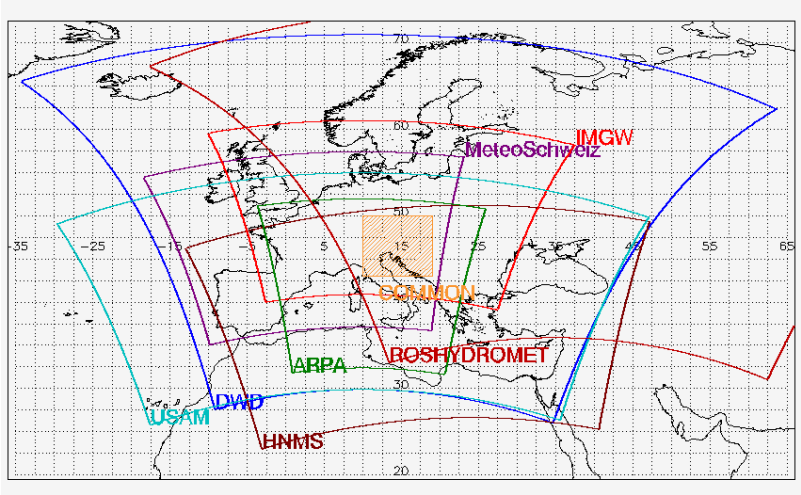

Figure 3: Domains of the various implementations of the COSMO model. A sub-area (yellow) in central Europe serves as a common basis for comparing scores from all the implementations.

Common Verification is performed for all the convection-parametrised versions of COSMO (5-7 km) as well as for some convection-permitting versions (2-3 km), which have a smaller region of overlap. The activity permits to identify model systematic errors which need to be addressed, such as the underestimation of the 2m temperature diurnal cycle (Figure 4), and the underestimation of 2m dew-point temperature over dry areas.

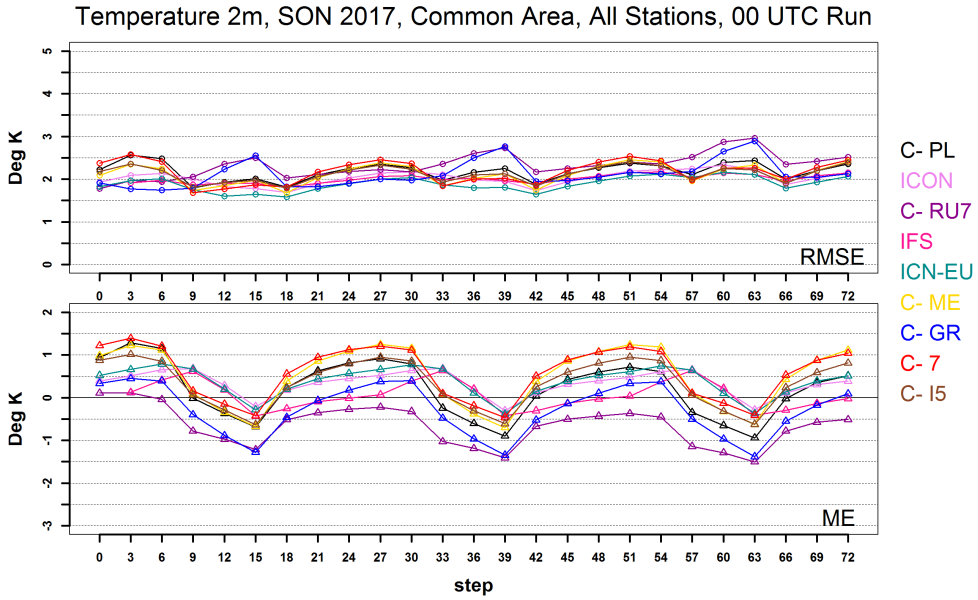

Figure 4: Example of the COSMO Common Verification for 2m temperature in autumn 2017. Shown are RMSE (top) and ME (bottom) of forecasts from 00 UTC as a function of the forecast range. All the lines labelled as C- are relative to COSMO model implementations either at 7 or 5 km horizontal resolution. The night-time warm bias and the daytime cold bias are evident in all the COSMO model

runs.

Conditional Verification results are also produced once there is an indication based on a prior analysis from a user, that there is an interdependency of variables or conditions on parameter values that is worth further investigation.

While the common verification of the COSMO consortium deals with limited area models only, it illustrates the need for, and potential benefits of, standardization of verification procedures for process-oriented model intercomparisons.

2.4 Major model intercomparison activities

The CBS and COSMO model intercomparison activities are focused on comparing scores from operational models and more geared towards administrative and/or user-oriented aspects of performance. For process-based intercomparison, there are several NWP projects with an approach more similar to CMIP. Examples (mostly GEWEX triggered and super-site based, such as GABLS3 and GABLS4) are:

