Met Office

Evaluation of Precipitation in EURO4M Reanalyses and Downscalers

Forecasting Research Technical Report 589

January 28, 2014

Peter Jermey and Richard Renshaw

Abstract

The European Reanalysis and Observations for Monitoring project has produced two reanalyses and two downscalers which include high resolution datasets of precipitation across the European domain. Representation of precipitation by the four models is compared to the global reanalysis ERA-Interim, which is considered the current gold-standard for reanalysis. The quality of the datasets is assessed via the points-based equitable threat and the new stable equitable error in probability space scores and also via the spatial fractions skill score. The datasets are also assessed on their ability to represent a number of monthly statistics useful for climate monitoring.

Inclusion of precipitation observations, as in the downscalers, greatly improves representation of precipitation at all scales and thresholds. Increasing resolution also improves representation of precipitation at higher thresholds. Four dimensional variational assimilation is better able to capture small scale events than three dimensional variational assimilation.

1 Introduction

Precipitation is a critical aspect of the climate. However, representation of precipitation in gridded datasets includes a number of challenging aspects that do not occur when considering fully continuous variables, such as temperature or pressure, which vary smoothly over a larger spatial scale. Precipitation events occur over a wide range of intensities, periods and spatial scales (25mm/minute localised flash-flooding to multi-year droughts covering several countries). Europe is less affected by long period events than other regions, but suffers from fluvial flooding and flash-flooding, in which very high rainfall occurs over a small domain and in a short period. Although such events may be representable by gridded datasets, localised extremes are usually softened by grid-box averaging and insufficiently short accumulation periods [King A.D. et al.(2012)].

The European Reanalysis and Observations for Monitoring project (EURO4M) has produced several gridded datasets of precipitation over Europe including two reanalyses and two downscalers. A High Resolution Limited Area Model (HIRLAM) reanalysis has been produced at 25km for 1993-2013 using 3DVAR (three-dimensional variational) data assimilation, [Dahlgren P. and Gustafsson N.(2012)]. This HIRLAM reanalysis has been downscaled at the surface to about 5km and combined with surface observations using two different optimal interpolation schemes (OI) to produce two higher resolution datasets for surface variables (MESAN and MESCAN) [Haggmark L. et al.(2000)]. Finally there is the 12km resolution Met Office reanalysis (MO), covering 2008-2009, using 4DVAR (four-dimensional variational) data assimilation [Rawlins F. et al.(2007)]. The project has also contributed to gridded observation datasets of precipitation including the Climate Research Unit at the University of East Anglia dataset (CRU/UEA) at roughly 60km resolution, covering 1901-2011 [Harris I. et al.(2013)], E-OBS at 25km resolution, covering 1950-2013 [van den Besselaar E.J. et al.(2011)], and the 60km Hamburg Ocean Atmosphere Parameters and Fluxes from Satellite Data (HOAPS/GPCC) gridded dataset which covers 1986-2006 [Andersson A. et al.(2010)].

Gridded observation datasets vary in quality. Satellite data is accurate over oceans, but less so over land.

Rain gauge data is generally more accurate over land, but requires strict quality control [Lopez P.(2013)] and a high-density network is required to produce an accurate gridded dataset. Reanalyses are useful for studying climate change and climate monitoring since they provide spatially continuous records of the atmosphere for lengthy periods that are consistent with both observations and physical laws [Kallberg P.(2010)]. Global reanalyses have been used to drive high resolution atmospheric models [Thieblemont R.Y.J. et al.(2013)], regional climate models [Chan S.C. et al.(2013)], hydrological models [Akhtar M. et al.(2009)], providing meteorological inputs to flood hazard mapping [Salamon A.L. et al.(2013)] and to study regional water storage changes [Yeh P.J.-F. and Famiglietti J.S.(2008)], amongst others. All of these applications could benefit from high resolution regional reanalyses/downscalers as provided by EURO4M. Many applications that currently use the relatively high density observations, that are available in the European domain, may also benefit from high resolution reanalysis/downscaler data. Given the variety of applications, representation of precipitation in the reanalyses/downscalers is evaluated here using a number of different measures.

These evaluation measures use rain gauge data as 'truth', using six or twenty-four hour accumulations for comparison. Methods for calculating these accumulations from the models are described in Section 2. Section 3 briefly describes the reanalysis/downscaler datasets and their annual error and bias is discussed in Section 4. Location-based verification is provided by the traditional equitable threat score (ETS) calculated at observation station positions in Section 5.1, and in Section 5.4, via the new stable equitable error in probability space (SEEPS) score [Rodwell M.J. et al.(2010)] which takes into account local climatology, developed at the European Centre for Medium-Range Weather Forecasts (ECMWF). Location-based verification can mask the improvements in precipitation representation due to high resolution [Mass C. et al.(2002)] and so spatial accuracy is also assessed via the fractions skill score (FSS), [Roberts N.M. and Lean H.W.(2008)], in Section 6. Representation of precipitation climate monitoring statistics in the reanalyses/downscalers are evaluated via coefficients of correlation between the reanalyses/downscalers and observation-based statistics in Section 7. Conclusions are in Section 8.

Prior to the EURO4M project, the state-of-the-art reanalysis over Europe was ECMWF's Re-Analysis Interim (ERA-Interim) [Dee D.P. et al.(2011)]. This is a highly accurate 80km reanalysis, out-performing satellite data in some studies [Pena-Arancibia J.L. et al.(2013)]. ERA-Interim provides essential boundary conditions to the EURO4M regional reanalyses. It is also used here as a challenging baseline against which to evaluate the higher-resolution regional reanalyses and downscalers.

The models are also compared to MO-variants. In the downscaler variant (MO-D), data assimilation is not included and so the dataset is simply the ERA-Interim reanalysis reconfigured to MO resolution, using the forecast model to produce precipitation accumulations. In the climate run variant (MO-C), cycling is also removed so that a single long forecast covers the entire period. In the 3DVAR variant (MO-3DVAR) persistence is used instead of a forecast in the assimilation effectively turning the system into a 3DVAR scheme.

SYNOP (surface synoptic observation) stations regularly provide weather prediction centres with rain gauge data. None of the reanalyses (ERA-Interim, HIRLAM or MO) assimilate these data, and so they provide a

useful independent measure of 'truth'. The MESAN and MESCAN downscalers use optimal interpolation (OI) to combine twenty-four hour precipitation accumulations from HIRLAM with rain gauge data and therefore rain gauges are not an independent measure for the downscalers.

2 Accumulations

Reanalysis systems incorporate numerical forecast models. These typically suffer from spin-up/down of precipitation rates i.e. for a short period at the start of the forecast there is systematically too little or too much rainfall. In ERA-40, the predecessor to ERA-Interim, a forecast offset is recommended to provide accurate accumulations of precipitation [Kallberg P.(2001)]. In ERA-Interim the first six hour accumulation has been found to be the most accurate, [Kallberg P.(2011)], and this is also the case for MO, see Appendix E. The first six hour accumulation is also considered relatively spin-up/down free in the HIRLAM analysis if taken from the 00Z or 12Z (re)analyses (which assimilate more observations than the 06Z or 18Z (re)analyses). The six hour accumulations used for evaluation of the three reanalyses are therefore the difference in accumulation from the 00Z (12Z) (re)analysis to the 06Z (18Z) forecast to enable a clean comparison to take place. The MESAN and MESCAN downscalers only produce twenty-four hour accumulations and so no comparison of six hour accumulations is possible.

The daily accumulation (24hr) on day x is defined as the accumulation between 06Z on day x until 06Z on day $x + 1$. For ERA-Interim, which has 00Z and 12Z (re)analysis times, this is calculated as the sum of the accumulations from twelve hour forecasts on day x. Daily accumulations from MO are calculated as the six hour forecast accumulations from 06Z, 12Z and 18Z on day x and from 00Z on day $x + 1$ minus the associated (re)analysis accumulations. MO has non-zero (re)analysis accumulations since its 4DVAR assimilation window starts three hours before the (re)analysis time. Daily accumulations are a product from the MESAN and MESCAN downscalers so no further manipulation is necessary.

3 Models

The five datasets compared here, ERA-Interim, HIRLAM, MESAN, MESCAN and MO, combine forecasts with observation data to estimate the atmospheric state. However, large differences may occur between the datasets and this is particularly so with precipitation which is not directly assimilated into ERA-Interim, HIRLAM or MO. These discrepancies may be due to differences in the assimilated observation sets, resolution, assimilation method and/or the forecast model used. A summary of the principal differences that are expected to affect precipitation representation is given in Table 1.

In the above table IFS is ECMWF's integrated forecast system and UM is the Met Office Unified Model.

4 Annual Error & Bias

ECMWF produced gridded twenty-four hour precipitation accumulations across most of the land area of Europe for 2002-2009, [Ghelli A. and Lalaurette F.(2000)]. This data is available on a 0.225 degree grid (approximately

Table 1: Summary of datasets.

Dataset	Resolution	Assimilation	Sat. Obs	Gauges	Model
ERA-Interim	80km	4DVAR	Y	N	IFS
HIRLAM	22km	3DVAR	N	N	HIRLAM
MESAN	5km	OI (approx. 270km)	N	Y	downscaled from HIRLAM
MESCAN	5km	OI (approx. 35km)	N	Y	downscaled from HIRLAM
MO	12km	4DVAR	Y	N	UM

30km) with each grid-box averaging data from several rain gauges to produce rainfall estimates more representative of the entire grid-box creating an observation dataset comparable with model data. Using this dataset as 'truth', RMSEs of twenty-four hour precipitation accumulations from ERA-Interim and the four EURO4M datasets has been calculated at 30km scale for 2008 and the results are displayed in Figure 1. Remapping of reanalysis/downscaler data from the native grid to the observation grid is carried out using the climate data operators (CDO) [Akhtar M. et al.(2011)]. This figure also displays the lower and upper bounds on the number of observations used per grid-box showing no data for Portugal, Spain and Sweden for at least some of the period and that, across many countries, as few as one observation may be used per grid-box. Grid-boxes containing single observations are unlikely to produce data that is representative of the whole grid-box. The most reliable values are over England, France and Germany where most of the grid-boxes contain at least three observations. Norway provided pre-gridded data for this project, see [Vormoor K. and Skaugen T.(2013)], and the number of stations used in this region is unknown. ERA-Interim, HIRLAM and MO have no assimilated precipitation data, but the downscalers MESAN and MESCAN include precipitation observations, likely to include the majority of those in the ECMWF gridded observation dataset.

Figure 1 demonstrates that the two EURO4M reanalyses, HIRLAM and MO, have a similar annual RMSE to ERA-Interim at this resolution (30km). The two EURO4M downscalers show marked improvement to ERA-Interim over France and Germany, where the rain gauge data is most dense, and show similar performance elsewhere. All five models show that the greatest RMSE occurs near the mountainous regions of the Alps, which is itself excluded from the observation dataset. However the 'truth' is also likely to be least accurate in these regions since variation in rain gauges will be less representative of the grid-box than in more level terrain.

The difference in RMSE between the EURO4M datasets and ERA-Interim is shown in Figure 2 which again demonstrates EURO4M models have annual RMSEs similar to that of ERA-Interim. HIRLAM has a slightly higher RMSE than ERA-Interim throughout most of the domain. MO has a generally lower RMSE throughout most of continental Europe with small areas where its RMSE is higher. The downscalers MESAN and MESCAN show considerable reduction in RMSE over ERA-Interim in France and Germany and perform similarly to MO elsewhere. RMSE is higher in MO, MESAN and MESCAN near coastal areas. This is likely to be a feature of remapping to the observation resolution, which is finer than ERA-Interim, but coarser than the EURO4M datasets. Calculating spatial averages of RMSEs across France and Germany, where the observation data is most trustworthy, yields 0.4mm for ERA-Interim and MO, 0.5mm for HIRLAM and 0.3mm for MESAN and MESCAN. This typical RMSE is not a very useful measure for comparing the systems since it assumes, for example, that the error between 10 and 11mm precipitation is as important to the user as the error between 0 and 1mm. It also gives no measure of extremes, does not assess the spatial distribution of precipitation and assigns better

Figure 1: *Twenty-four hour precipitation RMSE of reanalyses/downscalers compared to ECMWF gridded observation data for 2008. ERA-Interim (top), HIRLAM (L, middle-top) and MESAN (R, middle-top), MESCAN (L, middle-bottom) and MO (R, middle-bottom). The bottom row shows minimum(L) and maximum(R) number of observations used per grid-box.*

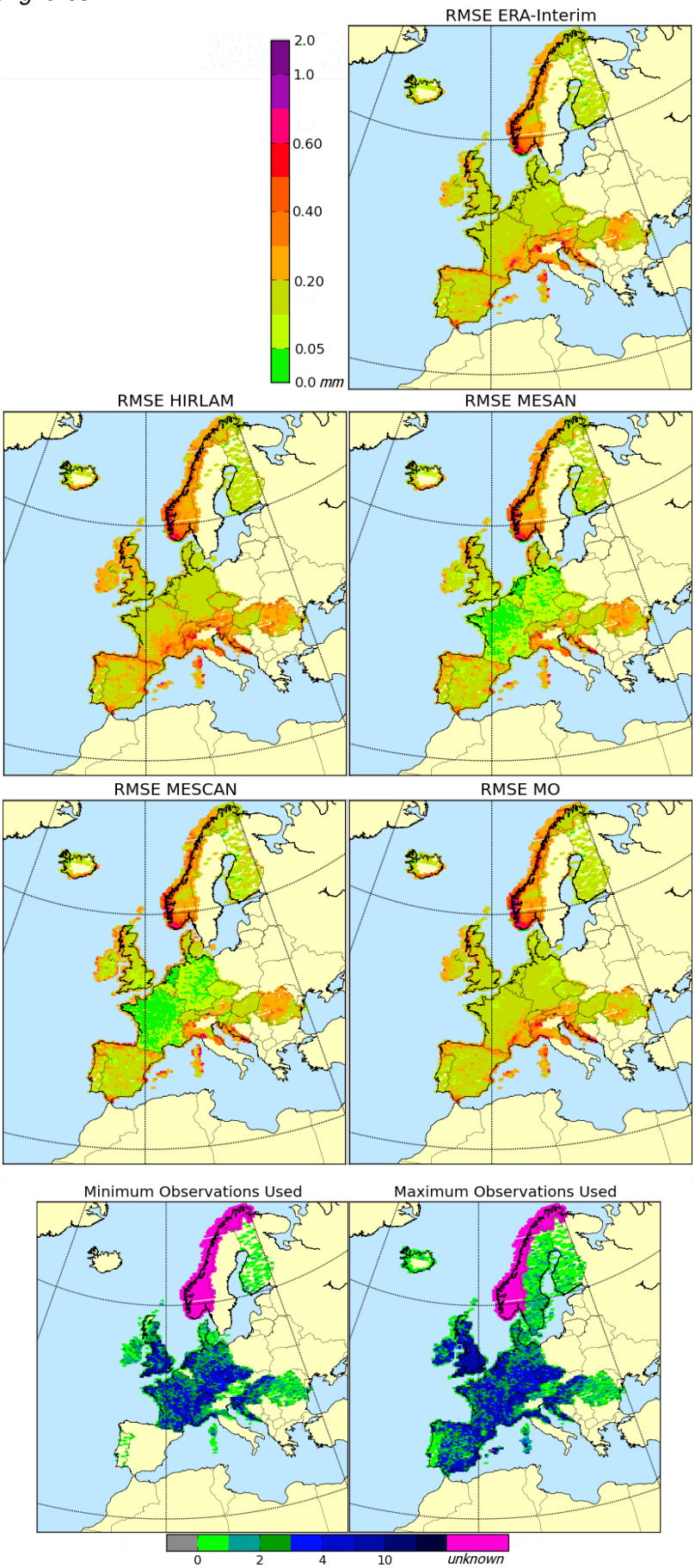
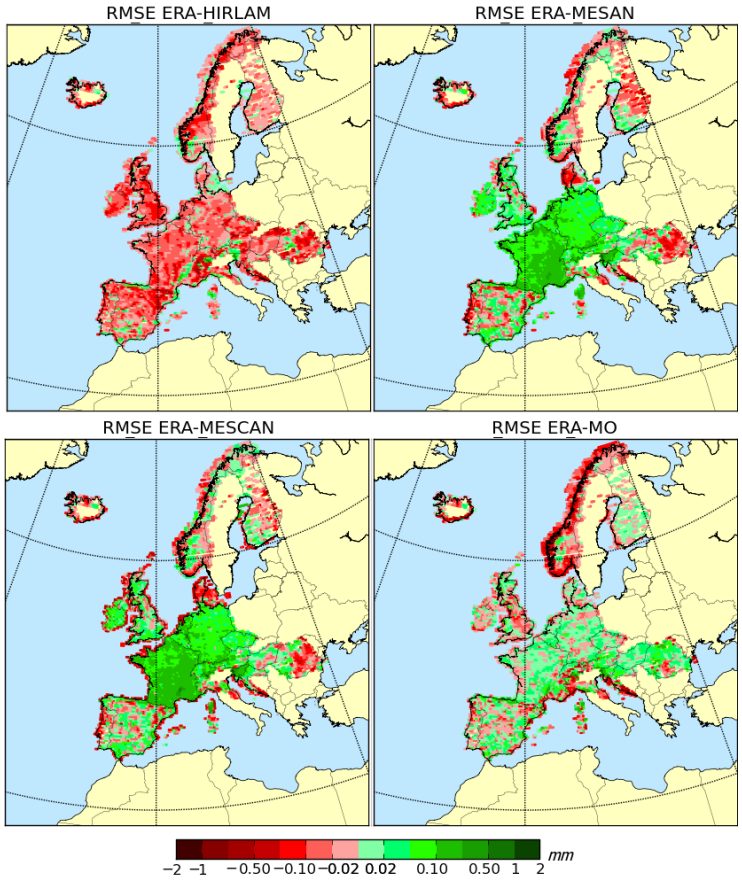


Figure 2: *Difference in precipitation RMSE of ERA-Interim compared to EURO4M reanalyses/downscalers for 2008. Positive or negative values indicate EURO4M has lower or higher, respectively, RMSEs than ERA-Interim. HIRLAM (L, top) and MESAN (R, top), MESCAN (L, bottom) and MO (R, bottom).*



scores to drier regions. However, since the results are similar, the typical RMSEs do indicate reasonable consistency between the different models.

Figure 3 shows the mean errors of each of the reanalyses/downscalers. The reanalyses, ERA-Interim, HIRLAM and MO, have regions of over-wetness and under-wetness. All three have larger areas of over-wetness than under-wetness, with HIRLAM the most biased and ERA-Interim the least biased in this way. The bias is lower in the downscalers MESAN and MESCAN, particularly over France and Germany.

5 Location-based Evaluation

5.1 Equitable Threat Score - six hour accumulations

The ETS is a commonly used measure to evaluate model skill at representing precipitation at specific points, comparing the gridded data with station-based observations [Gilbert G.K.(1884)]. This score evaluates the ability of the model to represent events where precipitation occurs above a certain threshold, adjusted according to its ability to predict where no event occurs (i.e. 'correct negatives'). ETS is described in more detail in Appendix A. Using six hour accumulations, HIRLAM and MO ETS are compared with ERA-Interim ETS for each month in the period 2008-2009 at thresholds 0.5mm, 1mm and 4mm. The threshold 0.5mm/6hr is indicative of 2mmday^{-1} which is approximately the average daily rainfall from European observation stations during the period, 1mm/6hr is indicative of 4mmday^{-1} which is slightly less than the mean precipitation on wet days in the observations and 4mm/6hr is indicative of 16mmday^{-1} which is approximately the average of the monthly one-day maximum precipitation in the observations. Therefore the thresholds 0.5mm, 1mm and 4mm indicate light, medium and heavy precipitation, respectively. ETS are also compared for the more extreme thresholds 8mm and 16mm. The results are displayed in Figure 4. The MESAN and MESCAN datasets comprise twenty-four hour accumulations only and cannot be included in this comparison.

