WORLD METEOROLOGICAL ORGANIZATION	DPFS/TT-SV/Doc. 4.1a
COMMISSION FOR BASIC SYSTEMS OPAG on DPFS	(X.IX.2014) _______
MEETING OF THE CBS (DPFS) TASK TEAM ON SURFACE VERIFICATION	Agenda item : 4.1
GENEVA, SWITZERLAND 20-21 OCTOBER 2014	ENGLISH ONLY

Implementation of global surface index at the Met Office

Submitted by Marion Mittermaier

Summary and purpose of document

This document provides background information on the implementation of a global surface index at the Met Office, focusing on five parameters: temperature, wind, precipitation, cloud base and total cloud amount. A variety of site lists were tested. The biggest challenge has been the equalization of site lists and ambiguities with the precipitation site-based climatologies required by the Stable Equitable Error in Probability Space (SEEPS).

Action Proposed

The meeting is invited to comment on the implementation and discuss the challenges of making results comparable between modelling centres as we foresee this to be the biggest issue.

Reference: -

1. Background

The Met Office has had forecast accuracy indices (and corporate performance targets based on these) for several decades. To date the global index has been based on upper-air parameters, whereas the focus of the UK index was on surface parameters over the UK. With the improvements in horizontal resolution to ~20 km, and the use of global model output for providing site-specific (post-processed) forecasts for the UK at longer lead times, and globally from day 0, there has been increasing interest in understanding how the raw model output is performing for surface weather.

The studies at the Met Office focused on adapting the existing UK index framework for this purpose. All the results shown below are from a variety of tests we have run in recent years. We have just delivered a trialling capability for model developers to test the impact of model changes on surface parameters, and this will be used to calculate a long time series of performance in the very near future.

2. First tests

2.1. Sampling and observations

One of the greatest concerns from the very beginning was the uneven sampling across the globe (in terms of land and sea areas) and the variations in density where surface land observations are available. Another issue was the potential influence of the coast in relation to the coarseness of the model grid. Earlier versions of our verification system were not checking whether a site ended up being represented by a model sea point, or be a mixture of land and sea points, through the process of interpolation. Of course, for upper air parameters this issue is immaterial. Another potential headache was sites located near areas of steep gradients in orography. Even with height adjustments, issues due to the severely smoothed model representation of the terrain could be problematic. Last but not least there is the issue of observations quality. Here we took the pragmatic decision that airports were likely to be the most reliable observing locations, due to international aviation requirements, and as a first attempt we restricted ourselves to airports only (i.e. sites that reported METARs). In the final selection we then restricted the sites further by excluding any airports that could potentially be or are influenced by sea points in the model, or near steep gradients in orography.

Figure 1: Map showing the 139 airport sites identified as verification locations.

2.2. Formulation

Scores are calculated from the following parameters:

- Near-surface (1.5m) temperature
- Near-surface (10m) wind speed & direction (as vector)
- Precipitation yes/no (equal to or greater than 0.5, 1.0 and 4.0 mm over the preceding 24 hours)
- Total cloud amount yes/no (equal to or greater than 2.5, 4.5 and 6.5 oktas)
- Cloud base height given at least 2.5 okta yes/no (equal to or less than 100, 500 and 1000 m above ground)

verified at

- at 6-hourly intervals up to 144 hours from the 00Z and 12Z initialisations.

Temperature and vector wind are verified using a *root-mean-square* (vector) error skill score (RMS(V)ESS) against 24h (observed) persistence. Cloud and precipitation are assessed using the Equitable Threat Score (ETS). Restrictions on manual and automated observations were relaxed as observing practices vary greatly from one country to the next. A montage of selected components is shown in Fig. 2.

