Met Office

MOSAC and SRG Meetings 2015

11-13 November 2015

MOSAC PAPER 20.19

A post-processing and verification strategy for the future

Nigel Roberts and Marion Mittermaier

1. Rationale: why do we need a new strategy?

Very little raw Numerical Weather Prediction (NWP) model output leaves the building- a fact that is often forgotten. The process by which the raw output is turned into a product is therefore very important. It has been shown that despite potentially superior raw NWP forecast skill; good post-processing algorithms can take less skilful raw NWP forecasts and produce superior post-processed forecasts. Such is the power of good post-processing.

Current Met Office post-processing is still too deterministic whilst the whole NWP suite is becoming increasingly probabilistic. Numerical Weather Prediction (NWP) is evolving quickly. Operational centres now run convection-permitting (C-P) models that are able to directly represent the local weather. With this capability has come the realisation that these forecasts still contain significant spatial errors and there is a need to run an ensemble to account for the spatial uncertainty. As a result, the Met Office has introduced its UK 12-member ensemble system MOGREPS-UK. The benefit of more frequently updated short-range forecasts has also been recognised with the move to an hourly cycling UK model. Convection-permitting models are not the only advance; global NWP models (deterministic and ensemble) are being run at ever finer resolution and for longer with the expectation that forecast skill will improve, especially for severe weather or persistent regimes. The new supercomputer at the Met Office will provide the capacity to improve our capabilities still further. This is very welcome, but will ultimately only be beneficial for society if it leads to noticeably better weather forecasts that people can use. The challenge ahead is to find a way to make the best possible use of the information these models provide; which will mean dealing with huge quantities of data from multiple forecasts to generate a wide range of innovative products for a considerable variety of customers. At the same time we need to understand the limitations of our forecasts so that we retain scientific credibility.

It is difficult to speculate where the IT infrastructure will be in 10 years from now and what it will enable us to do, but what has become clear from interactions with other Met Services and private enterprise is that this too is a very fast evolving area with many big players including IBM wanting to provide resource in the Cloud to conduct large experiments to optimise multi-model output for the benefit of the end user. In short, Met Office science needs to be well positioned to capitalise on any advances in IT that may come along for the generation of products.

It is not just NWP models that are advancing; a communications revolution is taking place in society. Individuals and businesses are now able to access weather information at any time of day on mobile computing platforms. To keep up with this change, much greater automation of phone-friendly products is essential, but this must not be detrimental to the high quality warnings of severe weather we provide to the public and government and constantly aim to improve.

This strategy is about reassessing what is required in the Public Weather Service (PWS) arena in the next decade. It is about improving and potentially repositioning our PWS output, and creating a distinction between the PWS and commercial offering. To begin with we need to ask ourselves what PWS needs. For example, we issue temperature forecasts to the nearest whole degree, so do we need to produce PWS temperature forecasts to 0.1 K accuracy? Whilst temperature forecasts should still be an essential part of what we provide, arguably what PWS needs is for us as a Met Service to fulfil our remit of protecting life and property, i.e. providing better warnings of severe weather and extremes. This has to date not been a focus of post-processing, though products such as first-guess warnings relevant for the Flood Forecasting Centre and National Severe Weather Warnings Service (NSWWS) have begun to be produced.

There is need to invest in a post processing system that can embrace these challenges in the next decade and beyond. It needs to be able to accommodate both changes in our NWP modelling system and new ways of disseminating forecast information, including uncertainty. With this in mind the initial draft strategy proposed here covers forecasts out to 14 days that, if implemented, will meet our needs for the next 10 years and beyond. This paper will give an outline of what a new system should look like, focussing on some of the scientific and practical issues that need to be considered

2. A new post processing and verification framework

From all the discussions held with a diverse group of people across the Met Office it is clear that post-processing means different things to different people. Loosely there are three groupings or

requirements: 1) the need for physical corrections; 2) the need for statistical corrections such as bias corrections for example, and 3) product generation or data mining. It is clear that we need all three, and we need improvements and/or enhancements in all three for the benefit of our Science and ultimately PWS.

Figure 1 shows a schematic of a proposed framework, which is seen to be a simple sequential chain with the different types of processing performed in the order shown. Each stage will have plug-and-play modules available and interchangeable, with verification performed before and after to assess the benefit. The framework will be fully probabilistic, using ensemble forecasts and generating probabilities. A definition of what each stage means is given below, with further discussion of some important aspects in subsequent sections. To begin with, we would envisage using a simple set of algorithms and move to greater complexity if new algorithms are cost effective and prove their worth. The key stages are discussed in more detail below.