The ETS of the three reanalyses is dependent on season, featuring high ETS in winter and lower ETS in summer. This pattern is expected in Europe since in winter weather patterns are dominated by large scale structures such as the North Atlantic oscillation which are easier to represent than the small scale structures that are the main cause of precipitation in the summer [Zolina O. et al.(2004)]. As expected the high resolution 4DVAR MO has consistently higher ETS than both ERA-Interim and HIRLAM. Perhaps unexpectedly, ERA-Interim has similar ETS as HIRLAM for these thresholds. Although HIRLAM features higher resolution than ERA-Interim, it is well known that location-based verification does not necessarily show the improvement in precipitation through higher resolution alone, see [Done J. et al.(2004)], etc. Therefore, since HIRLAM also features fewer levels than ERA-Interim and uses 3DVAR, it is perhaps not surprising that its ETS are not significantly higher than ERA-Interim. At the highest threshold, HIRLAM does show some improvement over ERA-Interim, demonstrating that the increase in resolution allows better representation of extremes.

Typically the difference in ETS between ERA-Interim and MO is 0.04, 0.04, 0.07, 0.10 and 0.11 for the 0.5mm,

Figure 3: *Twenty-four hour precipitation mean error of reanalyses/downscalers compared to ECMWF gridded observation data for 2008 (observation-minus-model). ERA-Interim (top), HIRLAM (L, middle) and MESAN (R, middle), MESCAN (L, bottom) and MO (R, bottom).*

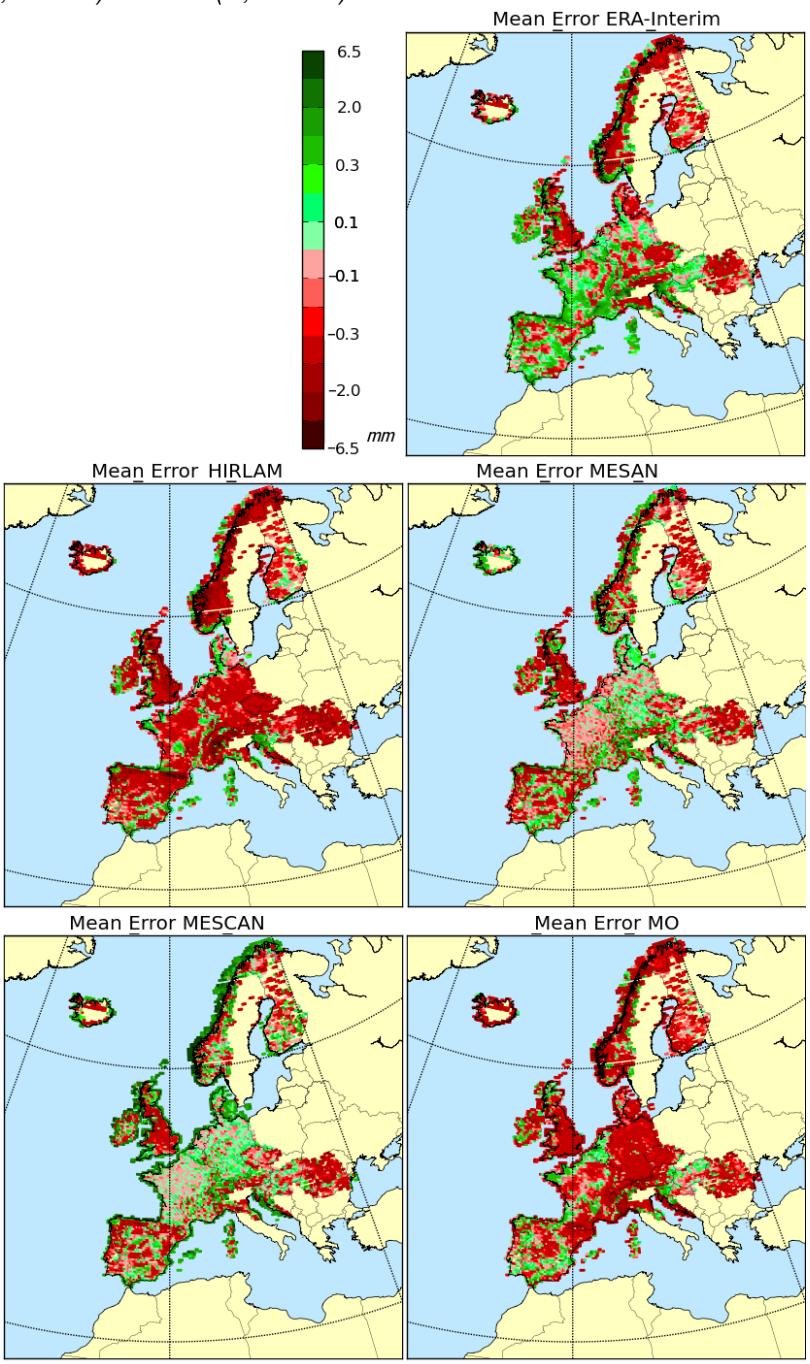
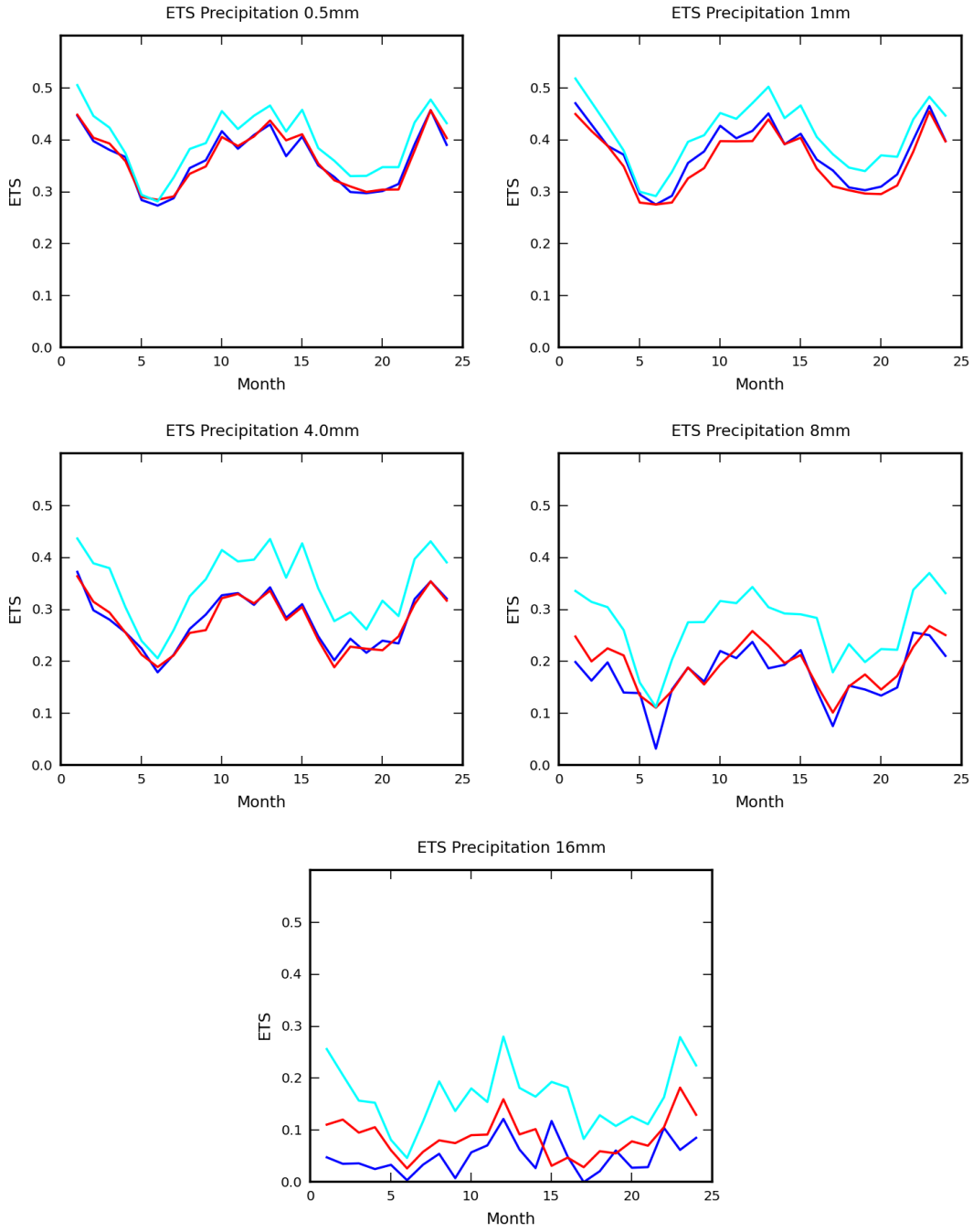


Figure 4: *Monthly Equitable Threat Scores 2008-2009 for ERA-Interim (darker blue), MO (lighter blue) and HIRLAM (red) reanalyses (6hr accumulations) at thresholds 0.5mm(L, top), 1mm(R, top), 4mm (L, middle), 8mm (R, middle) and 16mm (bottom).*



1mm, 4mm, 8mm and 16mm thresholds, respectively. Since the two systems both assimilate similar observations via 4DVAR, the majority of this score increase is likely due to increased resolution in the limited area model. Experiments using a 3DVAR version of MO and a version with no assimilated satellite data show that the increase in skill due to 4DVAR (over 3DVAR) is typically 0.03, 0.03, 0.03, 0.03 and 0.02 and due to satellite data 0.001, -0.01, -0.01, 0.01 and 0.0 for the 0.5mm, 1mm, 4mm, 8mm and 16mm thresholds, respectively. Therefore the inclusion of 4DVAR over 3DVAR has a similar impact to increasing the resolution from 80km to 12km. HIRLAM typical scores are similar to those of ERA-Interim, differing by 0.002, -0.01, -0.004, 0.02 and 0.04, for the 0.5mm, 1mm, 4mm, 8mm and 16mm thresholds, respectively.

Precipitation values from the reanalyses are grid-box averages. Therefore if a large amount of localised rain falls at a station its model grid-box equivalent is expected to be lower, spread throughout the box. This effect causes over-prediction of low threshold events and under-prediction of high threshold events, but increasing resolution should reduce this effect. The effect is demonstrated in Figure 5 which displays the frequency bias for the three thresholds.

The frequency bias is a ratio of the number of events modelled to the number of events observed, see Appendix A, so that values greater than one indicate over-representation in the model and values less than one indicate under-representation in the model. As expected when comparing gridded data to observed precipitation, the models over-represent the lower thresholds. At these lower thresholds HIRLAM is the least biased. The representivity decreases with increasing threshold so that, at higher thresholds ERA-Interim and HIRLAM are under-representative. At these higher thresholds MO is least biased, indicating that this, the highest resolution reanalysis, is best able to produce realistic levels of precipitation at the highest thresholds.

5.2 Equitable Threat Score - twenty-four hour accumulations

To compare the twenty-four hour precipitation accumulation products from the downscalers MESAN and MESCAN with the reanalyses, ETS of twenty-four hour accumulations are displayed in Figure 6. Figure 6 shows that MESAN and MESCAN perform much better than the other datasets at all thresholds. Although observations used here as truth are assimilated into the downscalers, and therefore are not independent, these high scores demonstrate the improvement possible through direct assimilation of precipitation observations. The ECMWF gridded observation dataset discussed in Section 4 is used as truth.

Figure 6 also shows twenty-four hour ETS for the four EURO4M datasets and ERA-Interim. Again the high resolution 4DVAR MO has consistently higher ETS than ERA-Interim and HIRLAM. HIRLAM performs worse than ERA-Interim for lower thresholds and better for higher thresholds, indicating that the higher resolution is beneficial in representing extreme events. The two downscalers, MESAN and MESCAN, out-perform the reanalyses, demonstrating the increased quality of representation available through precipitation assimilation. MESCAN, which concentrates on smaller scales than MESAN, out-performs the other downscaler at the highest threshold, which is associated with the smallest scale events.

Figure 5: *Frequency Bias 2008-2009 for ERA-Interim (darker blue), MO (lighter blue) and HIRLAM (red) reanalyses (6hr accumulations) at thresholds 0.5mm(L, top), 1mm(R, top), 4mm (L, middle), 8mm (R, middle) and 16mm (bottom).*

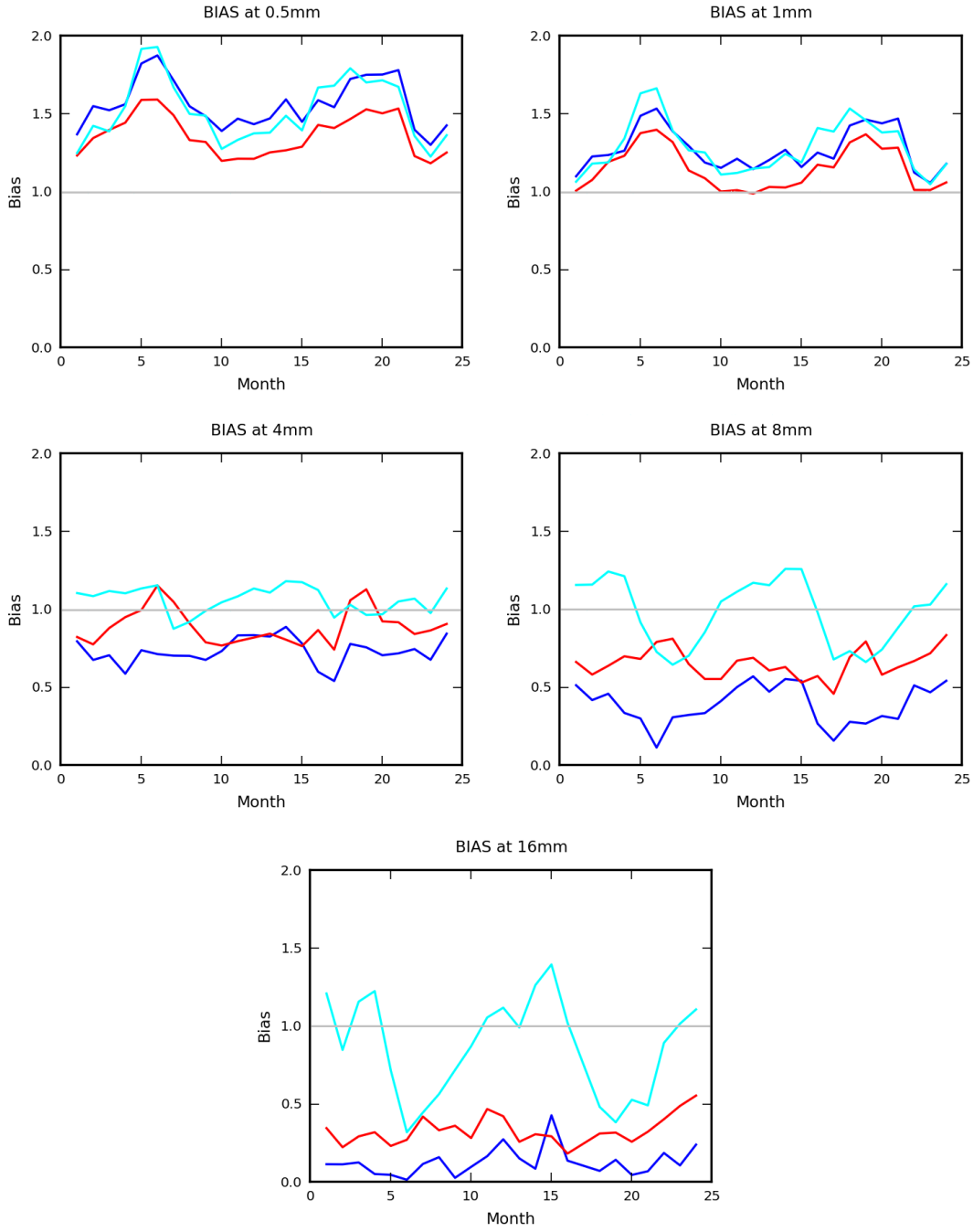
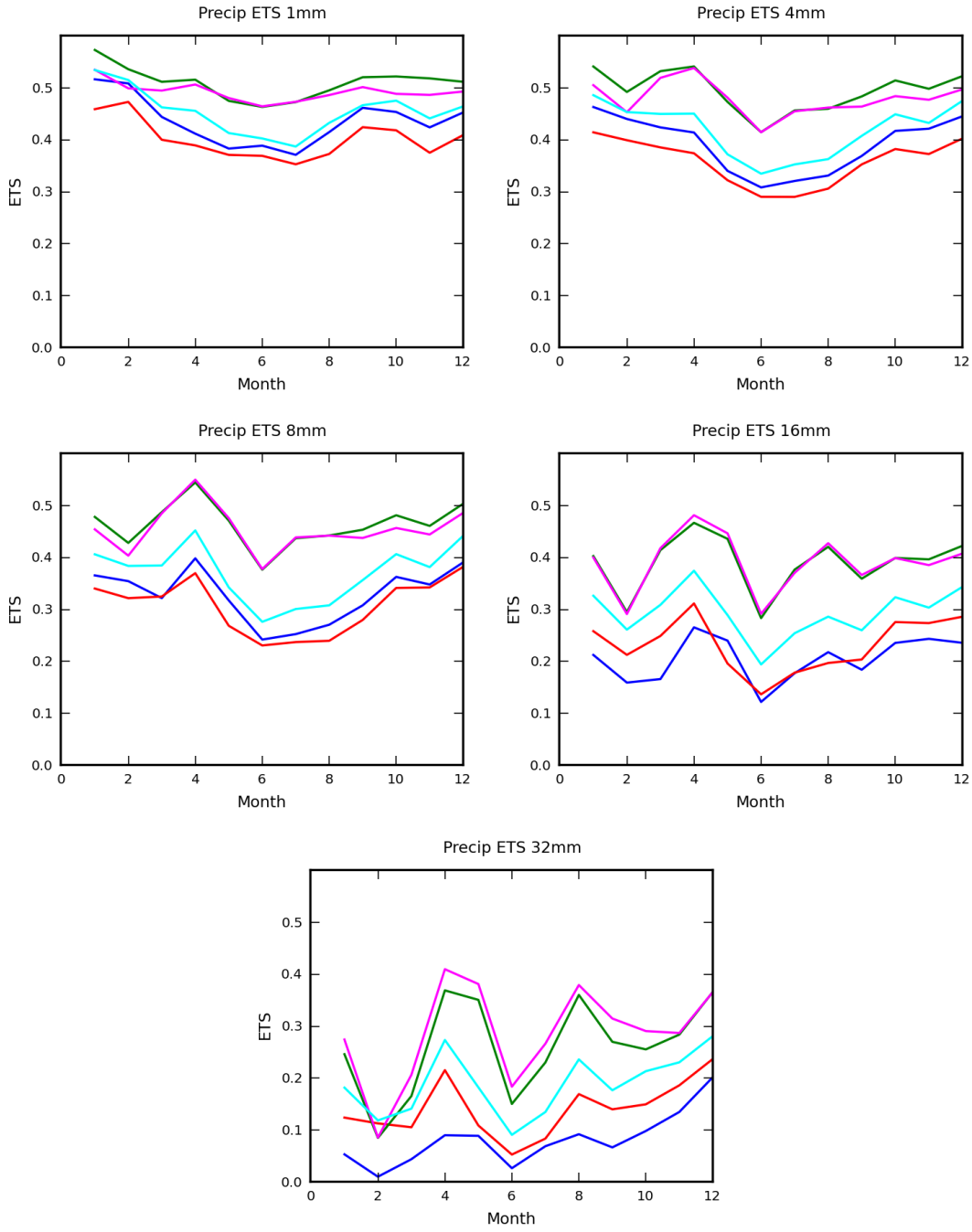


Figure 6: *Monthly Equitable Threat Scores 2008 for ERA-Interim (darker blue), MO (lighter blue) and HIRLAM (red) reanalyses (24hr accumulations) at thresholds 1mm(L, top), 4mm(R, top), 8mm (L, middle), 16mm (R, middle) and 32 (bottom).*



Twenty-four hour bias frequencies are shown in Figure 7. Again at lower thresholds the reananlyses are over-representative. This effect is greatly reduced by the downscalers. At high resolutions MO and HIRLAM are over-representative, but ERA-Interim is under-representative.