(i) Completed NWP intercomparisons:

a. GABLS-4: GEWEX Atmospheric Boundary Layer Study

http://www.umr-cnrm.fr/aladin/meshtml/GABLS4/GABLS4.html

b. Greyzone Cold Air Outbreak (CONSTRAIN)

http://appconv.metoffice.com/cold_air_outbreak/constrain_case/home.html

c. Arctic winter boundary layer over sea ice - SCM intercomparison

http://onlinelibrary.wiley.com/doi/10.1002/2016MS000630/epdf

d. GABLS-3: Intercomparison and evaluation study for single-column models - full diurnal cycle over Cabauw-

https://link.springer.com/article/10.1007/s10546-014-9919-1

e. GCSS, WGNE Pacific Cross-section Intercomparison,

Tropical and sub-tropical cloud transitions - GCM intercomparison

https://journals.ametsoc.org/doi/full/10.1175/2011JCLI3672.1

(ii) Ongoing NWP intercomparisons:

a. Precipitation Diurnal Cycle

http://www.gewex.org/panels/global-atmospheric-system-studies-panel/gass-projects

b. Surface Drag and Momentum Transport (COORDE)

http://www.gewex.org/panels/global-atmospheric-system-studies-panel/gass-projects

c. Greyzone EUREC4A -this pertains the representation of clouds at intermediate resolution, between explicit and implicit-

https://www.metoffice.gov.uk/research/collaboration/grey-zone-project/grey-zone-second-phase

d.WGNE Surface Flux Interomparison (no url at the moment)

e. Demistify: An LES and NWP Fog Modeling Intercomparison

http://www.gewex.org/panels/global-atmospheric-system-studies-panel/gass-projects

f. GEWEX Upper Tropospheric Clouds and Convection Process Evaluation Study

http://www.gewex.org/panels/global-atmospheric-system-studies-panel/gass-projects

One key feature of these intercomparison studies compared to, say, the CBS model intercomparison, is that they are targeting known major model weaknesses. Comparing output from various models helps to identify relevant processes and error sources. Thus they are naturally closer to being 'process-oriented'. An advantage of the CBS intercomparison, on the other hand, is that it is continuous over a long period of time. It could be useful to adopt some of the process-oriented diagnostics developed in dedicated projects within the long-term evaluation of CBS.

3. Process-oriented metrics and methods

In contrast to verification for performance monitoring purposes, where aggregation of scores into a small number of indices is required, process-oriented verification attempts to disaggregate and stratify results. It may use similar scores (bias, RMSE, correlation) as the performance monitoring but computes them over suitably defined sub-samples representing distinct atmospheric states, such as 'clear-sky' or 'well-mixed'. In a sense, it bridges the gap between the standard verification over all cases and individual case studies.

An American Meteorological Society (AMS) special collection of papers will be devoted to process-oriented evaluation of climate and Earth system models. The Model Diagnostics Task Force (MDTF) of the NOAA Modeling, Analysis, Predictions, and Projections Program (MAPP) is organizing the collection. The stated goal of the MDTF is to move beyond performance-oriented metrics toward process-oriented metrics of models, to aid current efforts to develop the next generation of climate and Earth system models including those related to CMIP, and to link model development and evaluation efforts across modelling centres.

One recent paper from this collection by Nearing et al. (2018) proposes an approach based on information theory to obtain a deeper insight into model performance and model realism. Applying the method to the PLUMBER land model intercomparison project (Best et al., 2015) they find that current operational land surface models do not make full use of the

information contained in the meteorological input provided and suffer from similar patterns of process-level structural errors. For tropical cyclones (TCs), Kim et al. (2018) use the method of compositing (on maximum wind speed and precipitation percentiles) to compare the structures of TCs of similar intensity between different models. They find that model differences can be traced to the sensitivity of convection to ambient moisture in the models, and to their representation of the surface heat flux feedback through low-level wind speed.

The MDTF is also developing an open-source, python-based software framework which integrates individual contributions of metrics developed by participating research groups and modelling centres. Some examples are

- Fast timescale diagnostics for tropical convection (D. Neelin)
- Diagnosis of warm rain processes using MODIS and CloudSAT (K. Suzuki)
- Diagnostics on MJO amplitude and propagation (X. Jiang)
- Diagnostics on MJO tropical-extratropical teleconnections (E. Maloney)
- Diurnal cycle, 2m temperature, and surface energy budget diagnostics (A. Dai, J. Wang)

Current participation in the process-oriented metrics framework is mainly by U.S. institutions. They are however interested in opening the activity to the wider global modelling community, for example via WGCM and WGNE. It is seen as complementary to, and extensible into, the existing CMIP Earth System Model eValuation Tool (ESMValTool) which is used by the international climate modelling community (https://www.esmvaltool.org/).

A recent overview by the NOAA MAPP Model Diagnostics Task Force about the common framework for process-oriented evaluation of climate and weather forecast models is given by Maloney et al. (2019).

As with verification in general, when using community tools, most of the programming work is in pre-processing the model data into the format required by the tool. The actual score or metric computation is usually comparatively straightforward. Another resource-intensive part of the work is analysis and interpretation, and gaining a deeper understanding of the results obtained. Thus, it is not always clear whether a net saving of resources is achieved by using such tools. Use of a common programming language (e.g. python) and data formats (e.g. netcdf) may already go a long way towards enabling increased collaboration between institutions.

A type of verification where the score computation itself is more complex, and where the development of community tools is most beneficial, is spatial verification (Gilleland et al., 2010). The SpatialVx package https://ral.ucar.edu/projects/icp/SpatialVx/ is a good example. Another example of using spatial verification methods for model intercomparison is given by Dorninger and Gorgas (2013).

4. Multi-dataset verification and up-scaling

Observations are estimates of the truth, and observation errors and representativeness issues inevitably add uncertainty to the verification results. As far as random observation errors are concerned, their effect can be reduced by looking at larger samples (longer time

intervals and/or larger regions). Systematic observation errors, however, can only be identified by using different, independent observational datasets.