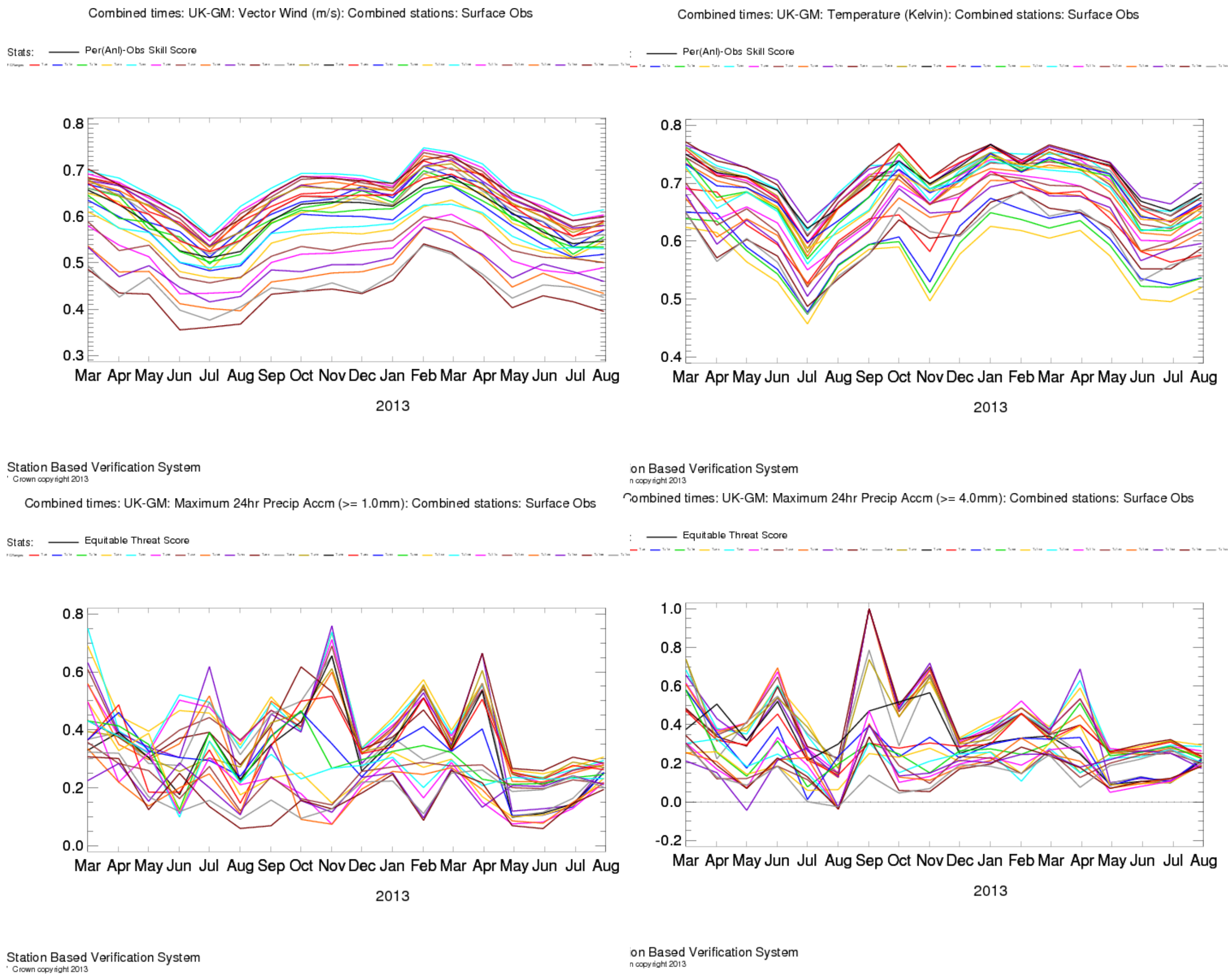

Figure 2: Monthly component scores as a function of lead time.