5.3 Met Office Downscaler and Climate-style runs

Figure 8 displays ETS for the EURO4M datasets, together with MO-C and MO-D. As expected, at most thresholds, MO-C performs much worse than the reanalyses, demonstrating that climate datasets cannot locally represent variables at relatively short temporal scales. At the highest threshold, however, MO-C is of similar quality to ERA-Interim. This suggests that high resolution models are required to accurately represent intense precipitation events. MO-D generally out-performs ERA-Interim and HIRLAM, but does not perform as well as MO. This shows that increased quality of precipitation representation may be achieved through increased resolution and also demonstrates the extra performance achieved in MO due to high resolution 4DVAR data assimilation.

Figure 9 displays frequency bias for the EURO4M datasets, MO-C and MO-D. This figure shows that MO-D behaves similarly to ERA-Interim. At lower thresholds it is over-representative of precipitation events and at higher thresholds it is under-representative of precipitation events. However in all cases it is considerably less biased than ERA-Interim. MO-C also behaves similarly to ERA-Interim at low thresholds and is of similar quality to ERA-Interim. However at higher thresholds, MO-C is much less biased and is most similar to MO. These results suggest that higher resolution leads to more realistic amounts of precipitation, especially at higher thresholds.

There is a comparison of the quality of representation of other variables in MO, MO-C and MO-D in Appendix F.

5.4 Stable Equitable Error in Probability Space

The Stable Equitable Error in Probability Space (SEEPS) was developed at ECMWF specifically to assess representation of precipitation in numerical models [Rodwell M.J. et al.(2010)]. Unlike other scores, the SEEPS event categories are defined by probability of occurrence based on station climatology, rather than absolute threshold. The threshold of each category, ‘dry’, ‘light’ and ‘heavy’, therefore, varies from station to station. It is detailed in Appendix B.

The 2008 SEEPS scores over the European domain for ERA-Interim and MO are shown in Figure 10. These are shown as one minus the SEEPS score, since SEEPS is an error score, so that the vertical axis represents increasing quality of the models. The figure shows that, unlike the ETS, SEEPS is not strongly dependent on season. The figure also shows that using this measure ERA-Interim and MO are of similar quality throughout.

Figure 11 displays the components of the SEEPS error score, averaged over 2008, for the two models. This Figure demonstrates that the error components of the two models are broadly similar. MO is less likely to wrongly classify dry conditions (as either light or heavy precipitation) and is less likely to wrongly classify heavy precipitation as light precipitation than ERA-Interim. ERA-Interim is less likely to wrongly classify either light or heavy precipitation as dry events.

Figure 7: *Frequency bias 2008 for ERA-Interim (darker blue), MO (lighter blue) and HIRLAM (red) reanalyses (24hr accumulations) at thresholds 1mm(L, top), 4mm(R, top), 8mm (L, middle), 16mm (R, middle) and 32 (bottom).*

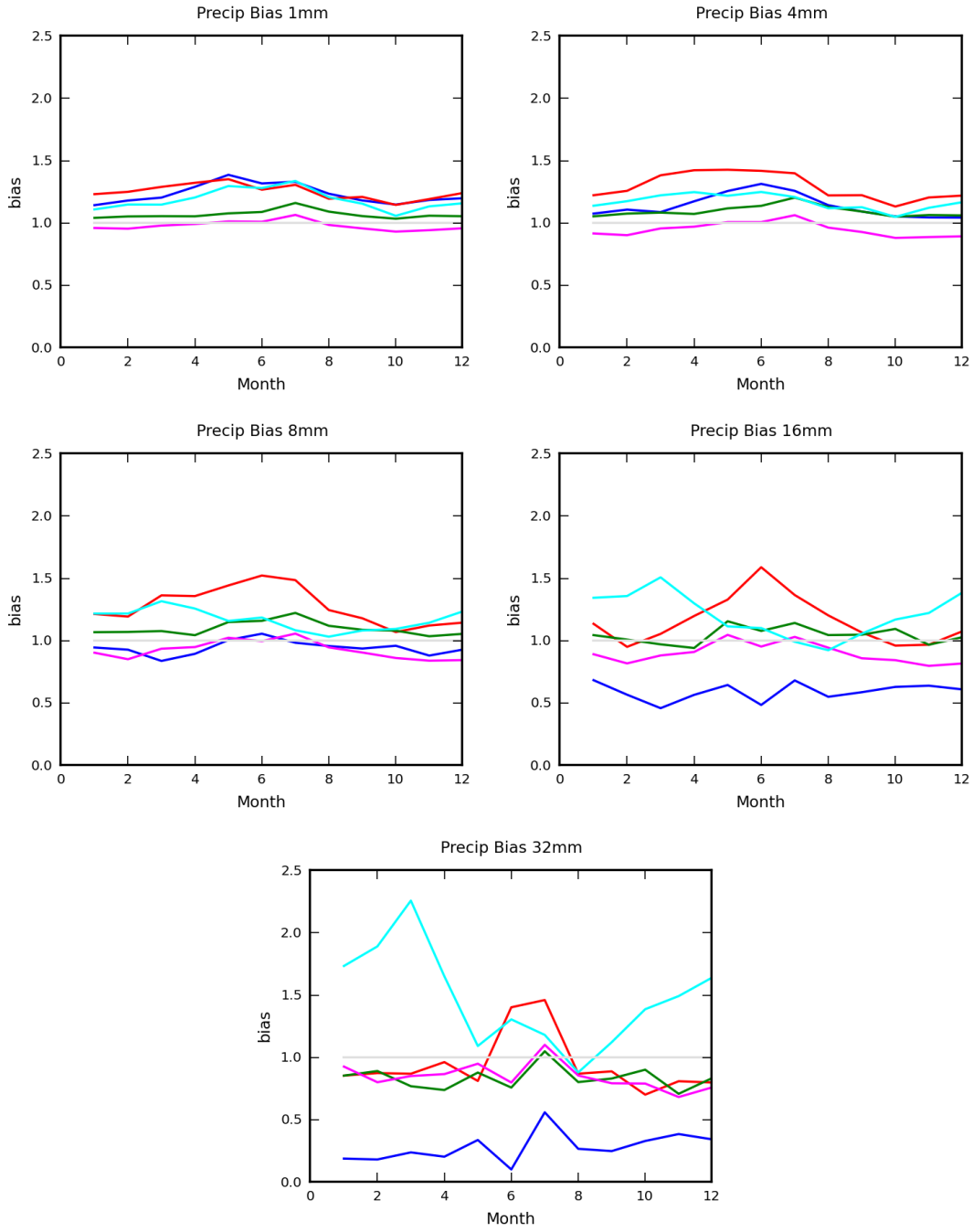


Figure 8: *Monthly Equitable Threat Scores 2008 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black), MO-D (grey) and HIRLAM (red) reanalyses (6hr accumulations) at thresholds 0.5mm(L, top), 1mm(R, top), 4mm (L, middle), 8mm (R, middle) and 16mm (bottom).*

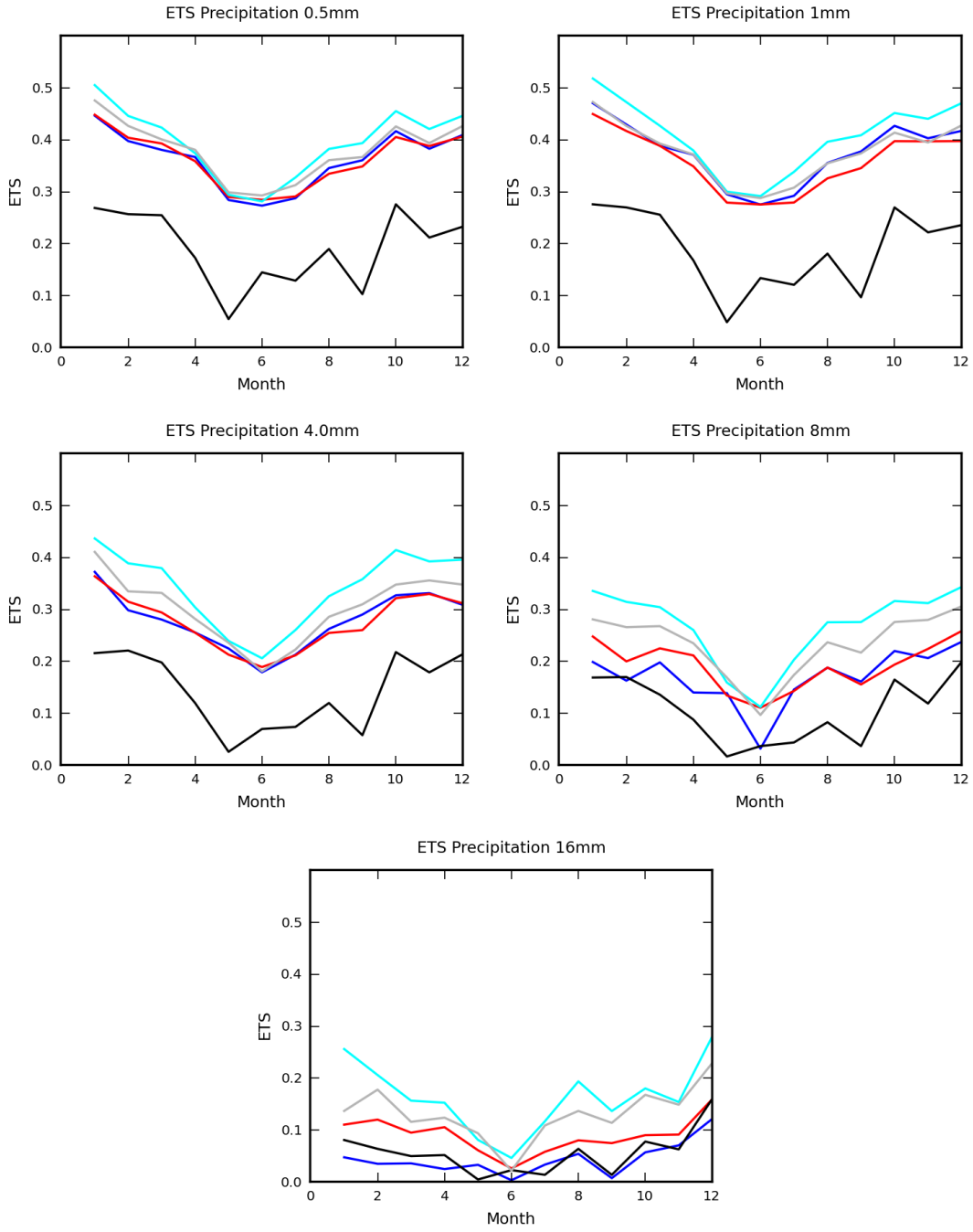


Figure 9: *Frequency Bias 2008 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black), MO-D (grey) and HIRLAM (red) reanalyses (6hr accumulations) at thresholds 0.5mm(L, top), 1mm(R, top), 4mm (L, middle), 8mm (R, middle) and 16mm (bottom).*

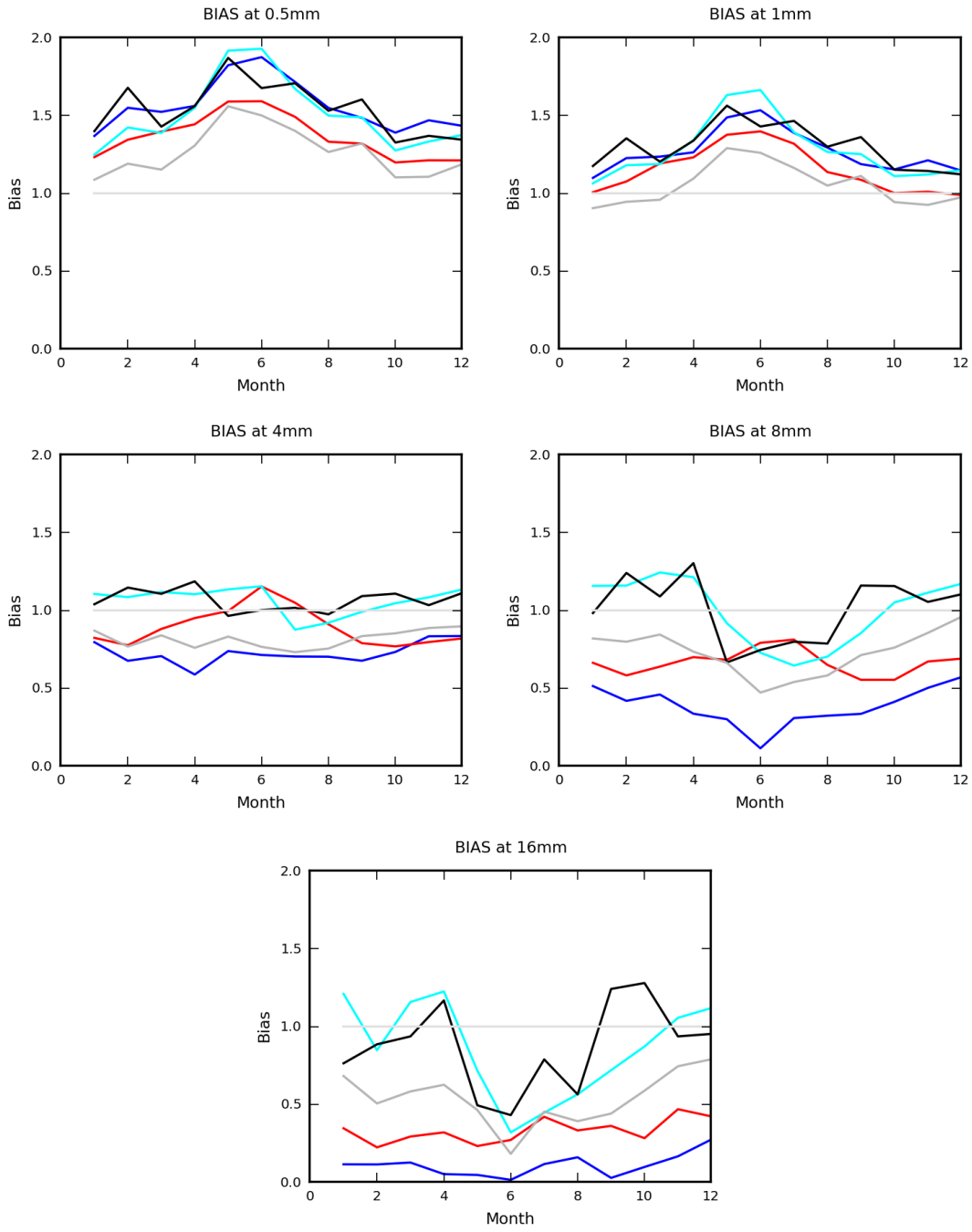


Figure 10: *1-SEEPS scores for ERA-Interim (dark blue) and MO (light blue). Solid lines - mean, dashed - maximum, dotted - minimum.*

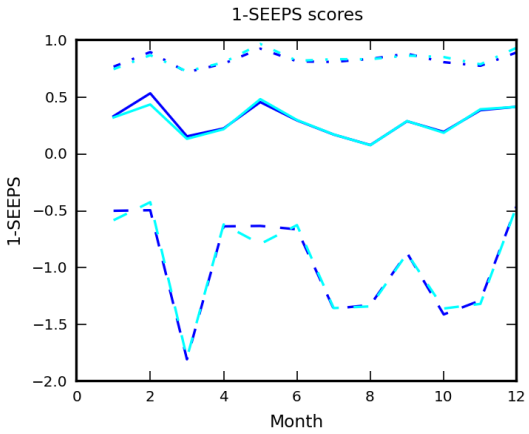


Figure 11: *SEEPS error for ERA-Interim and MO. Green - observed dry classified as light and heavy precipitation for light and dark, respectively. Orange - observed light precipitation classified as dry and heavy precipitation for light and dark, respectively. Red - observed heavy precipitation classified as dry and light precipitation for light and dark, respectively.*

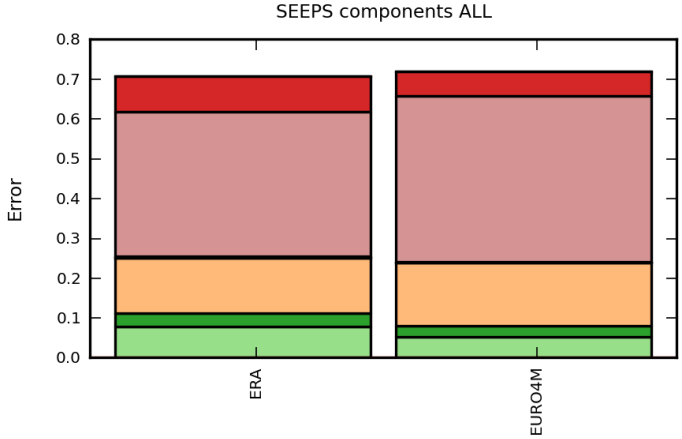
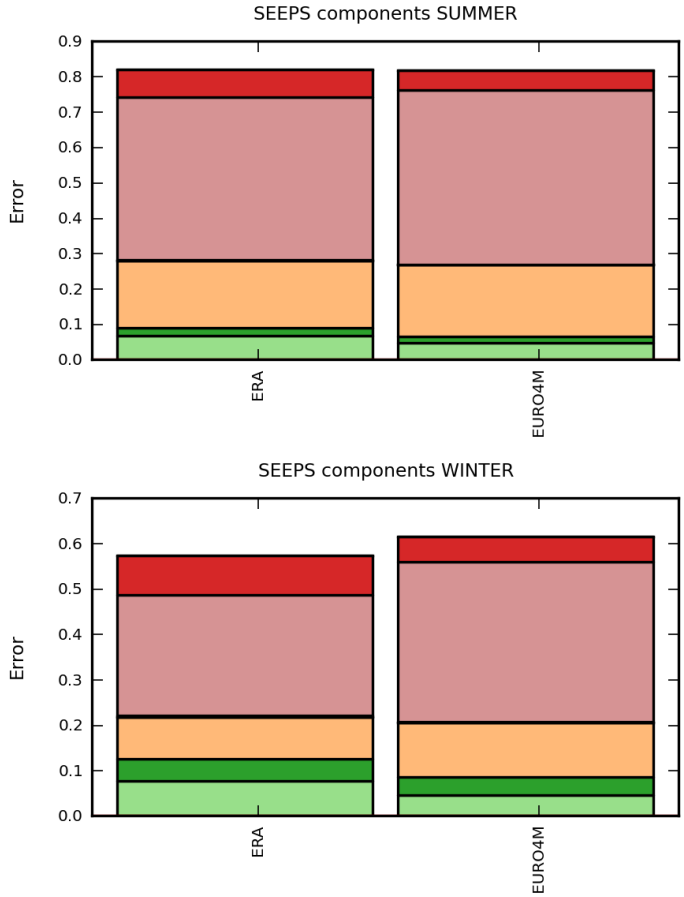


Figure 12: *SEEPS error for ERA-Interim and MO for summer months (June, July, August) (L) and winter months (December, January, February) (R). Green - observed dry classified as light and heavy precipitation for light and dark, respectively. Orange - observed light precipitation classified as dry and heavy precipitation for light and dark, respectively. Red - observed heavy precipitation classified as dry and light precipitation for light and dark, respectively.*



Similarly, Figure 12 displays the components of the SEEPS error score, averaged over summer and winter months. In summer months the models perform similarly well, but in winter months, ERA-Interim has the smaller error. In winter months events wrongly classified in the models as heavy precipitation are increased and events wrongly classified as dry are decreased compared to summer. This second category has a larger decrease in ERA-Interim than in MO and this is the major contributor to the former's relatively good performance in winter months. Events wrongly classified as light precipitation increase in winter in ERA-Interim, but not in MO.