Physical post processing [Physical]
Post processing that primarily makes use of physical relationships to account for model biases and representativity errors in each of the forecasts. This might be accounting for sub-grid or under-resolved topography or generating new diagnostics (e.g. wind gusts) or making adjustments to variables based on physical relationships to account for biases and representativity errors. Variables may be changed directly at each grid square or a probability density derived to account for uncertainty. This uncertainty or variability could come from sub-hourly observations as well as time-step information from the model.

Statistical post processing [Statistical 1]
This involves post processing that primarily makes use of statistical relationships to account for model biases and representativity errors in each of the forecasts. This might include methods such as Kalman Filtering (KF), Model Output Statistics (MOS), Ensemble Copula Coupling (ECC) or neural networks. Some methods will operate on a whole ensemble. Variables may be changed directly at each grid square or a probability density derived to account for uncertainty. Training data will usually be required.

Neighbourhood post processing [N'hood]
Neighbourhood processing is used to account for spatial uncertainties in forecasts by treating the weather at nearby locations as possible alternative scenarios. It is essential even when an ensemble is used because the ensemble may under-sample the range of possibilities at each location. The outputs from neighbourhood processing are probability forecasts from each of the ensembles.

Blending [Regrid and blend]
This is the process of blending probabilities between different ensembles and forecast lead times. This is discussed more below.

Probabilistic statistical post processing [Statistical 2]
Statistical methods to make the probability forecasts more skilful/useful. This might be re-calibration to make the probabilities more reliable (a 90% probability means an event happened 90% of the time).

Product generation [Gridded], [Text], [Spot]
A variety of products are produced. These may be gridded probabilities or other derived gridded products or forecast text or site-specific forecasts. Some of these products will be for internal use only, for example, to feed into hazard models. The products will be produced on the grids and in the formats required by customers.

Verification <Verify>
This is an integral part of the processing chain and is discussed in further detail below.

3. A fully probabilistic system

A new post processing system will carry the full set of deterministic and ensemble forecasts through the post processing chain and produce probabilistic forecasts at the end of the process. For each variable of interest, probabilities will be computed at every grid square for each ensemble. Neighbourhood processing will be employed to smooth the noisiness associated with under-sampling from a small ensemble. The neighbourhood processing can be applied to each variable independently since physical consistency between variables (e.g. rain and cloud) is maintained in the underlying

ensemble. Joint probabilities can be constructed later if needed. Neighbourhood processing may also be applied to generate probabilities from deterministic forecasts. Time lagging will be utilised, and may be particularly beneficial for short-range forecasts when the UKV is run with hourly cycling. An example of a probabilistic forecast produced from MOGREPS-UK is given in Figure 2.