4.1 Multiple datasets and cloudiness and solar radiation biases

To illustrate the benefits of, and indeed the need for, multi-dataset verification, an example from evaluating ECMWF cloudiness and radiation forecasts is discussed here. Figure 5 shows the Integrated Forecasting System (IFS) HRES wintertime short-range forecast bias for total cloud cover (TCC) in Europe.

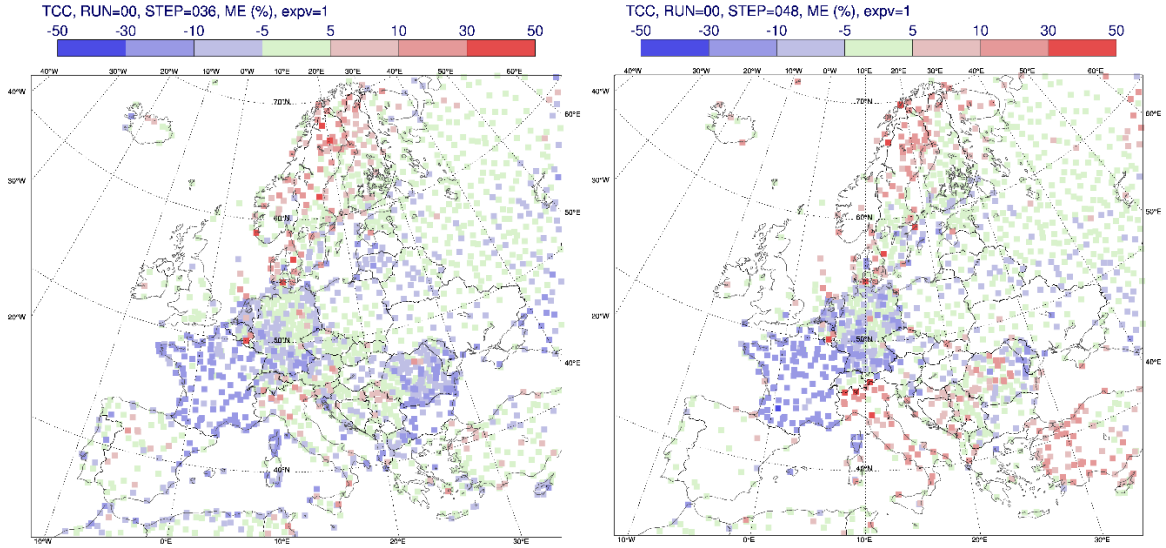

Figure 5: Bias in total cloud cover at forecast day 2 in NDJ 2017/18 from verification against SYNOP. Left panel: 12 UTC (+36 h forecast), right panel: 00 UTC (+48 h forecast). Note the jumps in bias across country borders caused by different characteristics in the observation methodology between national surface station networks.

While there is some indication of a generally negative bias in TCC in central Europe, and a positive bias in Scandinavia, strong jumps in bias across country borders (likely resulting from differences in observation methodologies) are apparent; hampering the quantitative evaluation and interpretation of the results. Differences between national surface station networks may be caused by different degrees of automated versus human total-cloud-cover observations, and associated differences in observation representativeness and error characteristics (Mittermaier, 2012).

The question arises as to what extent the apparent IFS underestimation of wintertime cloudiness in large parts of Europe is real and how much observational biases distort the result. An alternative is to look at SYNOP downward solar radiation measurements which are provided by some countries and available on the GTS. Evaluation of the forecast against this dataset shows a more uniform picture, with positive radiation biases (consistent with negative TCC errors) generally increasing from west to east (Figure 6a). To further increase confidence in these results, as well as to have more complete areal coverage, downward solar radiation has also been evaluated against a satellite-derived product (Figure 6b). These two independent datasets give very similar results for the bias both in terms of magnitude and geographical pattern. Having the SYNOP observations is important because they

represent ground truth for the satellite product, which is not a direct measurement but heavily processed using NWP information.