As with annual RMSE, see Figure 1, the SEEPS scores suggest similar performance between ERA-Interim and MO. Although the higher resolution of MO out-performs ERA-Interim in representation of extreme events, it performs similarly to ERA-Interim for low threshold events. These events are prevalent in Europe and therefore dominate the SEEPS scores, which divides all events into three categories of approximately equal probability of occurrence.

6 Fractions Skill Scores

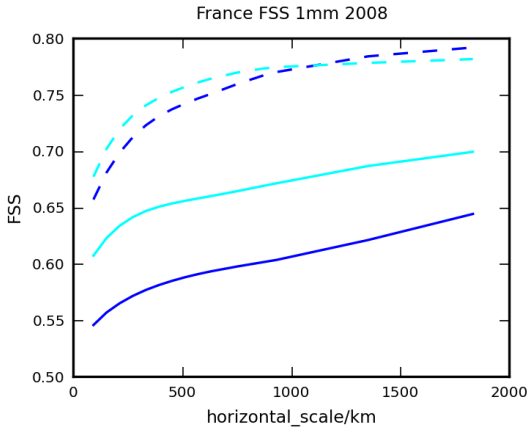
As noted previously, a reanalysis may accurately represent precipitation at a larger scale without accurately representing precipitation at specific locations (and potentially vice-versa). Although useful, station-based verification does not assess the reanalyses precipitation representation even at grid-box scale. The FSS, proposed by [Roberts N.M. and Lean H.W.(2008)] and detailed in Appendix C, measures the fraction of a subset of grid-boxes for which a certain threshold is passed. The fractions skill score is a measure of how well this fraction matches its observed equivalent, using gridded observation data. In this way the ability of the reanalysis/downscaler to accurately represent spatial density of precipitation events over a range of scales can be assessed.

A challenge for calculating FSS is obtaining high-resolution accurate gridded observations to use as truth. In other studies reliable precipitation radar data has been used as truth (e.g. [Roberts N.M. and Lean H.W.(2008)], [Stratman R. et al.(2013)], etc), but this is not readily available across Europe for the comparison year of 2008. Satellite data also has sufficient coverage, but does not accurately represent all forms of precipitation and may be less accurate than the reanalyses [Pfeifroth U. et al.(2013)], [Vila D. et al.(2009)]. The Global Precipitation Climatology Centre dataset (GPCC) dataset, [Becker A. et al.(2013)], has also been used as a truth for comparing reanalyses, e.g. [Betts A.K. et al.(2006)], but this is comparatively low resolution and is also not available for 2008. The ECMWF gridded observation dataset discussed in Section 4 is therefore used again here as the truth. To ensure accurate truth the scores are calculated for two subdomains covering France and Germany where the rain gauge network contributing to the dataset is most dense. The French and German domains are both approximately 600km by 700km which is of the order suggested by [Mittermaier M. and Roberts N.(2010)] as useful for mesoscale verification. Calculations were carried out using twenty-four hour accumulations (the minimum period in the observation dataset) and the 0.225 degree grid on which the dataset is based. Although scores must be at least 0.5 to be considered ‘useful’, scores less than this still have skill (i.e. the modelled precipitation is better than a random accumulation) [Mittermaier M. and Roberts N.(2010)] and so even low scores are considered for these comparisons.

As an example, Figure 13 shows FSS using a 1mm threshold for January and June 2008 across the French domain for ERA-Interim and MO. In January MO has a higher score than ERA-Interim at all scales. In June MO is more skillful than ERA-Interim at scales below approximately 1100km and less skillful above it. As shown in Figure 13, scores generally increase with scale and also decrease with threshold size (not shown), [Mittermaier M. and Roberts N.(2010)].

The comparison between two reanalyses covers horizontal scales from 90km (approximate size of three observation grid-boxes) to 930km (approximate size of 31 grid-boxes) and thresholds from 1mm to 16mm for each month in 2008, producing a large number of plots such as Figure 13. For brevity and clarity these results have been summarised by considering FSS relative differences to ERA-Interim for each of the EURO4M datasets, averaging over all months.

Figure 13: *Fractions Skill Scores over France for January 2008 (solid lines) and June 2008 (dashed lines). ERA-Interim - dark blue, MO - light blue.*



6.1 Relative Differences to ERA-Interim

Figure 14 displays relative difference in FSS with ERA-Interim for increasing threshold. This figure shows that the benefit of each of the regional models, over the global ERA-Interim, increases with increasing threshold. This may be expected since high threshold events are best represented by higher resolution datasets. The downscalers, MESAN and MESCAN, represent precipitation at all scales and thresholds better than the other datasets. The results shown here are not truly fair to the reanalyses, ERA-Interim, HIRLAM and MO, since the truth used here may be considered a subset of the observations assimilated into the downscalers, but not the reanalyses. However it is clear from these results that assimilation of precipitation leads to a large improvement in its representation.

Figure 14 also demonstrates that MO improves on HIRLAM and that this improvement also increases with threshold. This improvement is in part due to the increased resolution in MO, but is also due to using 4DVAR data assimilation instead of the 3DVAR assimilation used in HIRLAM.

Figure 15 displays relative difference in FSS with ERA-Interim for increasing scale. Again this demonstrates that the downscalers MESAN and MESCAN are the best datasets at representing precipitation at all scales and thresholds, but especially at the shortest scales. In Germany at 16mm threshold, the lowest scales feature unusually low relative skill for all the datasets. This feature is probably due to relatively few high threshold events occurring at these scales.

Figure 15 also shows that MO out-performs ERA-Interim and HIRLAM for all scales and thresholds. Its relative skill decreases slightly with increasing length scale. HIRLAM only improves on ERA-Interim for 4mm and 16mm thresholds (medium and heavy rain, respectively). For 1mm threshold (light rain) it is slightly worse at representing precipitation than ERA-Interim for most scales. Unlike the other EURO4M datasets, its skill relative to ERA-Interim increases slightly with horizontal length scale.

Figure 14: *Mean relative difference to ERA-Interim in FSS over France (L) and Germany (R) (2008) for increasing threshold. HIRLAM - red, MESAN - green, MESCAN - pink, MO - light blue. For horizontal scales 90km (top), 330km (middle) and 930km (bottom).*

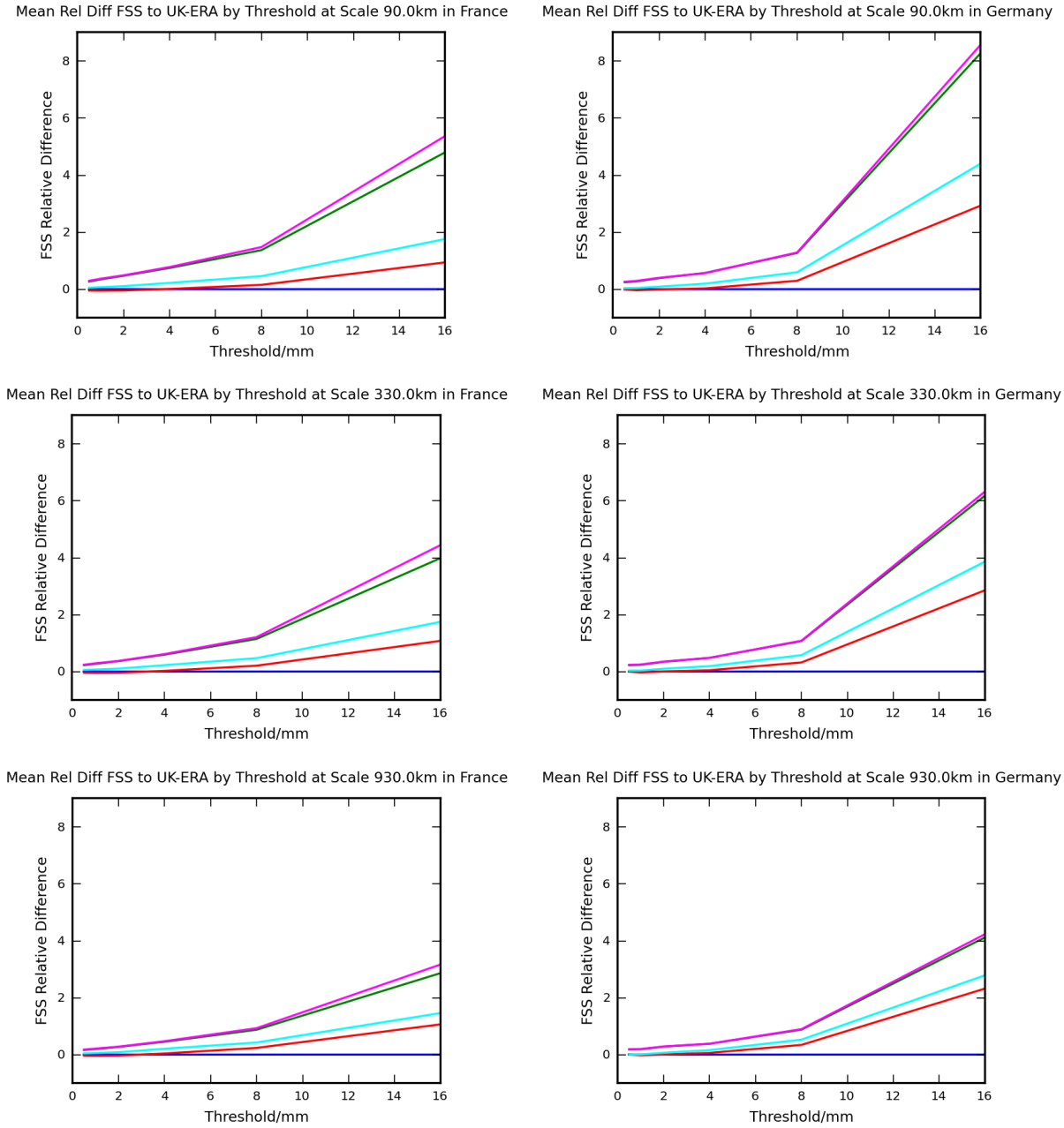


Figure 15: *Mean relative difference to ERA-Interim in FSS over France (L) and Germany (R) (2008) for increasing horizontal scale. HIRLAM - red, MESAN - green, MESCAN - pink, MO - light blue. For thresholds 1mm (top), 4mm (middle) and 16mm (bottom).*

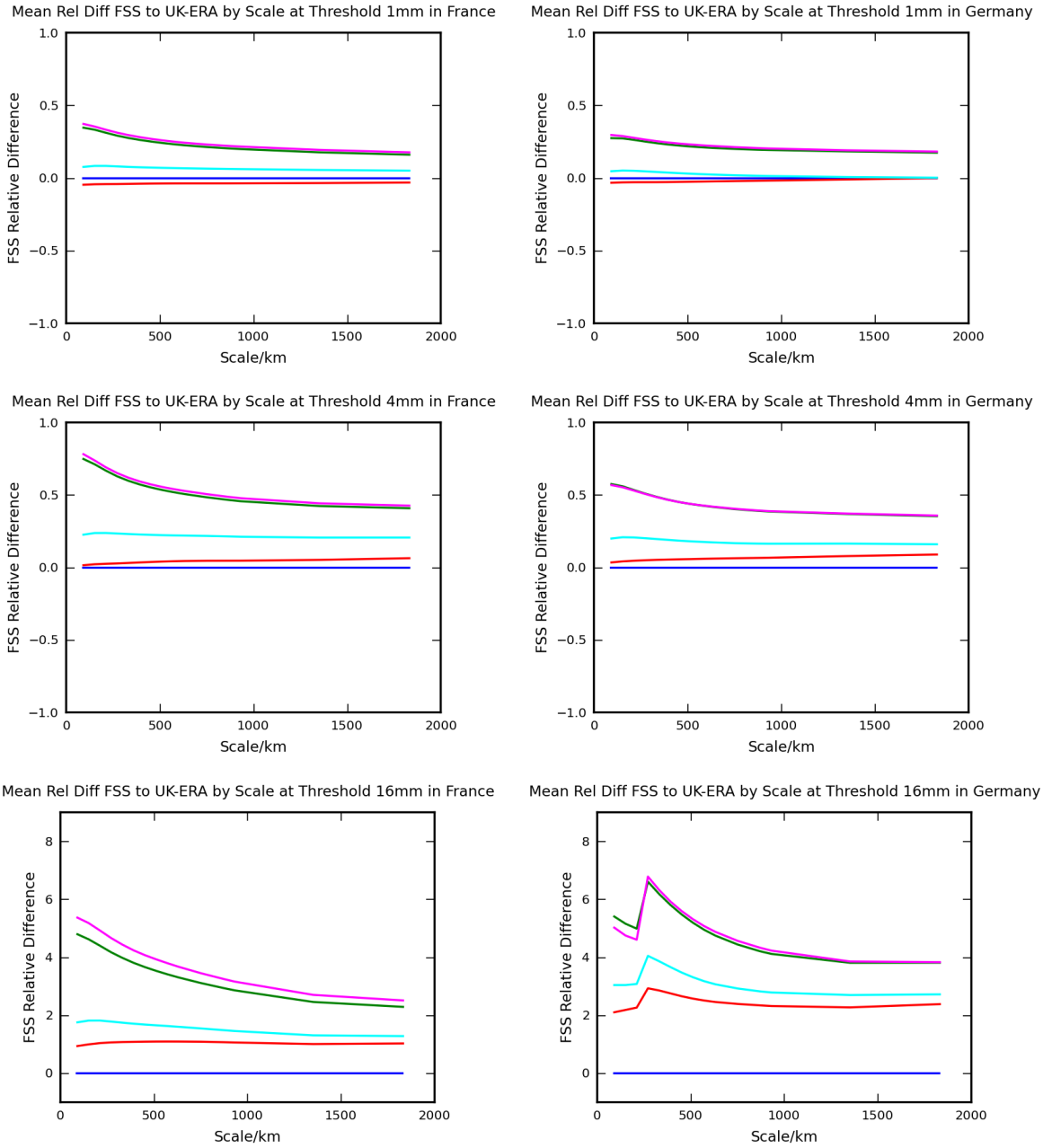


Table 2: Useful Scales

Dataset	France-1mm	Germany-1mm	France-4mm	Germany-4mm
ERA-Interim	≥570km	≥330km	none	none
HIRLAM	≥1350km	≥450km	none	≥1830km
MESAN	all	all	all	all
MESCAN	all	all	all	all
MO	≥210km	≥330km	≥750km	≥510km

Table 3: Useful Scales

Dataset	France-1mm	Germany-1mm	France-4mm	Germany-4mm
MO	≥210km	≥330km	≥750km	≥510km
HIRLAM	≥1350km	≥450km	none	≥1830km
MO-3DVAR	none	none	none	none

6.2 Usefulness

A model is considered useful at a certain scale if its FSS is at least $(1+f_0)/2$, where f_0 is the FSS of random data with the same proportion of events as the observations, [Roberts N.M. and Lean H.W.(2008)]. Table 2 shows the scales above which the models are useful for the French and German domains.

Table 2 shows that the downscalers MESAN and MESCAN are useful at all scales and therefore demonstrates the high quality achievable through rain gauge assimilation. At 1mm MO shows a benefit over ERA-Interim, but HIRLAM does not, demonstrating that ERA-Interim is of comparable quality to regional models for low-threshold events associated with broad scales. At 4mm both HIRLAM and MO have at least some scales for which they are useful, but ERA-Interim is not useful for any scale. Therefore useful representation of precipitation for any medium or large threshold events requires a regional model. None of the models have useful FSS at 16mm, indicating that further advances in regional reanalysis are required for useful data on such large threshold events.

6.3 3DVAR vs 4DVAR

To assess the impact of using 4DVAR assimilation over 3DVAR on precipitation representation, FSS are calculated for a variant of MO, called MO-3DVAR, which is driven by 3DVAR instead of the usual 4DVAR. MO-3DVAR is similar to MO except that in the assimilation, persistence is used to propagate the increment to the observation times, instead of the usual perturbation forecast model. In this section the FSS are shown as mean relative difference to MO-3DVAR and score differences for the months May, June and September 2008 are meaned to create the results.

Figure 16 shows the relative difference in FSS for MO compared to MO-3DVAR with increasing threshold. This demonstrates that the benefit of using 4DVAR increases with increasing threshold. This trend is expected since high threshold events are generally best represented by high resolution datasets. Even at 1mm threshold MO shows an improvement over MO-3DVAR. Figure 17 shows the relative difference in FSS for MO compared to MO-3DVAR with increasing horizontal scale, demonstrating that the benefit of 4DVAR increases with decreasing length scale. Even at the largest scales MO shows improvement over MO-3DVAR.