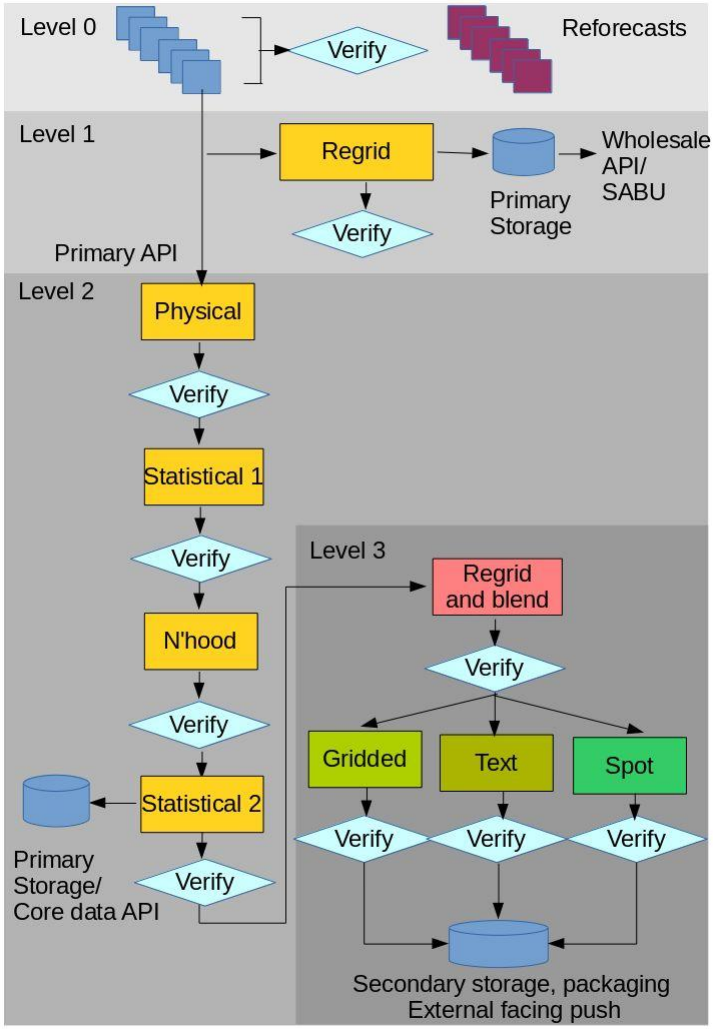

Figure 1: Schematic showing the proposed integrated post-processing and verification chain. Interfaces with the current terminology of "Best Gridded Data", i.e. levels 0, 1 and 2, and with the proposed new IT infrastructure (storage and APIs) are indicated.

4. Blending probabilities

Blending should be an integral part of the post-processing chain but should be used as sparingly as possible (for transparency reasons) and then it should operate in a dynamic way, ensuring that models can be removed or added with ease. Tools such as test harnesses or "testbeds" are required to check the added value of each candidate component of a blend, and candidates should not be included unless they add value, thus minimising the number of components and complexity.

Blending between different ensembles can be achieved by combining probabilities using appropriate weightings. All blending should be done in probability space and will combine the most recent forecast with older forecasts. This will have the benefit of reducing jumpiness from one forecast to the next, whilst still providing regular updates as new forecasts come in. Blending of probabilities will also be necessary to seamlessly blend the end of UK model forecasts into global model forecasts and blend probabilities at the edges of the UK domain with the global forecasts. See Figure 3 for an example of blending probabilities from MOGREPS-UK with probabilities from MOGREPS-G. In principle it would also be possible to allow an operational meteorologist to adjust the probabilities

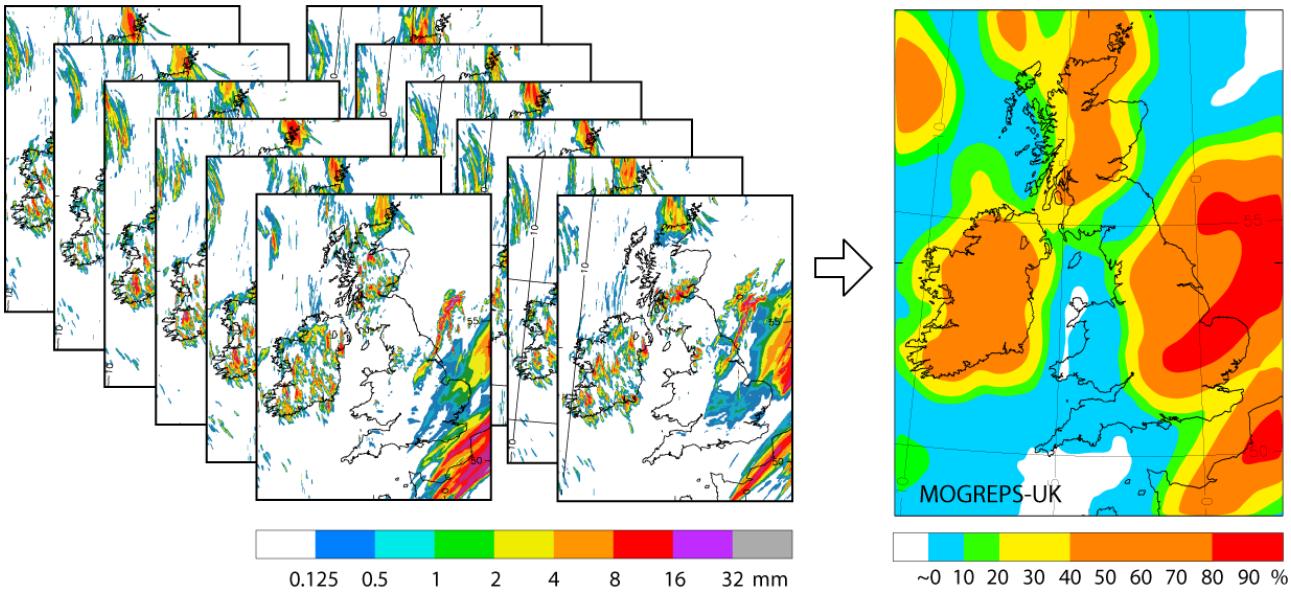

Figure 2: Example of MOGREPS-UK 6-hour rainfall accumulation forecasts (From 03 UTC 31/08/15) and the probabilistic forecast that can be constructed using neighbourhood processing.