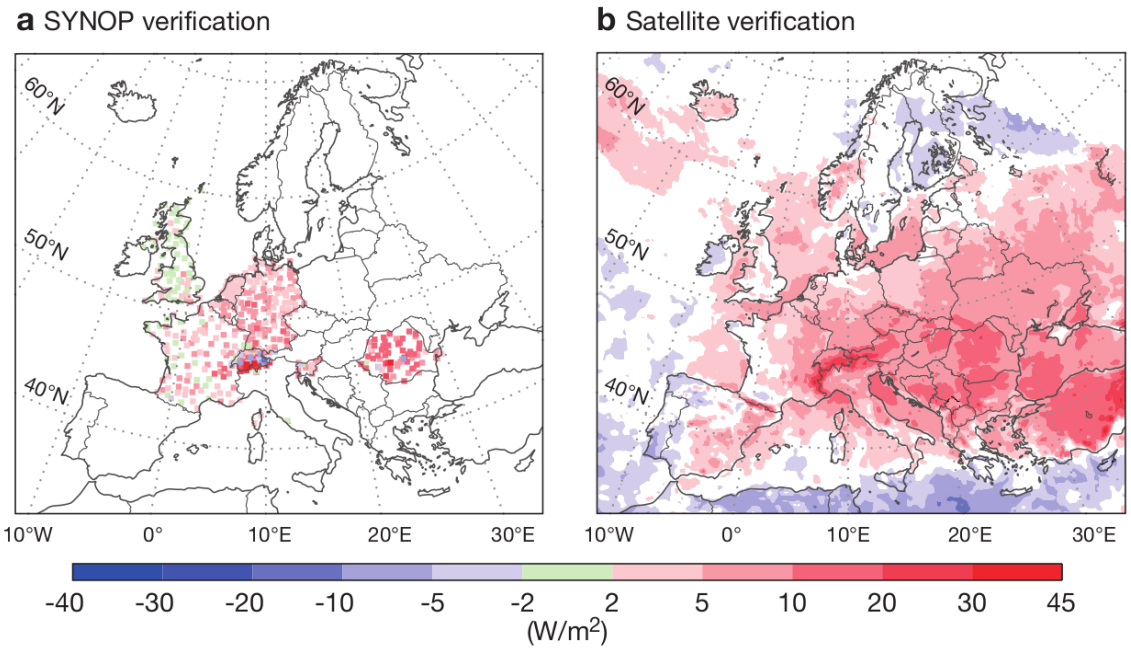

Figure 6: Bias in downward surface solar radiation (24-hour averages) at forecast day 2 in NDJ 2017/18 from (a) verification against SYNOP and (b) verification against the corresponding satellite product (right) from the Climate Monitoring Satellite Application Facility (CM SAF), from Haiden et al. (2018).

4.2 Up-scaling and the 'light convective precipitation' issue

As in the case of radiation, precipitation can be verified against ground-based point observations (rain gauges) and remotely sensed (radar, satellite) data. Ideally, an evaluation uses both types of datasets and compares results, taking into account the specific strengths and weaknesses of each individual dataset. One issue in the IFS and other global models is the overestimation of the frequency of occurrence of light convective precipitation. In the extra-tropics, this overestimation contributes to the annual drop of precipitation forecast performance during summer (Haiden et al., 2012). In the tropics, it lowers forecast performance throughout the year. However, from verification against rain gauges it is not clear what fraction of this overestimation is due to model issues, and how much is an artefact of the verification, arising from the representativeness mismatch between point observations and model grid scale.

Ideally, the precipitation forecast is verified against observations up-scaled to the native model grid; thereby requiring a density of at least several stations per grid box, which is not reached in any of the available datasets. However, a station density on the order of 0.01 km^{-2} as in the high-density rain gauge observations in Italy (distance of about 10 km between stations) does allow up-scaling to a coarser grid such as 0.5 deg.

Figure 7 shows the frequency bias for verification against up-scaled and point observations

in Italy. In summer (left panel), the frequency of light precipitation is overestimated by the model by a factor of up to 3 when verified against the point observations (red curve). Verification against the up-scaled observations shows that much of this bias is still present on a scale of about 50 km and is not merely an effect of the scale mismatch. In winter (right panel) the overestimation of light precipitation nearly disappears when verified against up-scaled observations. This comparison suggests that the winter issue is largely an artefact of the scale mismatch between model and observation, whereas the summer issue is an actual model problem. In the IFS, it appears to be related to the parametrization of rainfall evaporation from deep convection.

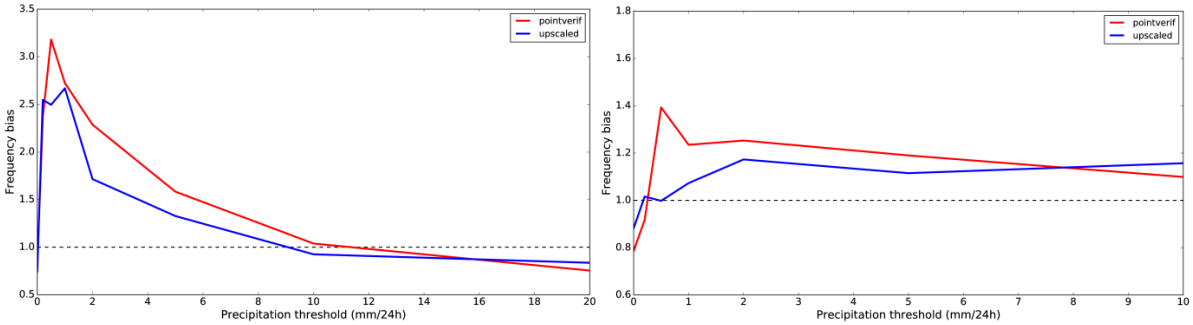

Figure 7: Frequency bias of the HRES short-range (30 h) precipitation forecast for Italy in summer (JJA 2017, left) and winter (DJF 2017-18, right) from verification against high-density observations. Red curves: verification against point observations; blue curves: verification against the same observations, up-scaled to a 0.5 deg grid. Note different scaling of axes in the two plots.

5. Conditional verification and machine learning

A standard technique used in process-oriented verification is conditional verification or, in other words, a stratification of verification results according to a control variable. In the simplest case, this variable is the verified quantity itself (e.g. cloud cover forecast error as a function of observed and/or forecast cloud cover). In the more general case, it is one or several variables which quantify the effect of specific processes on the verified quantity, thereby allowing a distinction between different 'regimes'. A typical example would be the evaluation of night-time 2m temperature biases as a function of cloud cover and wind speed.

Conditional verification can be used to quantify relationships between errors in different variables and can help to disentangle their sources. Figure 8a shows that the night-time negative 2m 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). However, when weighted by the frequency distribution of the cloud cover errors, shown as green bars in the plot, it turns out that cases where total cloud cover is underestimated and cases where it is nearly correct contribute about equally to the negative T2m bias (Figure 8b). This result indicates that the wintertime negative total cloud cover bias in the IFS over central Europe (on the order of 10% against SYNOP observations,

not shown) contributes to, but does not fully explain, the negative night-time T2m bias in this region (Haiden et al., 2018). In cases when the total cloud cover is correctly predicted, the negative T2m bias could result from other cloud errors, e.g. an underestimation of cloud optical depth, erroneous cloud type or erroneous cloud base height. It could also be from errors in processes not directly related to clouds, such as vertical mixing or coupling with the surface.