Table 3 shows the useful scales for MO, MO-3DVAR and HIRLAM. This demonstrates that the high resolution

Figure 16: *MO mean relative difference to MO-3DVAR in FSS over France (L) and Germany (R) against increasing threshold for 90km (top), 330km (middle) and 930km (bottom).*

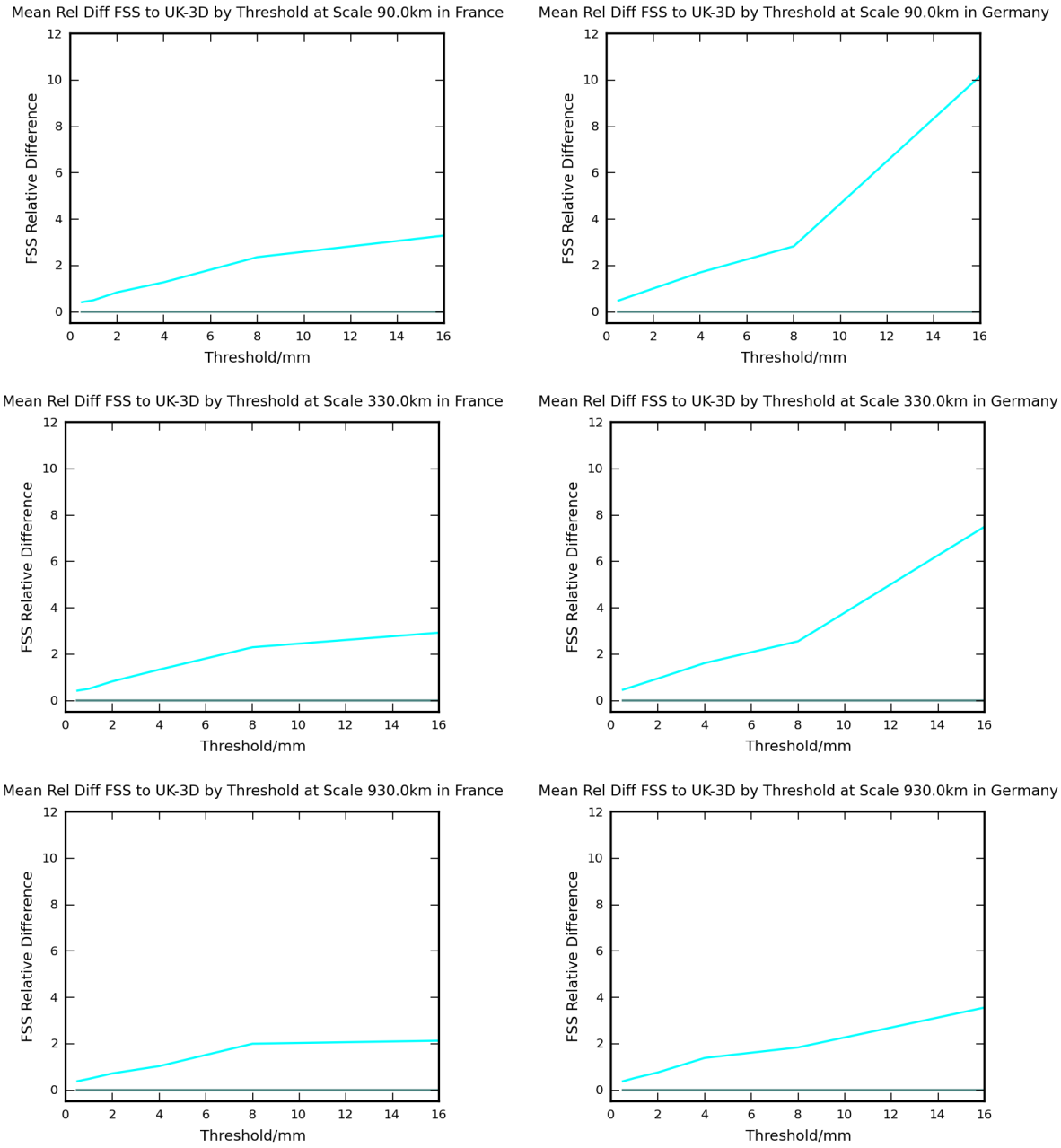
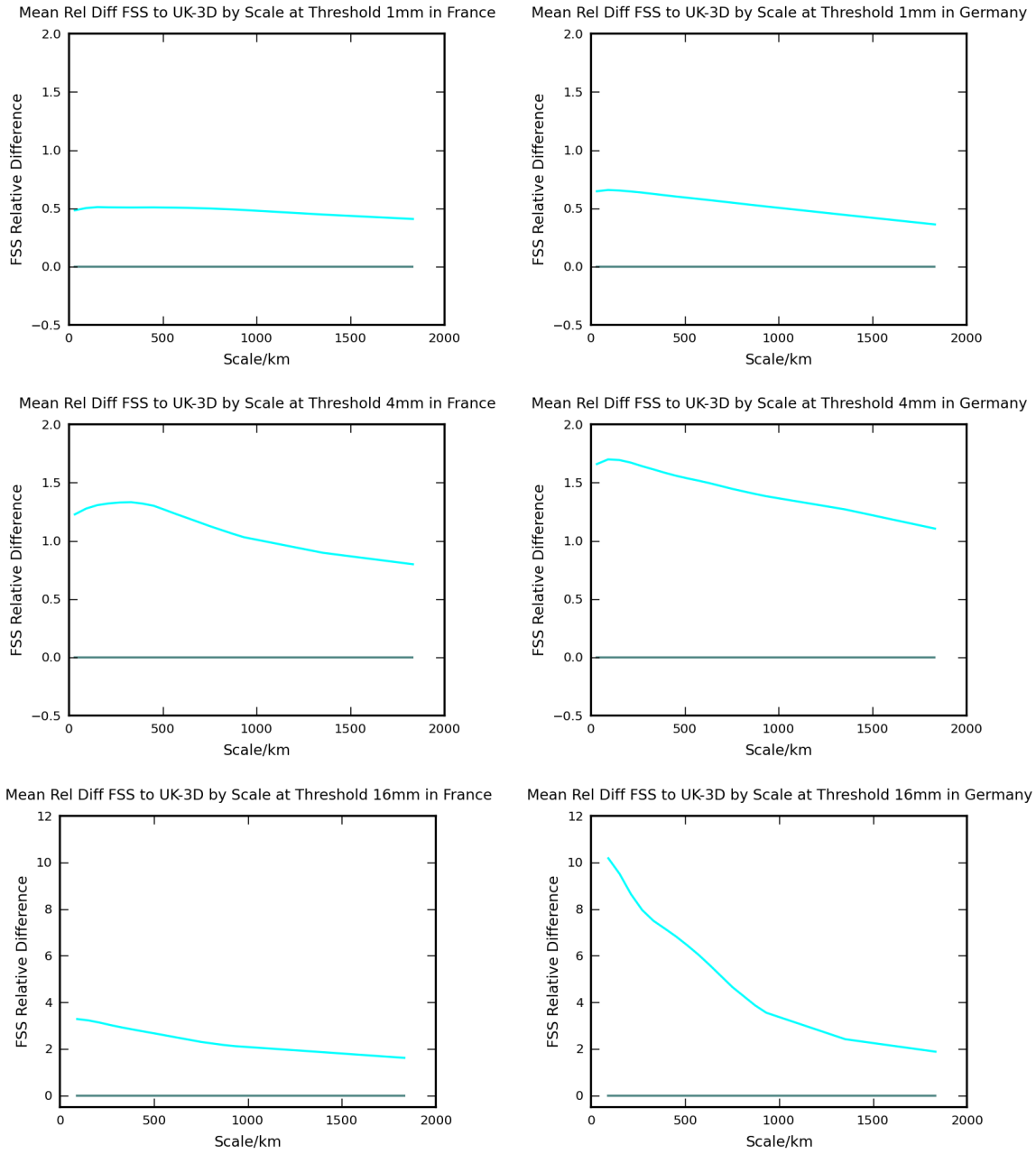


Figure 17: *MO mean relative difference to MO-3DVAR in FSS over France (L) and Germany (R) against increasing scale for 1mm (top), 4mm (middle) and 16mm (bottom).*



Met Office reanalysis is not useful at any scale without 4DVAR. MO-3DVAR is even out-performed by the lower resolution HIRLAM.

The FSS results demonstrate that the global reanalysis, ERA-Interim, is sufficient for representation of low threshold events associated with the broadest scales, but higher resolution, as in the regional reanalyses HIRLAM and MO, greatly improves representation of precipitation at higher thresholds/smaller scales. Using 4DVAR assimilation, over 3DVAR, is necessary to accurately represent these small scale events. Assimilation of precipitation observations greatly improves its representation at all scales and thresholds.

7 Representation of Climate Monitoring Statistics

The World Meteorological Organisation's Expert Team on Climate Change Detection and Indices (ETCCDI) have defined 27 core indices to describe and monitor climate change [Zhang X.(2013)]. Precipitation statistics include maximum daily precipitation accumulation, maximum accumulation across five consecutive days, the number of days with at least 1mm (wet days), 10mm and 20mm of precipitation, the maximum length of a dry spell, the maximum length of a wet spell and the total precipitation on wet days. These statistics have been calculated for each month 2008-2009 for each of ERA-Interim, HIRLAM, MESAN, MESCAN and MO, together with the monthly mean precipitation.

The European Climate Assessment and Dataset (ECA&D) includes these statistics for a range of observation stations across Europe [van Engelen A. et al.(2013)]. To compare representation of climate statistics, correlations between the ECA&D observation-based statistics and the model equivalents are calculated. These model equivalents are obtained by first remapping each of the models onto a common 0.1 degree global grid, using CDO [Akhtar M. et al.(2011)], to reduce the comparison dependence on grid difference [Lanciani A. et al.(2008)]. The common grid is a compromise between lower resolution ERA-Interim and the very high resolution downscalers, and is at the same resolution as the native MO grid. Since the native MO grid has a rotated pole, all datasets require interpolation to the common grid, i.e. MO is not given an unfair advantage. Simple bi-linear interpolation is then used to obtain each model value at station positions. Correlation coefficients are calculated for each month 2008-2009 and time-series are compared for each of the systems. The correlation coefficients are a measure of how well the models represent monthly variation in the climate monitoring statistics.

7.1 Monthly Mean Precipitation

Figure 18 displays mean daily precipitation correlation coefficients for the models comparing with ECA&D station data. The twenty-four coefficients, one for each month, are sorted into ascending order for each model (i.e. comparing the best performing month for each model). In this way, the left-hand-side plot of Figure 18 ranks the coefficients for each model, demonstrating that correlation coefficients of the downscalers MESAN and MESCAN are much higher than the reanalyses. The ECA&D observation data should be considered a subset of the

Figure 18: *Coefficients of correlation between reanalyses/downscalers and monthly mean precipitation at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*

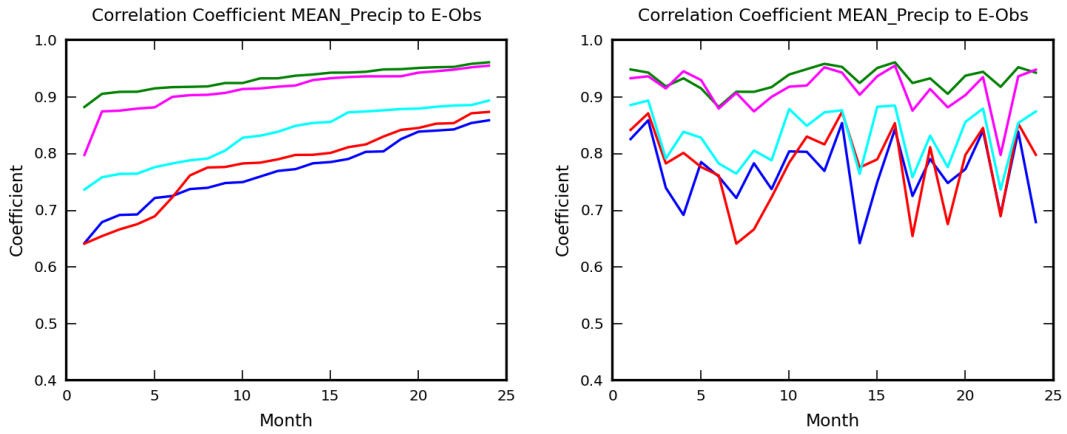
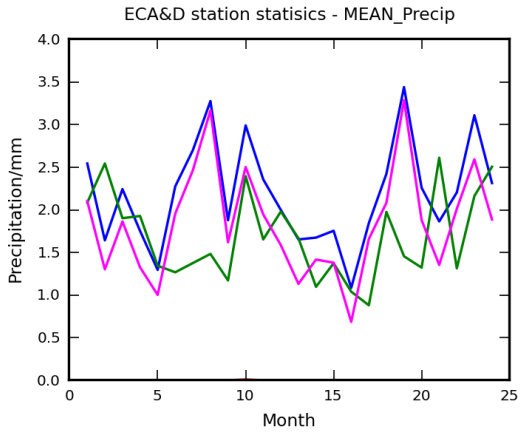


Figure 19: *Mean, median and maximum mean daily precipitation at ECA&D observation stations for 2008-2009. Dark blue - mean, green - maximum/10, pink - median*

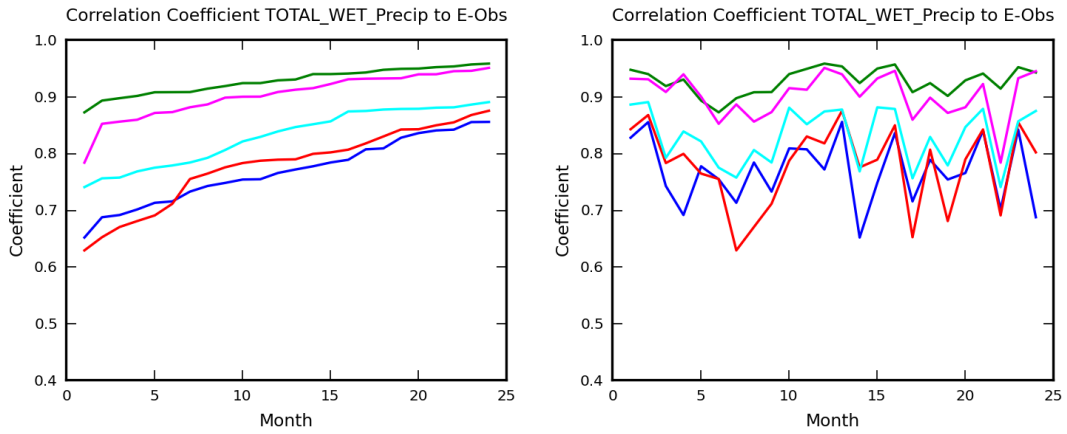


observations assimilated into the downscalers, but, even with this dependence, the downscalers demonstrate a step-change improvement in precipitation representation via assimilation of precipitation observations. HIRLAM is generally an improvement on ERA-Interim, demonstrating the value of increased resolution limited area re-analyses. MO has correlation coefficients which are an improvement on ERA-Interim and HIRLAM, but not as good as MESAN/MESCAN. This suggests that, although considerable improvement in precipitation representation is possible through increased resolution alone, very accurate precipitation representation is only possible through assimilation of precipitation observations.

The right-hand-side plot of Figure 18 orders the coefficients by calendar month. This shows that MESAN is better correlated than MESCAN to observation station mean precipitation for all bar three months - April 2008, May 2008 and December 2009. HIRLAM out-performs ERA-Interim on 16/24 months and out-performs MO for just one month in February 2009. MO out-performs ERA-Interim for the entire period and HIRLAM for 23/24 months, but performs considerably less well than MESAN or MESCAN throughout.

In October 2009 all models perform relatively badly, with MESAN the least affected. This effect is seen to a

Figure 20: *Coefficients of correlation between reanalyses/downscalers and total precipitation on wet days at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*



lesser extent in May 2009 and seems to be associated with months with low maxima that do not have dry mean stations. See Figure 19, which summarises the observation station values. In July 2008 HIRLAM performs much worse than the other models which is quite a wet month with low maxima. However attribution of relative performance to different weather types is not clear in these statistics.

MESAN, MESCAN and MO consistently improve on ERA-Interim's representation of monthly mean precipitation across Europe throughout 2008-2009, with MESAN or MESCAN the most accurate, dependent on weather. Also dependent on weather, HIRLAM out-performs ERA-Interim for just over half the period, but the benefits of its increased resolution are not strongly expressed through this statistic.

7.2 Total Wet Precipitation

Figure 20 displays correlation coefficients for the models comparing with the total precipitation on wet days (days with at least 1mm precipitation) in the ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that correlation coefficients of the downscalers MESAN and MESCAN are again much higher than the reanalyses with MESAN better than MESCAN. MO performs better than ERA-Interim and HIRLAM, which are of similar quality.

The right-hand-side plot of Figure 20 orders the coefficients by calendar month. For 21/24 months MESAN has highest correlation with the observations and MESCAN has highest correlations in the remaining 3 months. MO out-performs ERA-Interim on all months and HIRLAM on all but February 2009, which is a dry month, see Appendix D. HIRLAM out-performs ERA-Interim on 16/24 months suggesting their performance is similar.

7.3 Maximum Daily Precipitation

Figure 22 displays maximum daily precipitation correlation coefficients for the models comparing with ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that correlation coefficients

Figure 21: *Mean, median and maximum total precipitation on wet days at ECA&D observation stations for 2008-2009. Dark blue - mean, green - maximum/10, pink - median*

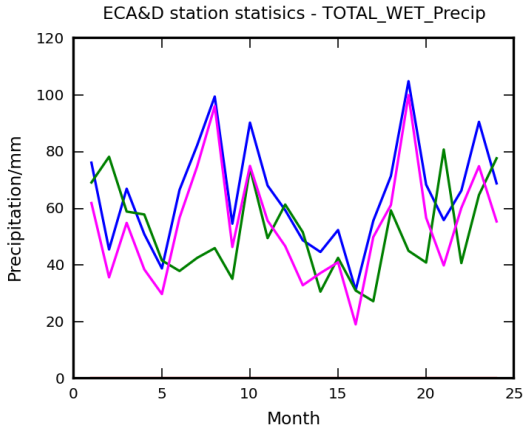


Figure 22: *Coefficients of correlation between reanalyses/downscalers and maximum daily precipitation at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*

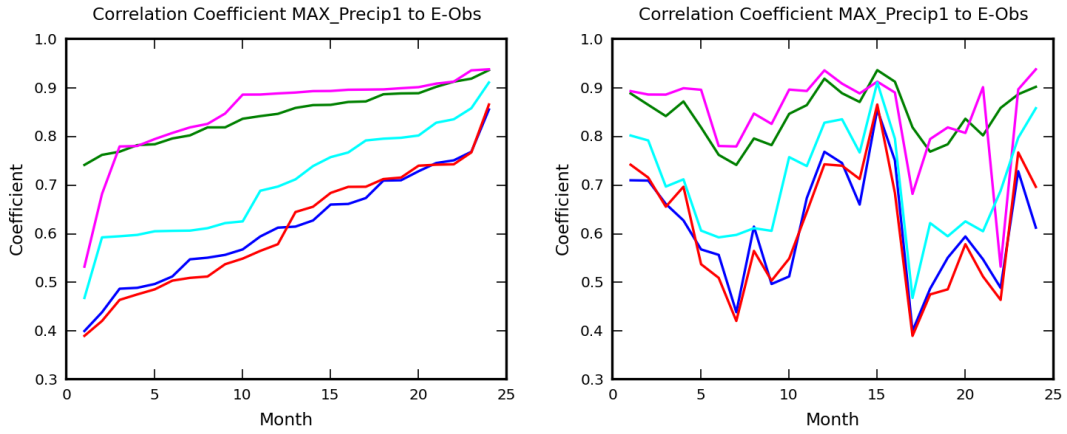
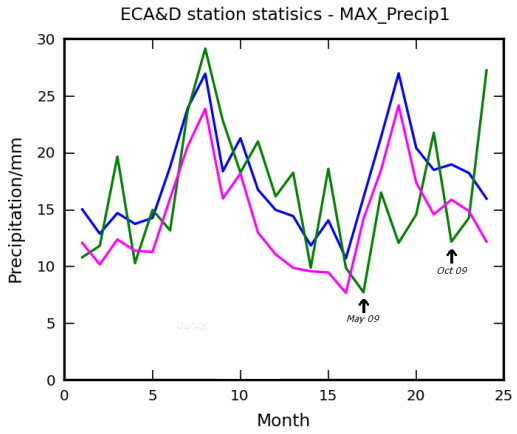


Figure 23: *Mean, median and maximum daily maximum precipitation at ECA&D observation stations for 2008-2009. Dark blue - mean, green - maximum/10, pink - median*



of the downscalers MESAN and MESCAN are again much higher than the other models. MESCAN has slightly higher correlations than MESAN except for its lowest correlations, which are much lower than those of MESCAN. HIRLAM is of similar quality to ERA-Interim. As with mean precipitation, MO has correlation coefficients which are better than ERA-Interim and HIRLAM, but not as high as MESAN/MESCAN.

The right-hand-side plot of Figure 22 orders the coefficients by calendar month. This shows that MESCAN is better correlated than MESAN to observation station maximum precipitation for all but five months. Although MESAN is generally best at representing mean precipitation, MESCAN is generally best at maxima, since more extreme values are associated with shorter length scales. In October 2009, however, MESCAN performs worse even than MO. HIRLAM out-performs ERA-Interim on 9/24 months. MO out-performs HIRLAM for the entire period and ERA-Interim for 23/24 months, but again performs considerably less well than MESAN or MESCAN throughout.