Products will be generated from the probability fields. They can either be probabilistic, or made deterministic by using a probability threshold (e.g. 90%) to discriminate between an occurrence or not. The important point is that any determinism is obtained right at the end of the process.

5. A single processing chain for gridded and site specific products

NWP models and nowcasting algorithms are subject to a range of errors. Loosely we refer to them as forecast errors though they consist of several components, including initialisation errors, errors through the boundaries, errors arising from the parameterisation schemes, numerical representation, grid aliasing etc. It is important to note that post-processing methods can help with removing systematic errors and biases, but are less influential for other forecast errors, except when an ensemble, neighbourhood or probabilistic approach is applied. These approaches can account for a range of timing and/or location errors, in keeping with aspects of predictability.

Inconsistencies can come in many guises. For media applications one of the main factors is maintaining a consistency of story. This is difficult when successive model runs flip flop between alternate solutions. This is one of the main reasons for blending, to smooth out the inconsistencies as far as possible. Inconsistencies also arise when different models are stitched together in space and/or time. They can also arise because spatial and site-specific forecasts are not produced together. Another “enemy” of consistency is a deterministic representation of forecasts. A classic example is the apparent flip flopping of weather symbols on the web, despite the use of blending. Currently gridded and site-specific post-processing chains are essentially separate such that the gridded forecast may not match the site-specific forecast at a given location.

The Met Office currently has separate post processing systems for gridded and site specific forecasts. There are good historical reasons for this. When NWP model grid squares were large there was a need to use local topographical information to reduce the representativity error (mismatch between point value and grid square average) at individual locations and account for model biases. Gridded post processing developed independently for the UK as convection-permitting models were introduced. The danger with having two separate systems is that inconsistencies between products emerge, and some of the post processing effort is duplicated unnecessarily. Now that we have a UK model with a 1.5km grid square the representativity error is greatly reduced. Individual weather phenomena such as showers or sea breezes or fog patches are explicitly represented. It now makes sense to have a post processing system that operates on the model grid and extracts site-specific forecasts at the end of the processing chain (based on probabilities). It is still true that even a 1.5km grid is not completely representative of every location within a grid square, especially for mountainous areas, but that variability can still be accounted for using this approach. It is also true that the global model and

MOGREPS-G grid squares are too large to be representative of point locations, but it is likely that this will not be the most significant problem when the forecasts contain large spatial errors. Representativity errors in the global model ensemble can still be accounted for in this framework and as resolution increases the justification for combining the gridded and site specific processing will become even more compelling.