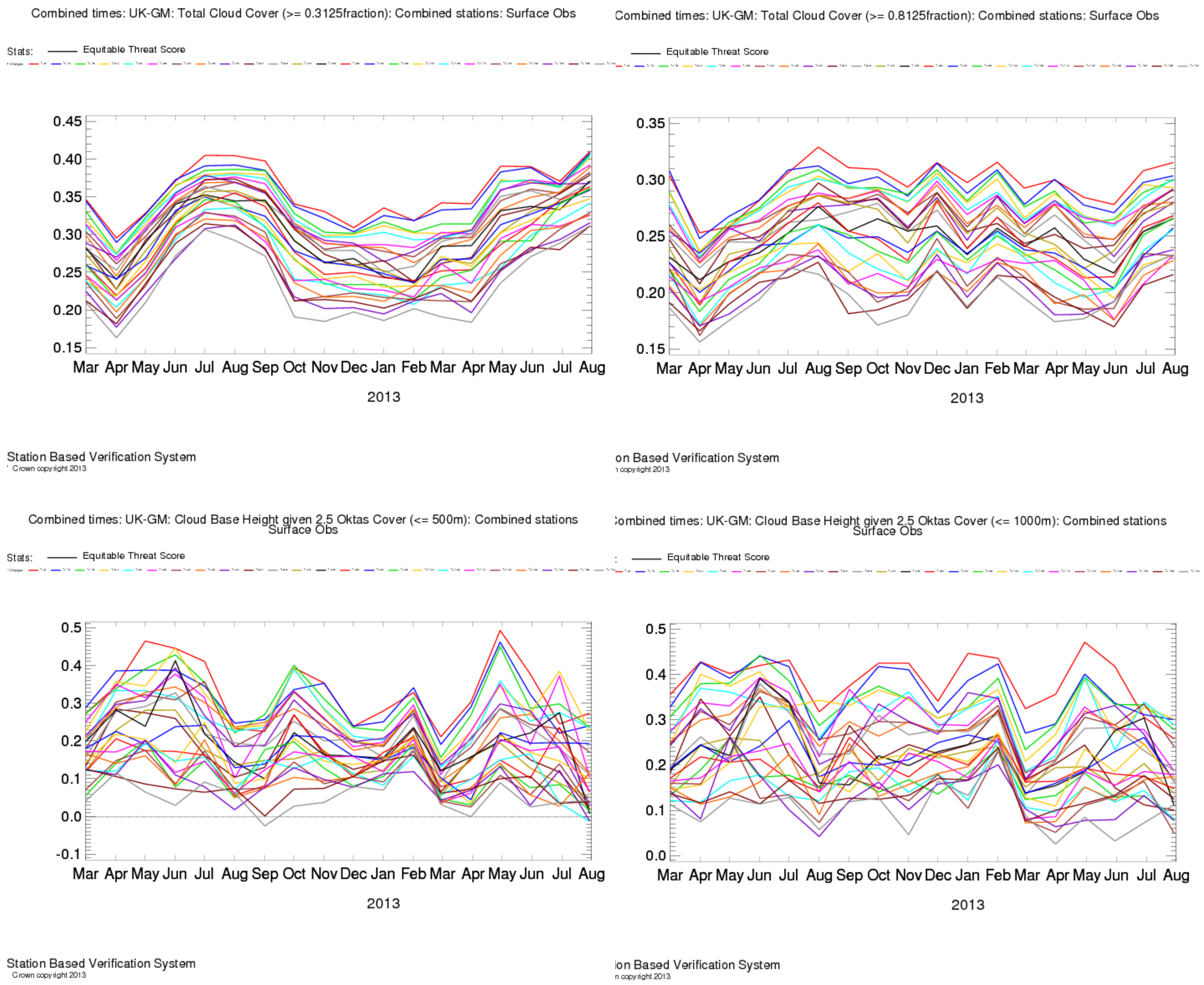

Figure 2 (cont.): Monthly component scores as a function of lead time.

One of the most striking aspects of Fig. 2 is the spread in results and the non-monotonic behaviour in skill for some parameters as a function of lead time. Clearly the t+6h global forecast is not necessarily the best performing, and whilst global data assimilation certainly improves the free atmospheric initial state, surface parameters appear to take some time to respond. For precipitation the first scores are at t+24h. For wind t+30h forecasts have the best scores, temperature forecasts from t+48h, 54h and 60h consistently scored better than earlier lead times. Precipitation performance is very mixed but again earlier lead times are not necessarily better. For the cloud parameters a more monotonic trend emerges. The relatively noisy time series would imply that the in-sample variability was large, and that 139 sites is too few to capture a robust signal.

3. Latest version

An aspect that was felt to be unsatisfactory of the first test was that model developers like to see scores broken down for different parts of the globe; a site list of only 139 stations was deemed to be too small for this purpose. The latest version of global surface verification now divides global observing sites into the traditional CBS regions, utilising all observations received via the GTS. Data quality control is determined by the model data assimilation scheme (surface observations are now assimilated into our global model). For the Met Office this yields 1029 sites for CBS Europe, 1613 in the Northern Hemisphere extratropics, 238 in the Southern Hemisphere extratropics and 348 in the tropics (20N-20S).

Having a fixed threshold for assessing global precipitation was clearly also unsatisfactory, so in the mean time work focused on our verification system being able to differentiate between land and sea points, and the implementation of the Stable Equitable Error in Probability Space (SEEPS). The SEEPS requires a monthly climatology of precipitation accumulations for each site. The implementation included significant improvements to enhance computational efficiency and attempting to make the code more robust, especially in the use of the climatologies. The files provided by ECMWF contain duplicate and triplicate entries, and we have devoted time to ensuring our code deals appropriately with these. We now have Python and Fortran code (for running on the HPC) to calculate the score, lodged in our reviewed code repository. Our latest version of the global surface index, with the new site lists, now has a 24h precipitation component using SEEPS, and some test results are shown in Fig. 3. Six-hourly results are aggregated into daily scores for days 1 to 6 (except precipitation). The results are broadly similar to the first test results, which used a much smaller site list. Total cloud amount and cloud base height show a clear decrease in the ETS with lead time. Temperature and wind on the other hand are still surprisingly non-monotonic, despite the significant increase in sites. The last row shows the SEEPS score for day 3 for all regions as well as the SEEPS score as a function of lead time for the Southern Hemisphere extratropics. This illustrates that the results are more similar as a function of lead time than for different parts of the globe.