In October 2009, ERA-Interim, HIRLAM, and especially MESCAN, perform relatively badly, with MESAN and MO unaffected. Again, this effect is seen to a lesser extent in May 2009 and seems to be associated with months with low maxima. See Figure 23 which summarises the observation station values. This is consistent with MESCAN being especially suited to representing extremes. MESAN, MESCAN and MO consistently improve on ERA-Interim representation of maximum daily precipitation across Europe throughout 2008-2009, with MESAN or MESCAN the most accurate. HIRLAM has similar performance to ERA-Interim so again the benefits of its increased resolution are not demonstrated by this statistic.

7.4 Maximum 5-Day Precipitation

Figure 24 displays maximum five-day precipitation correlation coefficients for the models comparing with ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that correlation coefficients of the downscalers MESAN and MESCAN are again much higher than the reanalyses. MESAN has slightly higher correlations than MESCAN especially at its lowest correlations, which are much lower than those

Figure 24: *Coefficients of correlation between reanalyses/downscalers and maximum five-day precipitation at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*

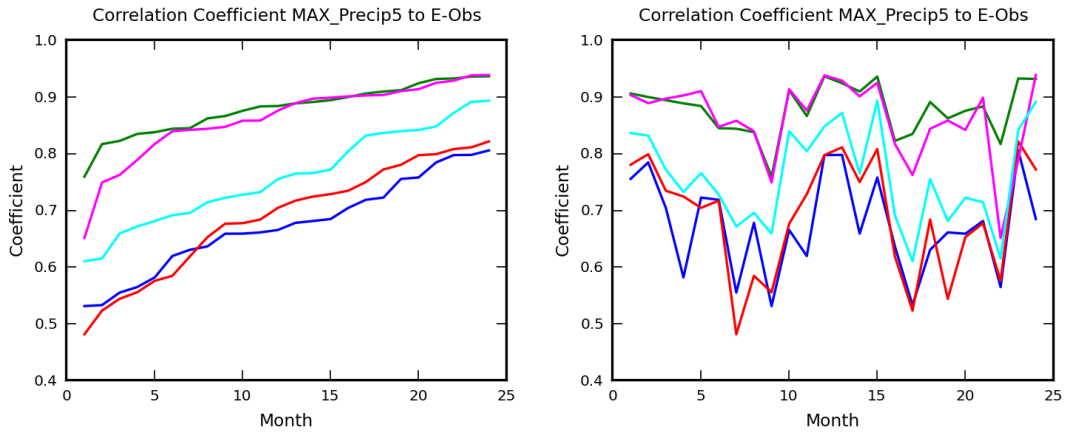
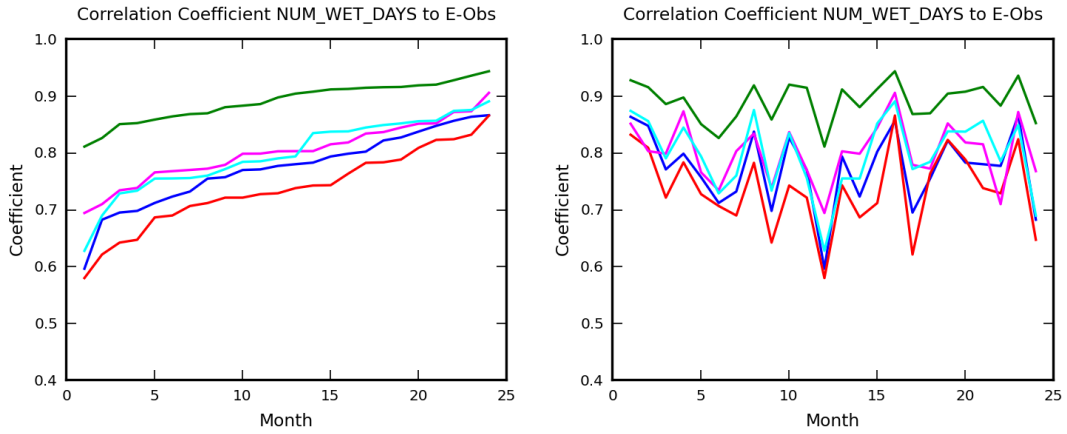


Figure 25: *Coefficients of correlation between reanalyses/downscalers and number of wet days at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*



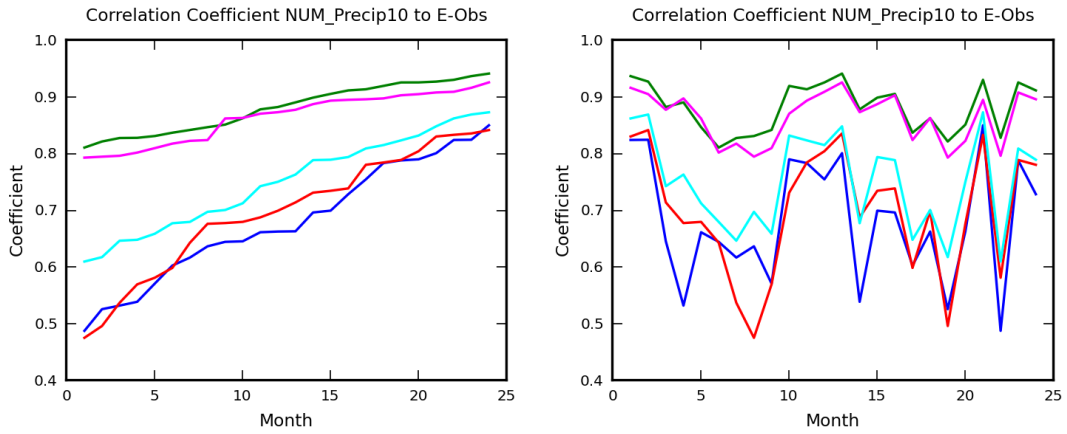
of MESAN. HIRLAM is of similar quality to ERA-Interim. As with other statistics, MO has correlation coefficients which are better than ERA-Interim and HIRLAM, but not as good as MESAN/MESCAN.

The right-hand-side plot of Figure 24 orders the coefficients by calendar month. This shows that MESCAN is better correlated than MESAN to observation station maximum five-day precipitation for 12/24 months, which suggests they are of similar quality. MO performs better than ERA-Interim and HIRLAM for the full period. HIRLAM performs better than ERA-Interim on 14/24 months, again suggesting that they are of similar quality in representing this extreme statistic.

7.5 Number of Wet Days

Figure 25 displays correlation coefficients for the models comparing with the number of wet days (days with at least 1mm precipitation) in the ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that MESAN has much greater correlation with the observation data than MESCAN and the

Figure 26: *Coefficients of correlation between reanalyses/downscalers and number of days with at least 10mm at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*



reanalyses. A threshold of 1mm over twenty-four hours is less than average for Europe and therefore is associated with longer length scales. MESCAN is concentrated on very short length scales and scores no better than MO. Likewise, HIRLAM is worse than ERA-Interim. The improved resolution of HIRLAM does not particularly help in resolving this broad scale feature which HIRLAM over-represents, see Section 5.1. MESCAN and MO have similar correlation coefficients which are higher than ERA-Interim and HIRLAM, but not as high as MESAN.

The right-hand-side plot of Figure 25 orders the coefficients by calendar month. For all months MESAN has the highest correlation with the observations. HIRLAM out-performs ERA-Interim on only 4/24 months. MO out-performs ERA-Interim on 21/24 months and out-performs HIRLAM for the entire period.

ERA-Interim performs relatively well for this longer scale statistic such that the benefits of the higher resolution HIRLAM reanalysis are not apparent. Likewise MESCAN, which concentrates on short scales, performs relatively badly, with MESAN consistently obtaining higher correlations.

7.6 Number of Days with at least 10mm

Figure 26 displays correlation coefficients between the models and the number of days with at least 10mm precipitation in the ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that correlation coefficients of the downscalers MESAN and MESCAN are once again much higher than the reanalyses, with MESAN generally better than MESCAN. HIRLAM has a small improvement on ERA-Interim. As with the other statistics, MO has correlation coefficients which are better than ERA-Interim and HIRLAM, but not as high as MESAN/MESCAN.

The right-hand-side plot of Figure 26 orders the coefficients by calendar month. For most months MESAN has greater correlation with the observations than MESCAN. In April 2008, May 2008 and June 2009 MESCAN out-performs MESAN. As shown in Figure 27, these are both months with few occurrences of 10mm precipitation days. In 16/24 months HIRLAM out-performs ERA-Interim and even out-performs MO in February 2009, again

Figure 27: *Mean, median and maximum number of days with at least 10mm at ECA&D observation stations for 2008-2009. Dark blue - mean, green - maximum/10, pink - median*

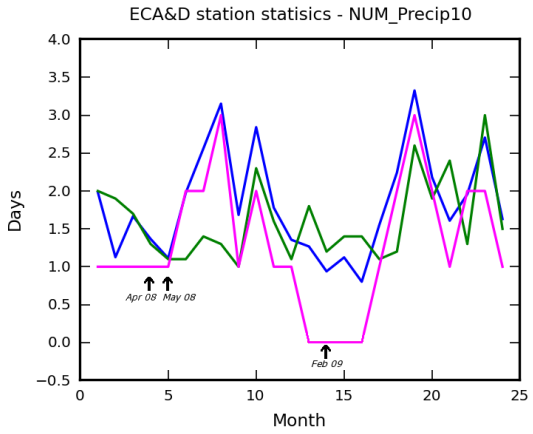
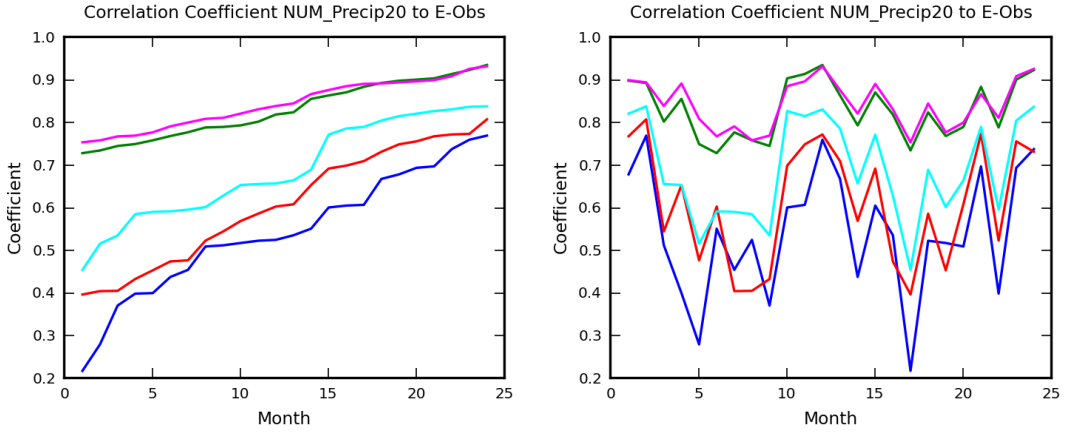


Figure 28: *Coefficients of correlation between reanalyses/downscalers and number of days with at least 20mm at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*



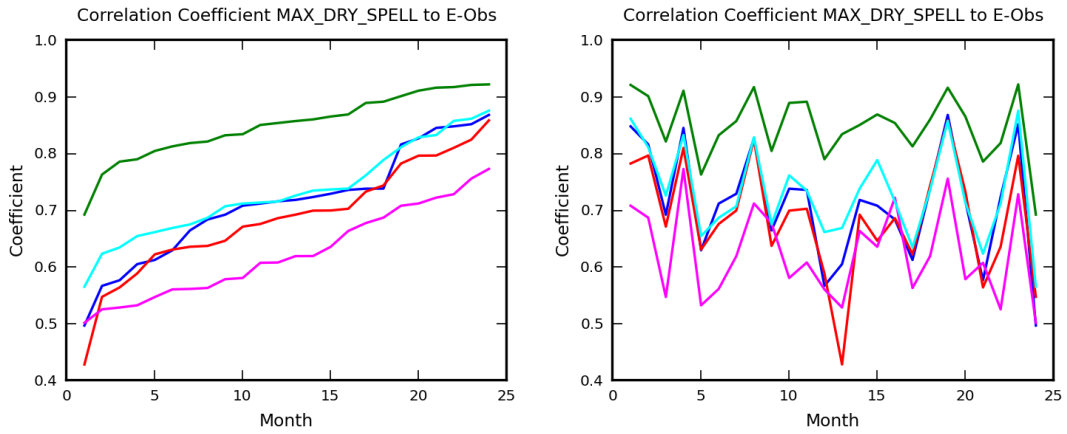
a month with few occurrences of 10mm precipitation days.

MESAN, MESCAN and MO consistently improve on ERA-Interim representation of monthly mean precipitation across Europe throughout 2008-2009, with MESAN generally having the highest accuracy. HIRLAM shows some improvement on ERA-Interim for this statistic.

7.7 Number of Days with at least 20mm

Figure 28 displays correlation coefficients for the models comparing with the number of days with at least 20mm precipitation in the ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that correlation coefficients of the downscalers MESAN and MESCAN are yet again much higher than the re-analyses, with MESCAN generally slightly better than MESAN. HIRLAM is an improvement on ERA-Interim. As with the other statistics, MO has correlation coefficients which are better than ERA-Interim and HIRLAM, but not as good as MESAN/MESCAN. The improvement of HIRLAM compared to ERA-Interim seems to show the benefit of increased resolution in representing extremes. Likewise the improvement of MESCAN, which concen-

Figure 29: *Coefficients of correlation between reanalyses/downscalers and maximum length of dry spell at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*



trates on shorter scales, over MESAN suggests that extreme values are associated with smaller scales.

The right-hand-side plot of Figure 28 orders the coefficients by calendar month. For most months MESCAN has greater correlation with the observations than MESAN. MESAN has higher correlations in just four months. In 19/24 months HIRLAM out-performs ERA-Interim and it even out-performs MO for one month, in June 2008.

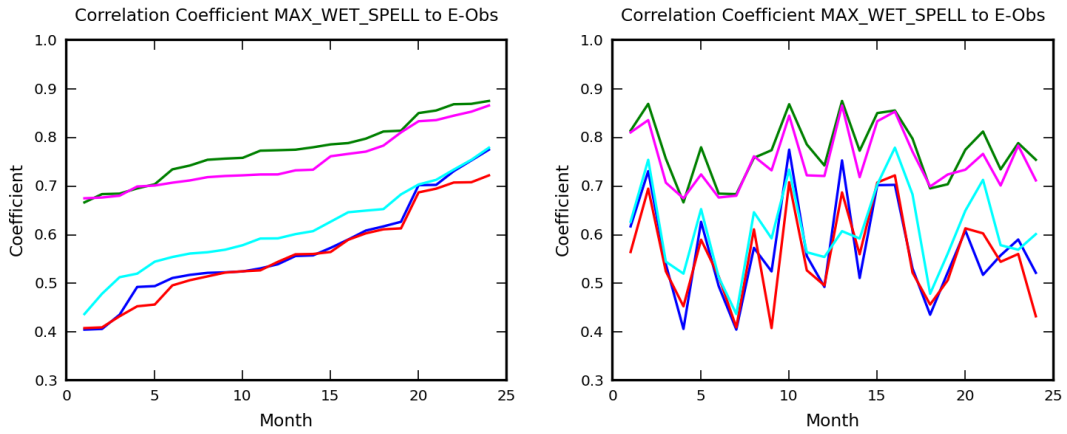
All EURO4M datasets consistently improve on ERA-Interim representation of 20mm days across Europe throughout 2008-2009 which demonstrates the particular improvement in representing extremes that the limited area models bring.

7.8 Maximum Dry Spell

The maximum dry spell (or consecutive dry days (CDD)) is the longest period of consecutive days for which precipitation was less than 1mm (dry day). It is useful for monitoring climate change since it is an important indicator of drought. Figure 29 displays correlation coefficients for the models comparing with the maximum dry spell in the ECA&D station data. The left-hand-side plot again ranks the coefficients for each model and shows that correlation coefficients of the downscaler MESAN is yet again much higher than the reanalyses, but MESCAN is generally worse than ERA-Interim, HIRLAM and MO. HIRLAM is slightly worse than ERA-Interim. As with the other statistics, MO has correlation coefficients which are generally better than ERA-Interim and HIRLAM, but not as high as MESAN.

The right-hand-side plot of Figure 29 orders the coefficients by calendar month. MESAN is the best system throughout the period. MO has similar results to ERA-Interim and both have higher correlations than HIRLAM on 21/24 and 17/24 months, respectively. HIRLAM performs better than MESCAN on 20/24 months. The regional models which better represent extremes of precipitation are less able to represent low or null precipitation events. Even so MESAN and MO perform better than ERA-Interim.

Figure 30: *Coefficients of correlation between reanalyses/downscalers and maximum length of wet spell at ECA&D observation stations for 2008-2009. Dark blue - ERA-Interim, red - HIRLAM, green - MESAN, pink - MESCAN, light blue - MO. Ordered by rank (L) and ordered by calendar (R).*



7.9 Maximum Wet Spell

Figure 30 displays correlation coefficients for the models comparing with the maximum wet spell in the ECA&D station data. The left-hand-side plot ranks the coefficients for each model and shows that correlation coefficients of MESAN and MESCAN are by far the highest, with MESAN slightly higher than MESCAN. MO performs better than ERA-Interim and HIRLAM performs slightly worse than ERA-Interim.

The right-hand-side plot of Figure 30 orders the coefficients by calendar month. MESAN is the best system for the full period, out-performing MESCAN on 20/24 months. MO out-performs ERA-Interim and HIRLAM on 21/24 months. HIRLAM has higher correlations on 13/24 months - i.e. it has similar ability to represent wet spells as ERA-Interim. Improved resolution is not in itself enough to improve representation of maximum wet spells.

Most of the precipitation-based climate statistics are represented much better by the high-resolution downscalers MESAN and MESCAN, which include precipitation observation assimilation, than by the reanalyses, which don't. MESAN is the best model for these statistics. For these statistics HIRLAM is of similar quality to ERA-Interim. MO scores overall better than HIRLAM or ERA-Interim, but not as well as MESAN or MESCAN.

For the two most extreme daily statistics, maximum of daily precipitation and number of days with at least 20mm precipitation, the models behave similarly except that MESCAN, which concentrates on the smallest scales associated with extremes, performs better than MESAN. For the number of wet days, in which low threshold precipitation is important, MESAN performs best, but MESCAN is no better than MO and HIRLAM is worse than ERA-Interim. For the maximum dry spell, MESCAN performs less well than the other four datasets.

8 Conclusions

Precipitation representation in the EURO4M reanalyses, HIRLAM and MO, and downscalers, MESAN and MESCAN has been assessed using a number of metrics, comparing against ERA-Interim which, prior to this project, was the 'gold standard' reanalysis for the European domain. Annual RMSE across the entire region indicates broad scale agreement between the EURO4M datasets and ERA-Interim. Although this document's primary aim is to assess precipitation in the EURO4M datasets, it is worth noting the high quality of precipitation data from ERA-Interim. ERA-Interim is of similar quality to the high resolution HIRLAM reanalysis across many of the measures shown here. It is especially good at representing low threshold, i.e. broad scale, events. Such broad scale events are heavily influenced by large scale weather features that are best represented in a global model.

HIRLAM features higher resolution than ERA-Interim, but 3DVAR assimilation. It is of similar quality to ERA-Interim in representing lower threshold events, but surpasses it in representing higher threshold events. HIRLAM's correlation coefficients with the observed number of days with at least 20mm of precipitation are slightly higher than those of ERA-Interim. The improvement in representation of large thresholds demonstrates the need for high resolution regional reanalysis to study extreme precipitation events. The comparatively poor performance at lower thresholds perhaps shows the need for 4DVAR assimilation when studying small scale variables, including precipitation. HIRLAM is a useful dataset for studies of precipitation extremes across a wide domain, but is less useful than other EURO4M datasets for light rain or for more localised studies.