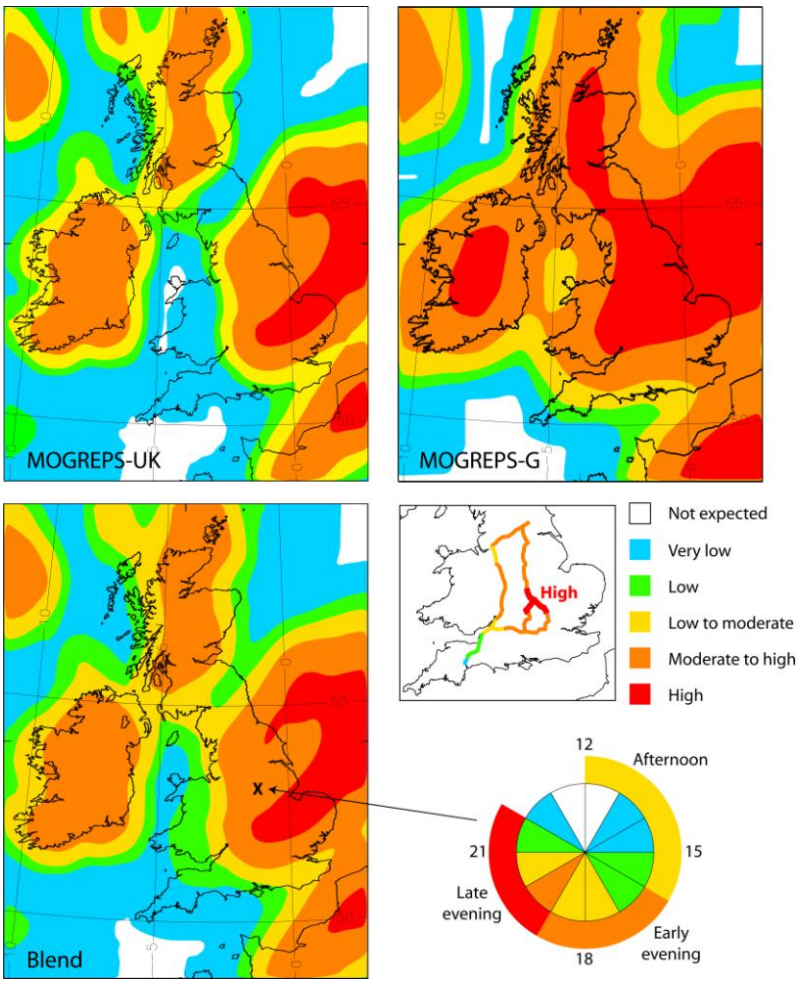

Figure 3: An example of blending probabilities from MOGREPS-UK (top left) and MOGREPS-G (top right) to give the blended result (bottom left) using a weighting of 80% MOGREPS-UK and 20% MOGREPS-G. Bottom right shows how the probabilities can be converted into a route-based or location forecast. The chance of occurrence can be provide for different periods in the day.

6. Verification at every stage

The post processing will have to make use of different methodologies to address different forecast issues. These methods have different levels of complexity and may or may not be compatible with each other. For example, some statistical approaches may require a considerable quantity of training data, whereas others are considerably cheaper. The post processing should deliver improved forecasts using the most beneficial and cost-effective algorithms. We can only make intelligent decisions about which ones to use if there is evidence to show that one out-performs another. A verification step is required between each process in the chain, both to inform which statistical, physical or neighbourhood algorithms work best individually, and also to determine which combinations are the most effective when put together.

The verification must be consistent throughout the system so that one part of the chain can be directly compared with another. Therefore the methodology and metrics must be compatible with the model data at every stage. A probabilistic verification approach will be employed using the same metrics that are used to verify ensembles. The objective is to assess the improvement in forecast skill measured in terms of the ‘reliability’ and ‘resolution’ of the probabilities from the raw ensemble/deterministic

forecasts through to the final probabilistic products. This will be an extension of the HiRA verification framework (Mittermaier 2014)

A difficulty with probabilistic verification is the need for a large sample to obtain meaningful results. This is particularly true if there are few observations and for more extreme weather events. The expectation is that useful signal can be obtained for "ordinary weather". For the more extreme cases a case-study approach is necessary. The verification should be increasingly regime based. Algorithms will be incorporated or developed that objectively detect meteorological scenarios. For example: cyclonic or convective or cloudy or stable conditions. This will provide insight into when and why particular aspects of the post processing, or changes to the underlying model, perform better or worse. These algorithms are also likely to be valuable for statistical methods that take useful predictors from historical forecasts.

7. The distinction between "ordinary" and "extreme" weather

Most current post-processing algorithms act to reduce the magnitude of "extremes" in the raw forecast, as many of these algorithms are based on regression and will draw the outcome towards the mean, smoothing or removing the peaks, which may in fact be correct. Using distributions of values (pdfs), and reforecasts, a truer reflection of the "extreme" can be provided. The proposal is therefore that the Met Office explores, in research mode, a trial of producing retrospective forecasts over the UK, and the temporal frequency one would require to achieve adequate sampling. The potential benefit can then be weighed up against the cost to determine a future strategy. For example it has been shown that it may not be necessary to run reforecasts every day, but only every 4 days. Work by others (Candille et al. 2007) showed that the degrees of freedom are over-inflated because of correlations between forecasts for successive days.