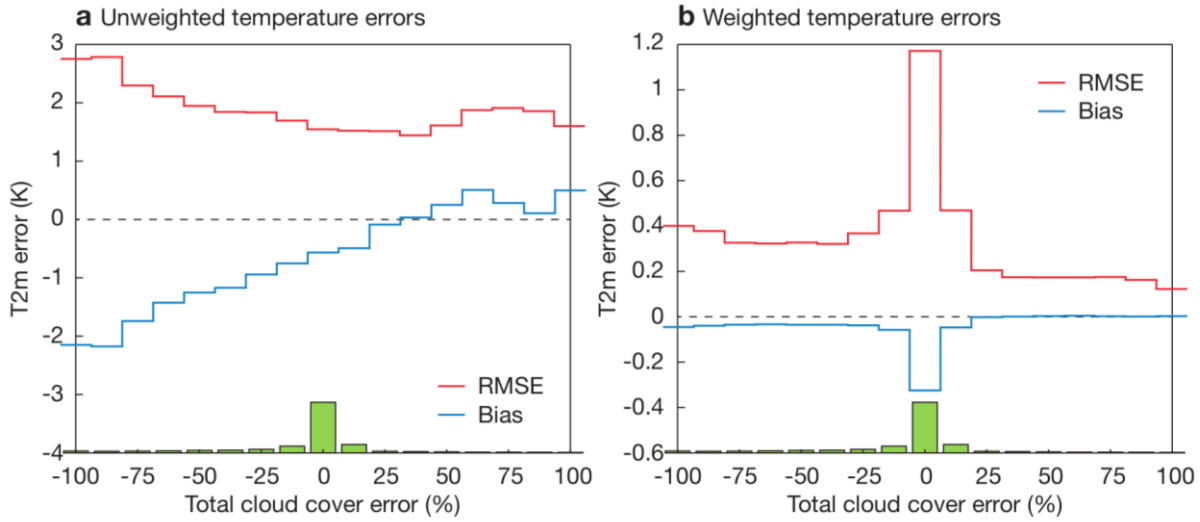

Figure 8: Root mean square error (RMSE) and mean error (bias) for T2m forecasts valid at 00 UTC as a function of the total cloud cover (TCC) error for December–January–February2016/17 in a central European domain (48–55°N, 0–15°E) at a lead time of 12 hours (a) averaged for each TCC error bin and (b) averaged for each TCC error bin and weighted by the TCC error relative frequency of occurrence. Green bars show the TCC error frequency distribution at an arbitrary vertical scale (from Haiden et al., 2018).

A stratification of T2m forecast errors for different categories of total cloud cover in the forecast and observations has also been found useful at Environment Canada (EC) where it is used to investigate differences in systematic error behaviour of forecasting systems with different spatial resolutions. In Figure 9 these stratifications are the Canadian Arctic Prediction System (CAPS, red line, 3km resolution), the Canadian Regional Deterministic Prediction System (RDPS, blue, 10 km resolution) and the Canadian Global Deterministic Prediction System (GDPS, green line, 25 km resolution). Figure 9 shows how the biases of the three systems change in relation to each other for different cloud (error) conditions.