The MESAN downscaler aims to improve representation of precipitation by using OI to merge HIRLAM with rain gauge data. The downscaler is not included in the ETS evaluation since its minimum accumulation period is twenty-four hours, which makes it less suitable for some short period applications including studies of flash-flooding. MESAN shows particular improvement in representing high threshold events and would be therefore very useful for studying fluvial flooding. It appears particularly good at representing precipitation, but the metrics shown here use precipitation observations as truth which are also assimilated into the dataset. Nonetheless, its clear improvement over the reanalyses ERA-Interim, HIRLAM and MO in FSS and monthly statistic correlations demonstrates the potential improvement achievable through rain gauge assimilation.

MESCAN is similar to MESAN, but focuses on narrower scales, and therefore performs similarly to MESAN in most metrics. The version used here is not the final version and features a lack of quality control and uses preliminary tuning of parameters. Even so, it performs similarly to or slightly better than MESAN in FSS and in representing monthly climate statistics, being especially good at representing statistics associated most with extremes: the maximum daily precipitation and the number of days with at least 20mm of precipitation. It performs least well at representing low-threshold statistics: the number of wet days and the maximum wet/dry spell. MESAN and MESCAN perform better than the reanalyses across all thresholds and scales and are therefore suitable for a wide variety of uses. However, the minimum accumulation period is twenty-four hours so they are less able to represent short period events.

Finally, MO uses high resolution 4DVAR, but it is costly so that only a two year period was run for this project.

MO performs consistently the best of the three reanalyses across most ETS. It performs favourably against both ERA-Interim and HIRLAM in FSS, especially at higher thresholds. It also consistently represents the monthly climate statistics well, performing better than ERA-Interim and HIRLAM. MO does not perform as well as MESAN or MESCAN, but the reanalysis could be further improved via precipitation assimilation. MO is useful for a wide range of applications across a range of scales. Its minimum accumulation period is an hour so it is also useful for studying short period events.

Of the four EURO4M datasets, the 'best' precipitation dataset to use will depend on the application. For many applications a combination of datasets is likely to provide better results than a single source, as suggested in [Pena-Arancibia J.L. et al.(2013)].

With thanks to Ric Crocker and Rachel North from the Met Office Verification team for advice and help with fractions skill scores and calculation of SEEPS.

We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (http://ensembles-eu.metoffice.com) and the data providers in the ECA&D project (http://www.ecad.eu).

The research leading to these results has received funding from the European Union, Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 242093 .

A Equitable Threat Score and Frequency Bias

The equitable threat score (ETS), or Gilbert score, is widely used to verify forecasts and was first proposed by [Gilbert G.K.(1884)] to assess forecast events against observed events. The score arises from the following contingency table which counts the occurrences of each event in both the forecast and the observations as follows

-	Event observed	Event not observed
Event forecast	A	B
Event not forecast	C	D

The number of correct forecasts expected by chance is then the number of observed events multiplied by the number of forecast events divided by the total number of events, i.e.

$$A_r = \frac{(A+B)(A+C)}{A+B+C+D} \tag{1}$$

The number of correct forecasts minus those expected by chance is then

$$ETS^* = A - A_r \tag{2}$$

This is normalised to give a score of 1 when $B = 0$ and $C = 0$ (there are no incorrect forecasts) and to be penalised by large values of B and C.

$$ETS = \frac{A - A_r}{A+B+C-A_r} \tag{3}$$

where

$$ETS = \frac{AD - BC}{(B+C)*n + AD - BC} \tag{4}$$

$$n = A+B+C+D \tag{5}$$

The associated frequency bias is simply the ratio of forecast events to observed events, i.e.

$$bias = \frac{A+B}{A+C} \tag{6}$$

B Stable equitable error in probability space

Unlike most other location-based scores, SEEPS is defined in probability space so that skill of representing a certain threshold is dependent on the climatology of the location and susceptibility to observation rounding is reduced [Rodwell M.J. et al.(2010)].

The probability of a dry day occurring is defined as, p_1, the probability of precipitation less than 0.25mm at each station location (note this is a much lower threshold than the climatological concept of wet and dry days). The probability of a wet day is then $\sum_{i=2}^{n} p_i = 1 - p_1$ with the wet events separated into $n-1$ bins of increasing

magnitude with equal probabilities, $p_i = \frac{1-p_1}{n-1}$. The bins are numbered from 1 (dry) to n (the wettest).

In this way wet events are defined using fixed probabilities instead of thresholds. The error of a model at a station location is then given by

$$s_{vf} = \begin{cases} (v-f)a + \delta_{1f}(c-a) & \text{if } v > f \\ (f-v)b + \delta_{v1}(d-b) & \text{if } v < f \\ 0 & \text{if } v = f \end{cases} \tag{7}$$

where v and f are the bin numbers into which the observation and the model are categorised, respectively and a and b are weights to penalise distance from the observed category. If $f = 1$ then $\delta_{1f} = 1$ and if additionally $v > 1$ then $s_{vf} = (v-2)a + c$ where c is a weight to correct for the differently-sized first bin. Likewise if $v = 1$ then $\delta_{v1} = 1$ and if additionally $f > 1$ then $s_{vf} = (f-2)d + c$ where d is a weight to correct for the differently-sized first bin. If $f > 1$ then $\delta_{1f} = 0$ and if $v > 1$ then $\delta_{v1} = 0$.

The constraints on equitability require

$$\sum_{v=1}^{n} p_v s_{vv} = 0, \tag{8}$$

which always holds since $s_{vv} = 0$ for all v, and

$$\sum_{v=1}^{n} p_v s_{vf} = 1 \tag{9}$$

for all f.

Using (7) a table can be constructed of all possible error scores

-	$v=1$	$v=2$	$v=3$	$v=4$	...	$v=n$
$f=1$	0	c	$a+c$	$2a+c$	...	$(n-2)a+c$
$f=2$	d	0	a	$2a$	...	$(n-2)a$
$f=3$	$b+d$	b	0	a	...	$(n-3)a$
$f=4$	$2b+d$	$2b$	b	0	...	$(n-4)a$
...		$(f-2)b+d$	$(f-2)b$	$(f-3)b$	$(f-4)b$	...
$(n-f)a$						
$f=n$	$(n-2)b+d$	$(n-2)b$	$(n-3)b$	$(n-4)b$	...	0

If $n = 2$, (9) implies $c = \frac{1}{p_2}$ and $d = \frac{1}{p_1}$ and, if $n = 3$, (9) implies

$$b = \frac{ap_3}{1-p_3} \tag{10}$$
$$c = \frac{1-ap_3}{1-p_1} \tag{11}$$
$$d = \frac{1-ap_3}{p_1} \tag{12}$$
$$\tag{13}$$

with a unconstrained. If $n = 4$, (9) then implies

$$1 = cp_2 + ap_3 + cp_3 + 2ap_4 + cp_4 \tag{14}$$
$$p_1 = 1 - 2ap_4 - ap_3 \tag{15}$$
$$(b+d)p_1 = 1 - bp_2 - p_4 \tag{16}$$
$$(2b+d)p_1 = 1 - bp_3 - 2bp_2 \tag{17}$$
$$\tag{18}$$

Combining (16), (17) and (18) leads to $(a+b)p_3 = 0$ which is not possible. Therefore SEEPS is only equitable with the relatively trivial two-category case or the three-category case. Using the three-category case, the categories are defined as 'dry', 'light' and 'heavy'. The weighting a is chosen so that the smallest error of a model that never predicts an event in bin 1 or 3 is maximised. This lowest error occurs when the model is always correct at the other event bins and models bin 1 or 3 observations as bin 2 events (probability of such events occurring is p_1 or p_3, respectively). This leads to an expected error of $p_1 d = 1 - p_3 a$ and $p_3 a$, respectively. Then $\min(1 - p_3 a, p_3 a)$ is maximised if $a = \frac{1}{2p_3}$. And so (7) becomes

$$s_{vf} = \begin{cases} (v-f)\frac{1}{2p_3} + \delta_{1f}(\frac{1}{2(1-p_3)} - \frac{1}{2p_3}) & \text{if } v > f \\ (f-v)\frac{1}{2(1-p_3)} + \delta_{v1}(\frac{1}{2p_1} - \frac{1}{2(1-p_3)}) & \text{if } v < f \\ 0 & \text{if } v = f \end{cases} \tag{19}$$

C Fractions Skill Score

The fractions skill score (FSS) was developed by [Roberts N.M. and Lean H.W.(2008)] to assess a model's skill in representing precipitation events at a given spatial scale. The FSS uses gridded observation data as truth and the scale assessed is based on $n \times n$ observation grid-boxes where n is a user-determined integer less than or equal to $2N - 1$ where N is the number of grid-boxes on the longest side of the observation grid.

The model data is reconfigured to observation resolution then, for a given threshold, the fraction of events modelled at scale $n \times n$ centred on grid-box i, j is given by

$$M_{ij}^n = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_M \left[i + k - 1 - \frac{n-1}{2}, j + l - 1 - \frac{n-1}{2} \right] \tag{20}$$

where $I_M[i,j]$ is an integer - 1 if there is an event at i, j and 0 otherwise. The observed fraction is similar, but a convolution kernel is applied to smooth the data to the appropriate scale. The kernel used here, following [Roberts N.M. and Lean H.W.(2008)], is a $n \times n$ mean filter, K_n, but any appropriate smoothing filter should produce similar results.

$$O_{ij}^n = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} I_O \left[i + k - 1 - \frac{n-1}{2}, j + l - 1 - \frac{n-1}{2} \right] K_n(k,l) \tag{21}$$

The mean square error between observed and model fractions is then

$$MSE^n = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(M_{ij}^n - O_{ij}^n \right)^2 \tag{22}$$

A reference error, which is the expected error from randomly rearranging the model's grid-boxes, is

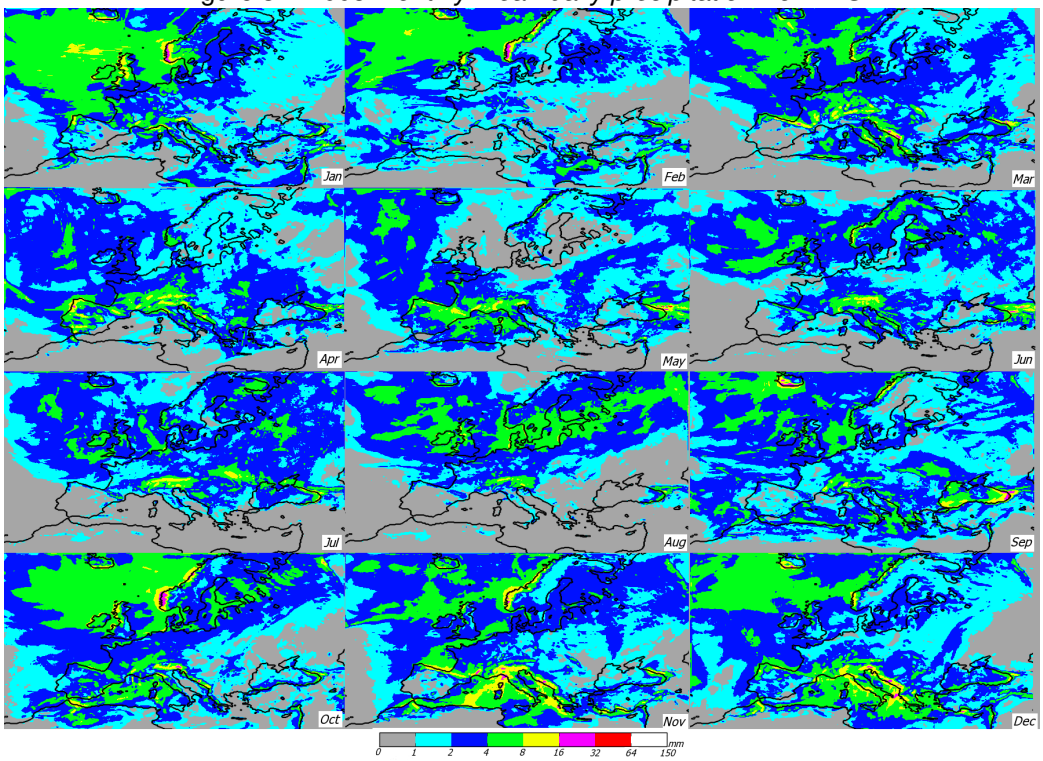
$$MSE_r^n = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(M_{ij}^n \right)^2 + \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \left(O_{ij}^n \right)^2 \tag{23}$$

The fractions skill score is then given by

$$FSS^n = 1 - \frac{MSE^n}{MSE_r^n} \tag{24}$$

which is one minus the mean fraction error normalised by the largest possible fraction error obtainable using this scale, domain and fraction of events. Useful skill is achieved when the FSS is at least $0.5(1 + f_0)$ where f_0 is FSS for a random forecast with the same fraction of events as the truth.

Figure 31: *2008 Monthly mean daily precipitation from MO*



D Monthly Statistics

Figures 31 and 32 show the mean daily precipitation from MO for 2008 and 2009, respectively. Similarly, Figures 33 and 34 show the maximum daily precipitation from MO for 2008 and 2009, respectively. Within 2008, particularly wet months are August, October and July, and particularly dry months are May, February and April. The four driest months in 2008-2009 are April 2009, May 2008, February 2009 and February 2008 and the wettest are July 2009, August 2008, November 2009 and October 2008.

E Spin-up/down in MO

A common problem in numerical weather prediction is that of spin-up/down, in which unrealistic levels of precipitation occur at the start of the forecast. This is sufficiently problematic in ERA-40 that the initial accumulation period does not produce the best representation of precipitation and an offset is suggested for reliable precipitation products [Kallberg P.(2001)]. Representation of precipitation in ERA-Interim is a substantial improvement on this so that the recommended accumulation period is the initial (re)forecast period [Kallberg P.(2011)].

To assess potential spin-up/down in MO, typical three-hourly accumulations of total precipitation and its components for land and sea are displayed for this reanalysis and ERA-Interim, against increasing (re)forecast time, in Figure 35. Typical values were created by taking the mean value from each analysis time on the fourth day of every month in 2008. Figure 35 suggests that both reanalyses feature spin-up over land and spin-down over sea. MO features smaller spin-up over land and larger spin-down over sea than ERA-Interim, but overall this is less than that of ERA-Interim.