With the greater emphasis on forecasting hazards and associated impacts a repositioning of Public Weather Service requirements is potentially also necessary. The post-processing will produce two types of outputs. The first is the automated products that can go directly to our web site and app or to customers. The second are products that are solely designed to be used by operational meteorologists to help construct the forecast message. Any automated product that leaves the building without human scrutiny should not contain information about extreme weather otherwise there would be a serious risk of public misunderstanding or a spurious signal getting through unchecked. A predetermined set of thresholds are required to partition what is deemed to be high impact as opposed to typical conditions. For example wind gusts > 60mph might be deemed high impact for southeast England and not used as a threshold for automated products, whereas a lower threshold might be used to indicate a windy day. Note that in this situation an automated forecast of a high chance of a windy day would not be incompatible with a warning of strong winds issued by the Operations Centre.

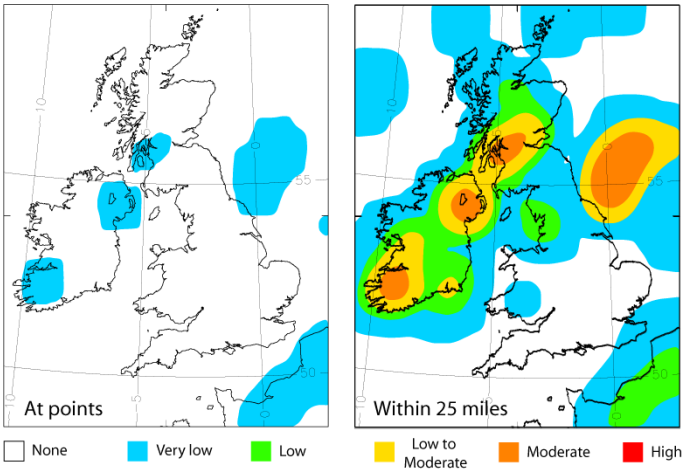

Figure 4: Probability of exceeding 30 mm at each pixel (left) and within 25 miles of each pixel (right).

Probabilistic forecasts of typical weather at a location should have a sufficiently large range (vary from day to day) to be used for decision making (e.g. should I bring an umbrella). However, for more extreme or more localised weather, the probability at any single location is often very low. This is not necessarily because the chance of that event happening is somewhere small, but sometimes just

because the chance of it happening at any particular location is small. Consistently low probabilities are not helpful for producing warnings if they mask the signal that something is likely to happen somewhere. For that reason, probabilities of high-impact, or localised weather conditions will be computed, to give the chance within the area surrounding each grid square, or within regions of interest. An example of this is shown in Figure 4, in which the chance of rainfall > 30mm is moderate somewhere in central Scotland or Northern Ireland even if very small in any particular place.

8. An Operations Centre dashboard

As NWP models go to higher resolution, ensemble get bigger and forecasts are run more frequently for longer on larger domains, it is becoming increasingly difficult to quickly extract the most salient information. It will soon become too time-consuming to trawl through every variable in every ensemble member every time a new ensemble is run. Summary information is required to highlight the main weather signals in the forecasts to help operational meteorologists quickly focus on what matters. Using the probabilistic framework, a summary dashboard can be constructed that provides an indication of the likelihood of different types of weather. An example of how this might look is shown in Figure 5. The dashboard contains an indication of the chance of both ordinary and severe weather for the whole UK. More detail about the geographical variation and specific weather can be obtained by "tunnelling down", as shown in this example with the coarse grained maps of UK regions. A route through to the raw ensemble forecasts should be available.