Fig. 3: Scores for a selection of parameters and CBS regions

4. Our current UK forecast accuracy metric

In April 2014 we introduced a new UK forecast accuracy metric which is called the “relative index” and measures the value added by the 1.5 km version of the Unified Model over the global model, now at 17 km. The relative index is based on the same parameters listed above, but also includes visibility and considers 6h precipitation instead of 24h. It also only covers the first 36h of the forecast and uses a UK site list. As only the first 36h are used the global update runs at 06 and 18Z are also included (which produce forecasts to t+60h), so all four main runs of the UKV are compared to global. Figure 4 shows the relative index performance showing an average of close to 10% added value of km-scale NWP over global NWP surface parameters.

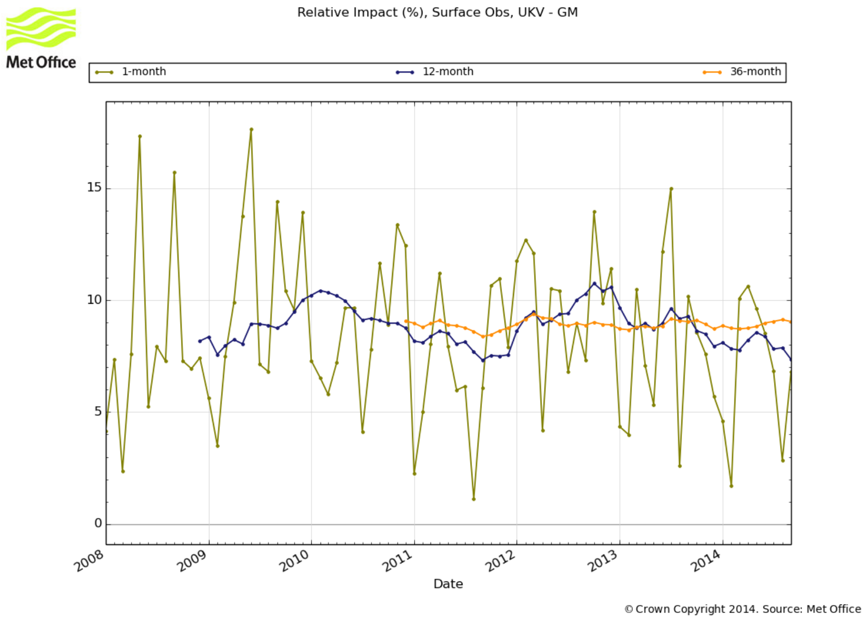

Fig. 4: UK relative index showing added value of km-scale NWP over global over the UK.

5. Concluding remarks

Undoubtedly some further refinements will still be made to the global surface index, not least because significant questions remain. On the whole there is the sense that these results are adding a great deal to our understanding of global model performance. The latest version will now be rolled out to model developers for use with trialling model changes. The back-processing of scores (as far as back as we can go), will commence soon, once we have figured out how to resolve the issues with changing site lists over time.

If the SEEPS were to be adopted as a metric that is exchanged between centres, then the question of who will maintain the climatology files in the future, and who ensures that all centres are using the same files, has to be resolved. The climatologies will need to be version controlled to reflect the closure of sites and the addition of new ones, so that once they have been established long enough to compute a climatology, forecasts can be evaluated at these locations. One thought on reducing the waiting time for including new sites could include using the climatology from an old site, if it was nearby, so that effectively the climatology can be “translated” or “imputed”. The disadvantage of this approach would then be that it requires a continual update of the climatology so that eventually the climatology at the new site would consist of only observations from the site, and not from the old site nearby. This would add a considerable overhead, but in our view, would be highly desirable. It may be though that the definition of “nearby” may be difficult. This needs to be discussed.

Aspects of data quality monitoring and how results can be compared between centres, remains an open question. The SEEPS comparison of a number of global models by ECMWF has been largely successful in that all the results were computed in one place using the same stations and computer code. This approach isn't necessarily sustainable and practical, and centres, like the Met Office, want to be able to calculate these metrics ourselves.