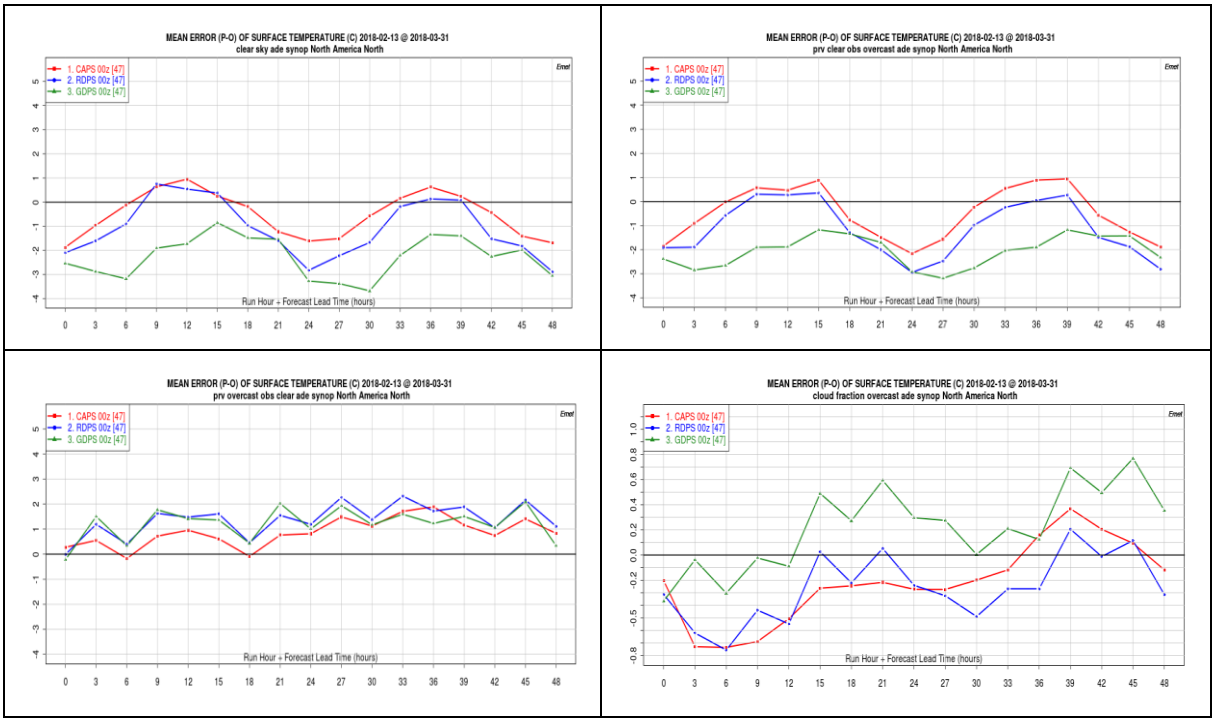

Figure 9: 2m near-surface air temperature bias in three different forecasting systems run by Environment Canada, evaluated conditioning on total cloud cover. Top-left panel: both forecast and observations are clear sky; top-right panel: forecast is clear sky but observation is cloudy; bottom-left panel: forecast is cloudy but observation is clear sky; bottom-right panel: both forecast and observations are cloudy.

EC has also investigated the effect of various options for computing 2-m temperature in the model as a function of lowest model level and skin temperatures, and for a given roughness length. The representation of surface inhomogeneities in the form of tiles raises the question concerning which method of 2-m temperature computation best matches the verifying station observation. In many high latitude land areas, for example, model tiles have a significant fraction of water resulting from the presence of many small lakes. The choice of different roughness lengths does not affect the verification results as much as the soil-only versus aggregated soil and water tile air temperature. In the IFS, 2-m temperature is computed for the low vegetation tile even in generally forested areas, to match the SYNOP observations which are usually located in locally forest-free areas. However, this method has its limits, as the thermal and dynamic effects of nearby forests is not fully taken into account.

While there is a wide range of potentially useful ways in which verification results may be stratified, it is important to consider the statistical effects of any such sub-sampling on the results. A well-known issue of great practical relevance is the focus on observed extreme events, which distorts verification results and may discredit skillful forecasts, as discussed by Lerch et al. (2017). In general, sub-sampling based on observations *and* forecasts is a good strategy for avoiding such issues.

Specific verification metrics highlight specific aspects of forecast quality. For a continuous variable, the most basic distinction is between systematic and 'random' errors, which are

(for a deterministic forecast) described by the mean error or bias, and the error standard deviation. The word random has been put in quotes here because a substantial part of the apparently random error may in fact be a state-dependent systematic error. This type of question is where conditional verification can provide additional insights, and a better estimation of the fraction of truly non-systematic errors.

How can we more efficiently identify patterns and processes related with specific model errors? In some cases, physical considerations and/or forecaster's experience may already point towards specific patterns and relationships, such as in the 2m temperature vs cloud error example. However, in other cases relationships will be more complex, possibly more non-linear, and multivariate, making the search for error sources more difficult and time consuming. Machine learning (ML) algorithms may help to identify and quantify the most important (and sometimes possibly unexpected) connections between state variables, tendencies, and errors with a reasonable investment of time. This could give model developers useful information for further in-depth investigations.

ECMWF has started to look into ML methods to study biases in wintertime 2m temperature at high latitudes. For the Sodankyla supersite, 2m temperature biases were related to a range of other variables using deep learning techniques, in particular random forest (Figure 10). One of the important differences between this method and multivariate linear regression is that there is no a-priori assumption about linearity, or the shape of a relationship, so that also rather non-linear relationships can be identified. The ranking of the predictors in terms of importance for correcting the forecast can serve as a basis for a physical interpretation of the forecast errors' potential sources. With this approach, model diagnostic can directly benefit from state-of-the-art ML techniques applied to post-processing problems (Taillardat et al. 2016, Ben Bouallègue 2016, Rasp and Lerch 2018).

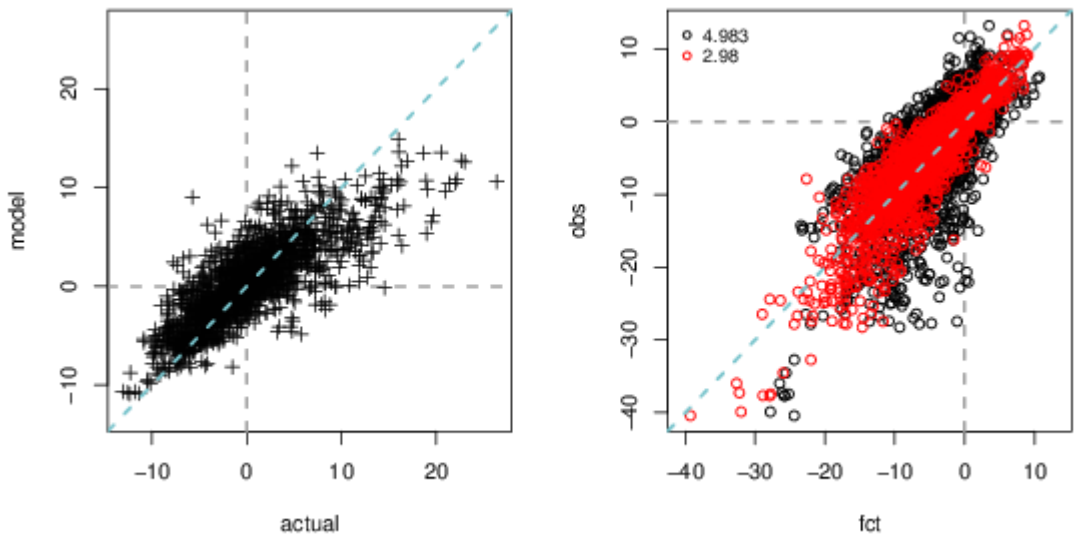

Figure 10: Left: scatterplot of actual vs predicted 2 m temperature forecast errors at Sodankyla. The prediction is based on a random forest algorithm. Right: scatter plot of 2 m temperature forecast vs observation before (black) and after (red) conditional bias correction with random forest.