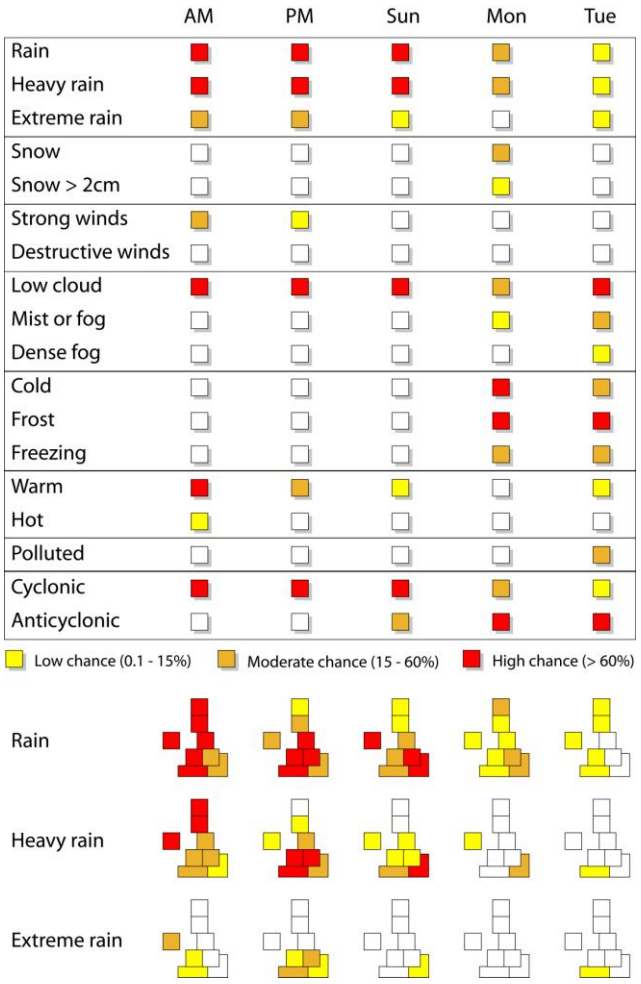

Figure 5: An example of a top level "screen" for the Ops Centre, highlighting aspects of the forecast that the operational meteorologists should focus on. Ideally this would be an interactive screen with the capability of clicking on any of the colour boxes to get further information.

9. Requirements and predictability

If weather forecasts are going to be both accurate and useful, there must be an understanding of both the user requirements and our forecast capability. We should not try to forecast beyond the limits of predictability. A forecast that it will rain at 2pm at a particular location in 5 days time is very likely to be wrong and therefore worthless (except as entertainment!). We need to know the limits of predictability and make sure forecasts do not convey more accuracy than they really have. Again, this means

moving to more probabilistic forecasts and presenting information as the chance of occurrence within time windows or spatial areas (e.g. figures 3 and 4). It will require verification of the spatial and temporal detail that can be used before skill is lost. A forecast of a high chance of rain this afternoon or ground frost tomorrow tonight is helpful even though it isn't delivered in hourly intervals. The understanding of predictability goes hand-in-hand with an understanding of the level of detail and accuracy that is expected.

10. Research and development areas

An initial R&D activities list is summarised in Table below. Activities are grouped by time scales, based on how much time would elapse before a benefit will be seen (subject to the investment in terms of funding). Activities in different rows could be done in parallel, again subject to investment.

1-3 years	*3-5 years*	*Beyond*
Underlying software developments	Software infrastructure consolidation and maintenance	Software infrastructure consolidation and maintenance
Develop neighbourhood processing for all variables.	New physical and statistical methods. Advanced neighbourhood methods.	Further new methods and enhancements
Basic spatial blending and calibration of ensembles, evaluate the value of reforecasts	Further development of spatial blending and calibration	Enhancements
Investigate new ways of extracting and conveying probabilistic information, inc text generation	Introduce new ways of conveying probabilistic information	Continued development of probabilistic products
Verification methods, inc spatial ensemble methods	Regime-based verification and post processing	Verification of predictability limits
Basic dashboard for Ops Centre Consultation and development	Enhancements to dashboard and maintenance. Interactive.	Enhancements to dashboard and maintenance
Survey of public requirements Scoping of required diagnostics (including sub hourly)	Exploring the use of sub-hourly variability in the model and observations	Continued monitoring of requirements

11. Conclusions and recommendations

A new post-processing and verification framework is proposed which is in keeping with the overall scientific strategic direction of the Met Office. Investing in implementing this strategy would lead, over a period of time, to a step-change in capability in this essential and crucial area of Science. We invite the committee to provide comments and feedback, and if possible an endorsement on the proposed:

- Single processing chain, grid-based and probabilistic;
- A repositioning of the PWS offering, with a shift in emphasis towards the development of both products for high-impact weather and rapid refresh automation for "ordinary" weather; and
- Research and development goals in the short- to medium-term.

12. Acknowledgements

Many thanks to all those who have contributed in the 1-2-1s and workshops. Also a particular thank you to the virtual project team: Piers Buchanan, Teil Howard, Nina Schuhen and Helen Titley, as well as Bruce Wright, Simon Jackson and Andrew Bennett.

13. References

Candille G., C. Côté, P. L. Houtekamer, and G. Pellerin, 2007: Verification of an Ensemble Prediction System against Observations. *Mon. Wea. Rev.*, **135**, 2688–2699.

Mittermaier, M.P., 2014: A strategy for verifying near-convection-resolving forecasts at observing sites. *Wea. Forecasting*. **29(2)**, 185-204.