6. Supersites

One of the main problems in NWP model development, especially in the area of physical parameterizations, is compensating errors. Their presence hampers progress by creating situations in which an improved representation of some physical process cannot be implemented in a model because it no longer compensates a (previously hidden) bias caused by other parts of the model. A typical example is the parameterization of turbulent mixing near the surface under very stable conditions, which is known to overestimate vertical exchange rates compared to observations. This situation would lead to a warm bias in 2m temperature. However, in the IFS, for example, the warm bias is compensated in winter by an underestimation of cloudiness, which by itself would lead to a cold bias. Thus, a long-term strategy of NWP model development must include the identification, and where possible removal, of compensating errors. This procedure requires constraining the parameterizations involved in as many ways as possible. A powerful resource for such an approach is the use of supersite datasets, which are obtained in highly instrumented locations.

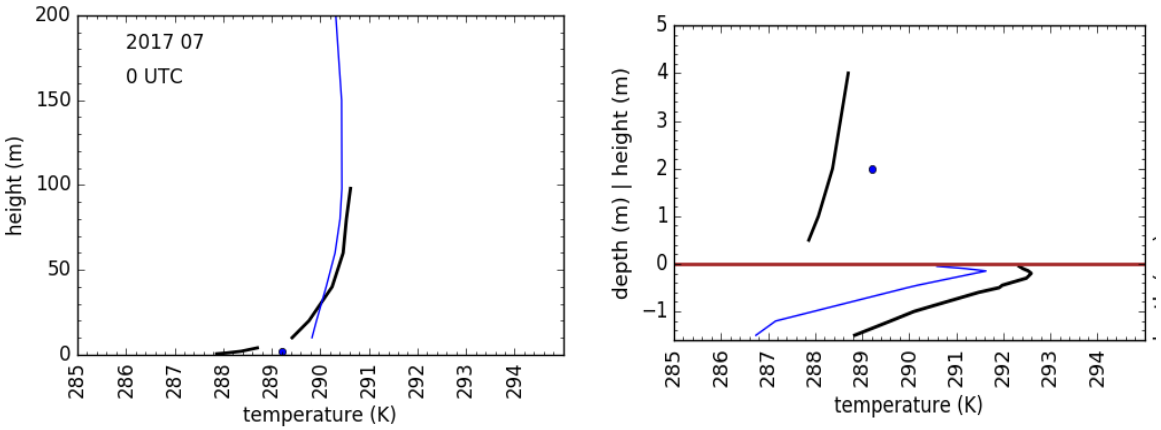

Figure 11: Comparison of IFS atmospheric and soil temperature profiles (blue) with observations from the Lindenberg supersite (black). Left panel: comparison with tower (99 m) and mast (4 m) data; right panel: comparison with mast (4 m) and soil data. The blue dot shows the forecast 2m temperature.

Measurements at supersites such as Lindenberg (Germany), Cabauw (Netherlands), or Sodankylä (Finland) typically encompass atmospheric profile data from masts and towers, soil profile data, as well as surface radiation and turbulent fluxes. They may also include cloudiness parameters, e.g. from lidar measurements. Having concurrent observations of atmospheric and soil temperature profiles helps to identify the source of 2m temperature biases. In the IFS, for example, night-time 2m temperature in summer is overestimated, whereas the temperature in the soil layers is underestimated (Figure 11). During daytime (not shown) conditions are reversed; suggesting that the atmosphere-surface thermal coupling is too strong in the model.

In the framework of the Short-Range Numerical Weather Prediction cooperation (SRNWP) of EUMETNET, a Data Exchange Programme is maintained among European Meteorological Services. The goal of this activity is to support the development of soil-vegetation-

atmosphere transfer models within the SRNWP community, by providing good quality operational data from a limited set of well instrumented and high quality observation sites. Soil, surface, and boundary layer data are provided and collected on a dedicated website (http://srnwp.cosmo-model.org/). Participating sites are Lindenberg (D), Payerne (CH), Capofiume (I), Sodankylaa (FI), Cabauw (NL), Toulouse (F) and Cardington (UK).

A type of supersite specifically focussed on radiation, cloud, and aerosol processes is provided by the Atmospheric Radiation Measurement (ARM) facility (https://www.arm.gov/). Collocated measurements of a comprehensive set of radiation and cloud parameters allows to constrain microphysical parameterizations, for example the representation of rain formation and sub-cloud evaporation (Ahlgrimm and Forbes, 2014), or cloud condensate variability (Ahlgrimm and Forbes, 2016).

7. Conclusions and outlook

Process-oriented verification is an integral part of model development and it is a term which encompasses a wide range of verification activities and methodologies. There are ongoing efforts to coordinate such activities on a national level (e.g. the Model Diagnostics Task Force (MDTF) in the United States) or internationally (e.g. within the CMIP framework in the climate modelling community, and through various model intercomparison activities in NWP) and to develop common toolboxes and software repositories. The goal of such initiatives is to avoid duplication of work and to make use of the existence of a range of models with different formulations in the search for error sources. Given the fundamental similarities and overall convergence between climate and NWP models towards Earth-system modelling, it makes sense for the NWP and climate modelling communities to combine efforts in this area.

Recommendation: Use model (conditional) biases as a common topic of interest to initiate closer collaboration in verification activities between the climate and NWP modelling communities.

Spatial verification methods are a prime example of an area where the development and use of community tools has been beneficial. The more complex score computation in this case (as compared to classical point verification) means that a larger fraction of the verification procedure is non-centre-specific and can be performed with the same software package. The same applies, to some extent, to process-oriented verification. However, while spatial verification already encompasses a wide spectrum of methods, process-oriented verification includes a potentially even larger set of approaches, so the challenge will be to develop routines that are both practical and specific, but also sufficiently general to be of use for the wider NWP community. One candidate method could be conditional verification, because an unknown (and potentially substantial) part of what are considered 'non-systematic' errors in NWP may actually be flow- or situation-dependent systematic errors.

Recommendation: Investigate whether a generic 'conditional verification tool' can be created (and integrated with existing packages) which is sufficiently flexible to be used across centres.

With regard to the CBS exchange of scores, it would be a major step forward if more global centres would participate in the exchange of surface scores. However, to guarantee a fair comparison, care needs to be taken in the spatial aggregation of the station-wise scores which are exchanged. For example, due to differences in observation quality control procedures at different centres, the number of available data values may differ.

Recommendation: Encourage global forecasting centres to provide WMO CBS surface scores (currently DWD, UKMO, JMA, ECMWF).

Machine learning methods hold promise for detecting patterns and non-linear relationships between model state variables that have been overlooked or less explored. While the nature of such algorithms does not allow a direct pinpointing of the source of a model error, they may efficiently point modellers to areas for further investigation by other (physically based) methods.

Recommendation: Encourage the use of machine learning in process-oriented verification studies.

Acknowledgements

We would like to thank Ron McTaggart-Cowan (Environment Canada) and Zied Ben Bouallegue (ECMWF) for helpful suggestions and contributions to the text.

References

Ahlgrimm, M., and R. M. Forbes, 2014: Improving the representation of low clouds and drizzle in the ECMWF model based on ARM observations from the Azores. *Mon. Wea. Rev.,* **142,** 668-685.

Ahlgrimm, M., and R. M. Forbes, 2016: Regime dependence of cloud condensate variability observed at the Atmospheric Radiation Measurement sites. *Q. J. R. Meteorol. Soc.,* **142,** 1605-1617.

Ben Bouallègue, Z., 2016: Statistical post-processing of ensemble global radiation forecasts with penalized quantile regression. *Meteorol. Z.,* **26**, 253–264.

Best, M. J., G. Abramowitz, H. R. Johnson, A. J. Pitman, G. Balsamo, A. Boone, M. Cuntz, B. Decharme, P. A. Dirmeyer, and J. Dong (2015): The plumbing of land surface models: benchmarking model performance. *J. Hydromet.,* **16,** 1425-1442.

Dorninger, M., and T. Gorgas, 2013: Comparison of NWP-model chains by using novel verification methods. *Meteorol. Z.,* **22,** 373-393.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.,* **9,** 1937–1958.

Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Met. Soc.,* **91,** 1365-1376.

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

Kim, D., Y. Moon, S. J. Camargo, A. A. Wing, A. H. Sobel, H. Murakami, G. A. Vecchi, M. Zhao, and E. Page, 2018: Process-oriented diagnosis of tropical cyclones in high-resolution GCMs. *J. Climate,* **31,** 1685-1702.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science,* **32,** 106-127.

Maloney, E. D., and 28 co-authors, 2019: A framework for process-oriented evaluation of climate and weather forecasting models. *Bullet. Amer. Met. Soc.,* (submitted).

Mittermaier, M., 2012: A critical assessment of surface cloud observations and their use for verifying cloud forecasts. *Q. J. R. Meteorol. Soc.,* **138,** 1794–1807.

Nearing, G. S., B. L. Ruddell, M. P. Clark, B. Nijssen, C. Peters-Lidard, 2018: Benchmarking and process-diagnostics of land models. *J. Hydromet.,* **19,** doi: 10.1175/JHM-D-17-0209.1.

Rasp, S., and S. Lerch, 2018: Neural networks for post-processing ensemble weather forecasts. *Mon. Wea. Rev.,* **146,** 3885-3900.

Rodwell, M. J., L. Magnusson, P. Bauer, P. Bechtold, M. Bonavita, C. Cardinali, M. Diamantakis, P. Earnshaw, A. Garcia-Mendez, L. isaksen, E. Källen, D. Klocke, P. Lopez, T. McNally, A. Persson, F. Prates, and N. Wedi, 2013: Characteristics of occasional poor medium-range weather forecasts for Europe. *Bull. Am. Meteor. Soc.,* **94,** 1393-1405.

Skamarock, W. C., S.-H. Park, J. B. Klemp, and C. Snyder, 2014: Atmospheric kinetic energy spectra from global high-resolution nonhydrostatic simulations. *J. Atmos. Sci.,* **71,** 4369-4381.

Stouffer, R. J., V. Eyring, G. A. Meehl, S. Bony, C. Senior, B. Stevens, and K. E. Taylor, 2017: CMIP5 scientific gaps and recommendations for CMIP6. *Bull. Am. Meteor. Soc.,* **98,** 95-105.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.,* **144,** 2375–2393.