Figure 32: *2009 Monthly mean daily precipitation from MO*

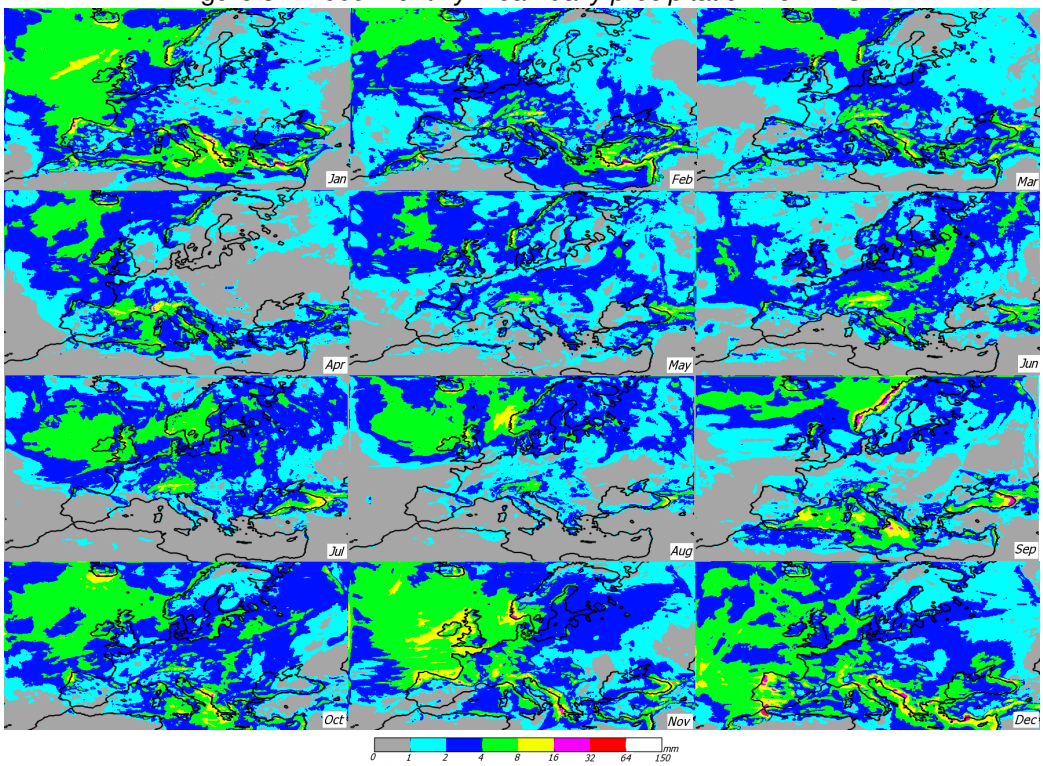


Figure 33: *2008 Monthly maximum daily precipitation from MO*

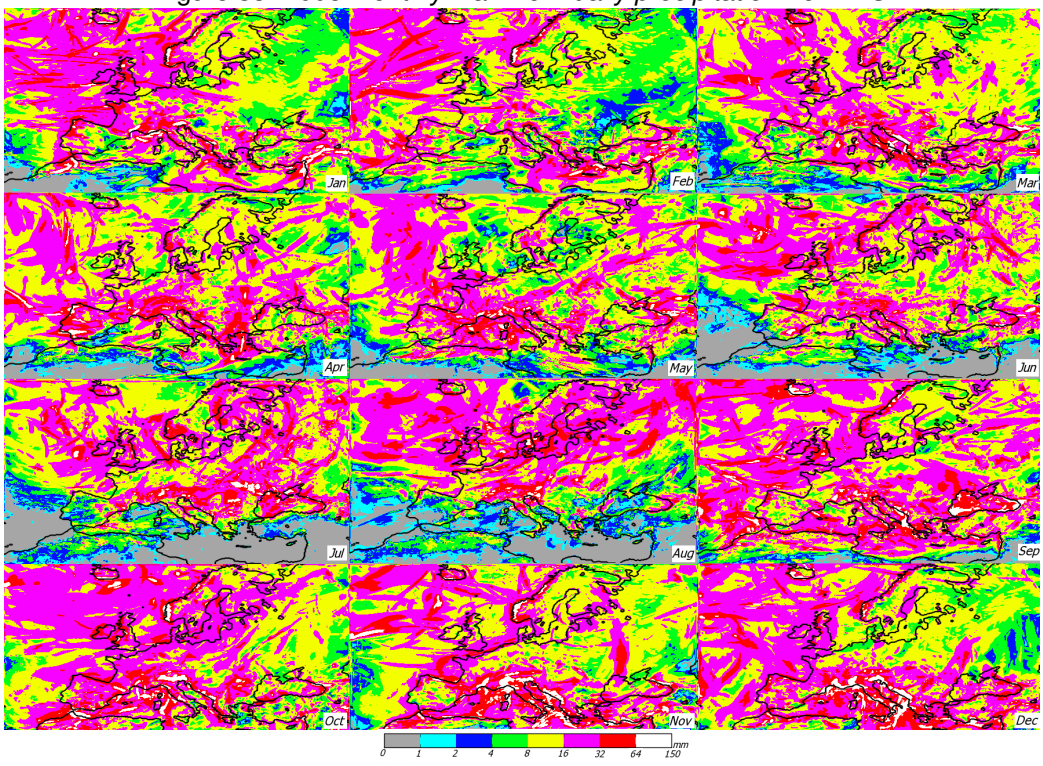


Figure 34: *2009 Monthly maximum daily precipitation from MO*

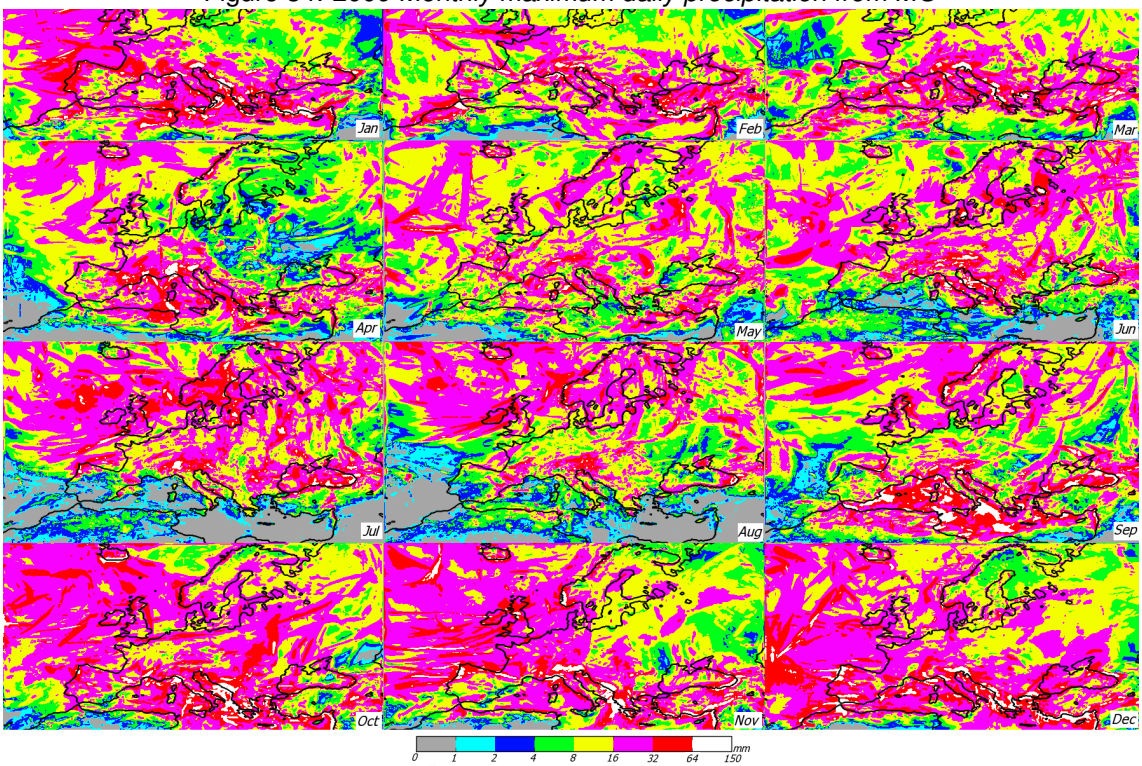


Figure 35: *Typical accumulations of precipitation against forecast time: three-hourly accumulations of MO (red) and ERA-Interim (blue) for land (dashed), sea (dotted) and both (solid). The plot also includes the equivalent value from the median station of the ECA&D observations (green).*

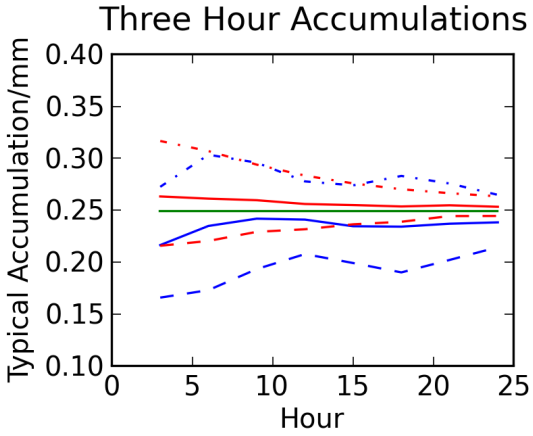


Figure 36: *Bias (L) and ETS (R) for MO. January (Red) and July (Green) 2008.*

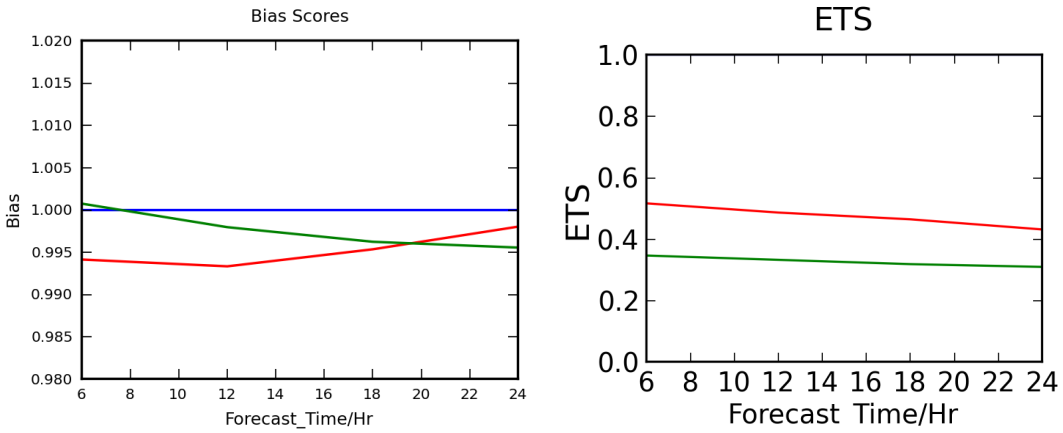


Figure 36 displays the ETS and frequency bias of six-hour accumulations from MO at 4mm threshold against increasing forecast time for January and July 2008. This figure suggests that the bias does significantly vary with increasing forecast time and also that ETS decreases with increasing forecast time. Although there is some spin-up/down associated with MO it is not sufficiently large that the initial forecast period does not produce the highest quality precipitation accumulations. Overall spin-up/down is smaller than in ERA-Interim. Therefore accumulation periods which start at the analysis time are recommended for most products.

F Met Office Downscaler and Climate-style runs

Comparison of MO-D with MO demonstrates the effects that high resolution 4DVAR data assimilation has on the quality of the reanalysis. Comparison of MO-C with MO, or ERA-Interim, demonstrates the necessity to constrain the reanalysis to observations either through the global ERA-Interim reanalysis (used as the background in MO and MO-D) or through assimilation. A comparison of representation of precipitation in MO and its variants is given in Section 5.3. A comparison of representation of other variables is given here.

To compare the quality of the continuous variables temperature, wind speed, relative humidity and pressure, the skill of the six hour forecast is used. This is defined as

$$\text{skill} = \frac{\text{RMSE(persistence)}^2 - \text{RMSE(forecast)}^2}{\text{RMSE(persistence)}^2}. \tag{25}$$

F.1 Temperature and Wind Speed

Figures 37 and 38 show temperature and wind speed skill for MO, its variants and ERA-Interim for 1000HPa - 250HPa and at the surface. Generally MO out-performs ERA-Interim and, as expected, MO-C performs least well throughout. At the surface MO-D does not perform as well as ERA-Interim, but higher in the atmosphere it out-performs both ERA-Interim and MO. These results suggest that higher resolution improves temperature and wind speed representation, but downscaling with no additional observation assimilation will not produce

Figure 37: *Monthly temperature skill scores 2008 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black) and MO-D (grey). Surface (top L), 1000HPa (top R), 850HPa (middle L), 500HPa (middle R) and 250HPa (bottom)*

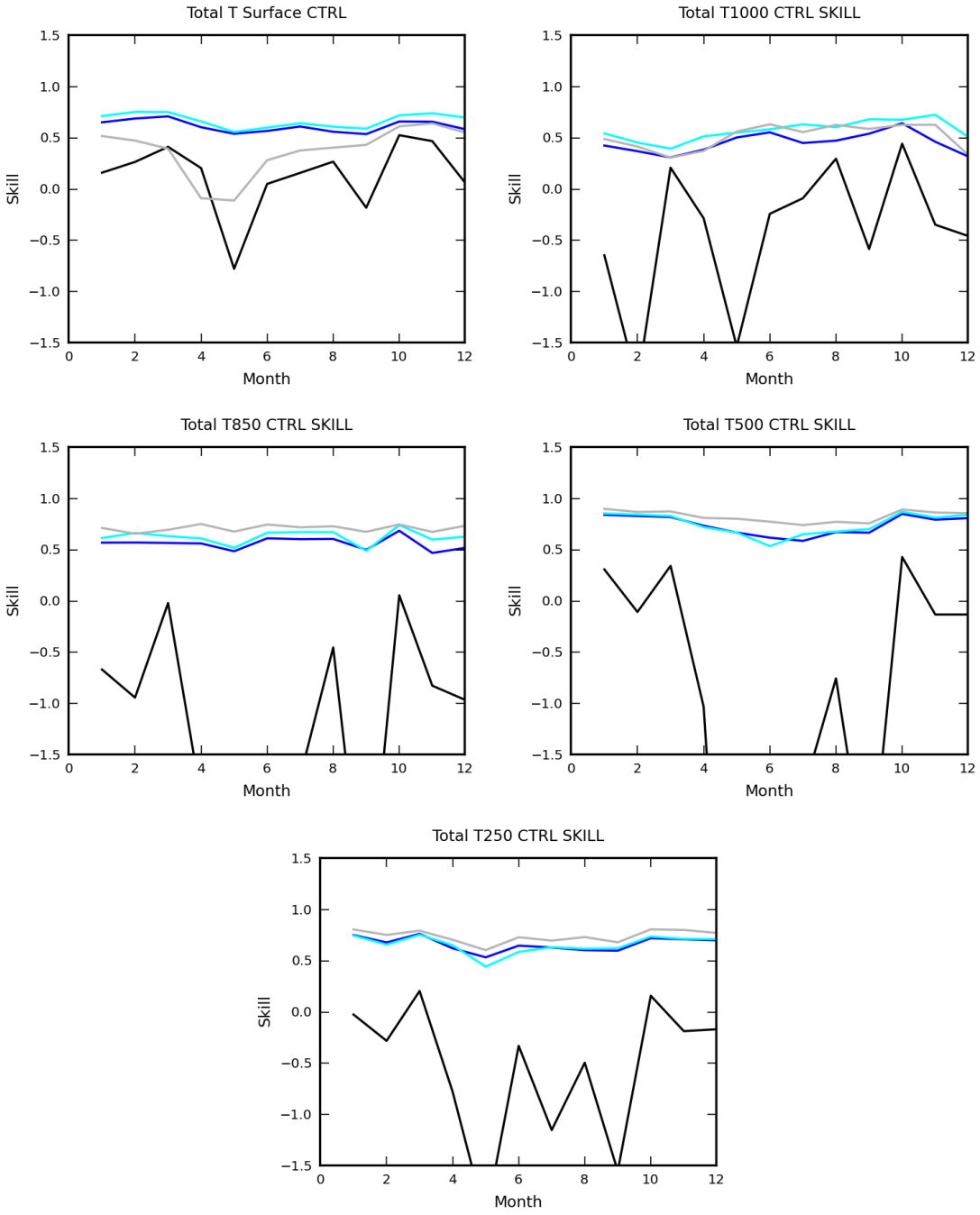


Figure 38: *Monthly wind speed skill scores 2008 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black) and MO-D (grey). Surface (top L), 1000HPa (top R), 500HPa (bottom L) and 250HPa (bottom R).*

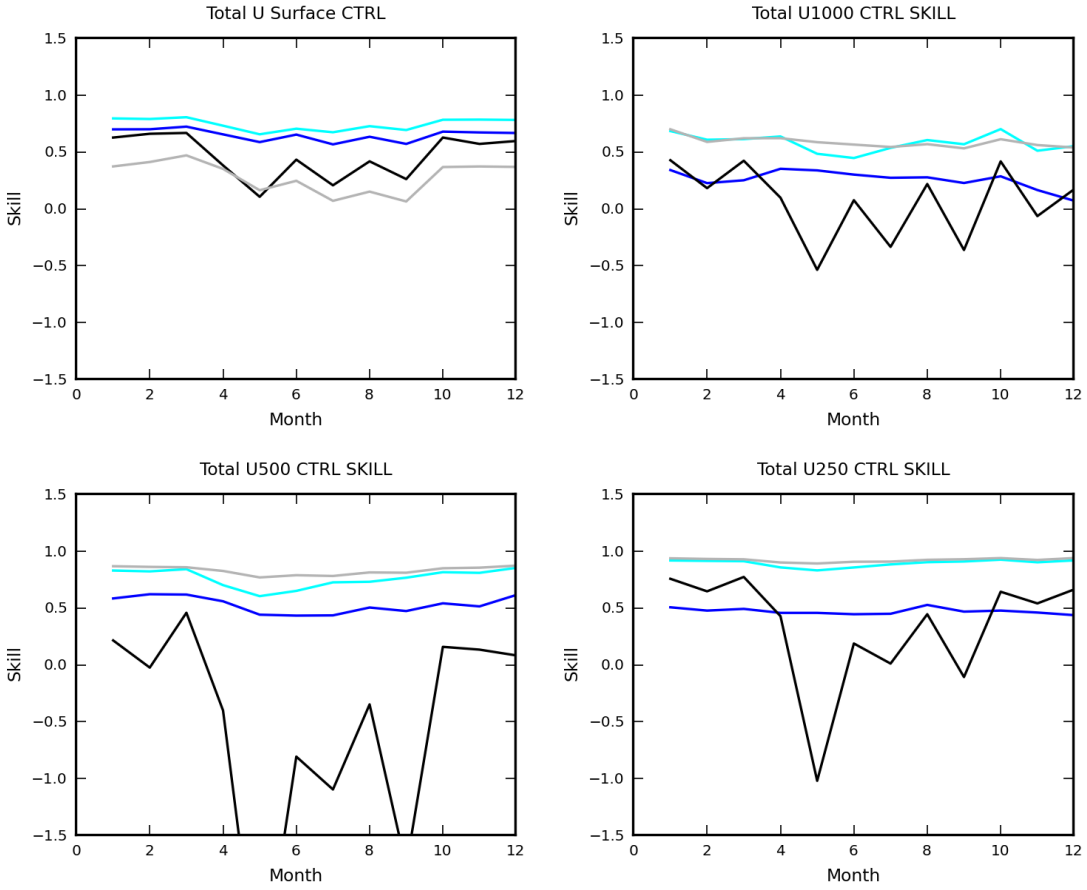
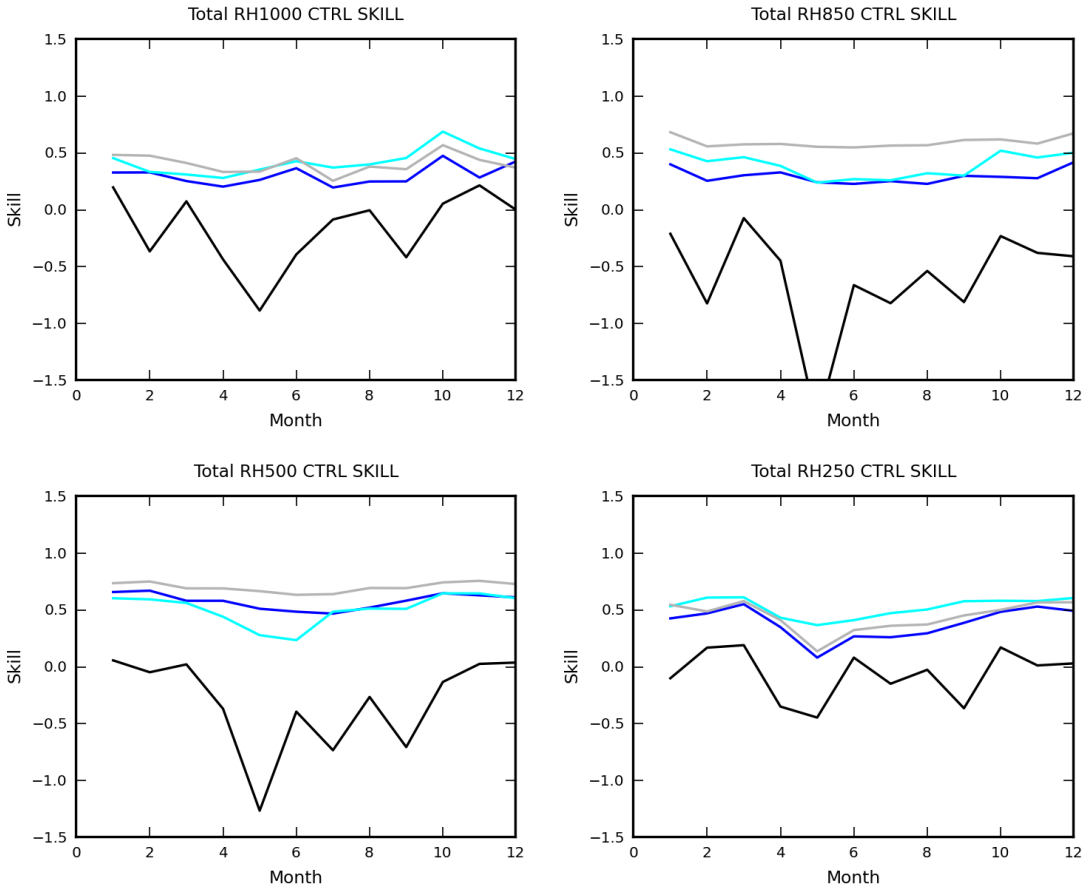


Figure 39: *Monthly relative humidity skill scores 2008 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black) and MO-D (grey). Surface (top L), 1000HPa (top R), 850HPa (middle L), 500HPa (middle R) and 250HPa (bottom)*



accurate results at the surface.

F.2 Relative Humidity

Figure 39 displays skill scores for relative humidity. Generally MO out-performs ERA-Interim except at 500HPa. MO-D improves on ERA-Interim at all heights and at 850HPa and 500HPa it also out-performs MO.

F.3 Pressure

Figure 40 displays skill scores for mean sea level pressure. MO and MO-D are of similar skill to ERA-Interim, indicating that regional models add nothing to representing this larger scale variable.

F.4 Cloud

Figure 41 shows ETS for cloud. As with other variables, MO is generally an improvement on ERA-Interim and MO-C and MO-D is the worst performing dataset throughout. When there is a large amount of cloud MO-D performs similarly to MO, but when there is a smaller amount of cloud it performs less well than ERA-Interim. This suggests that downscaling increases the amount of cloud in the dataset.

Figure 40: *Monthly mean sea level pressure skill scores 2008 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black) and MO-D (grey).*

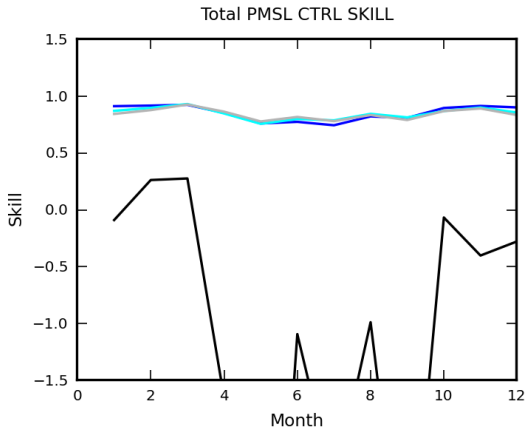
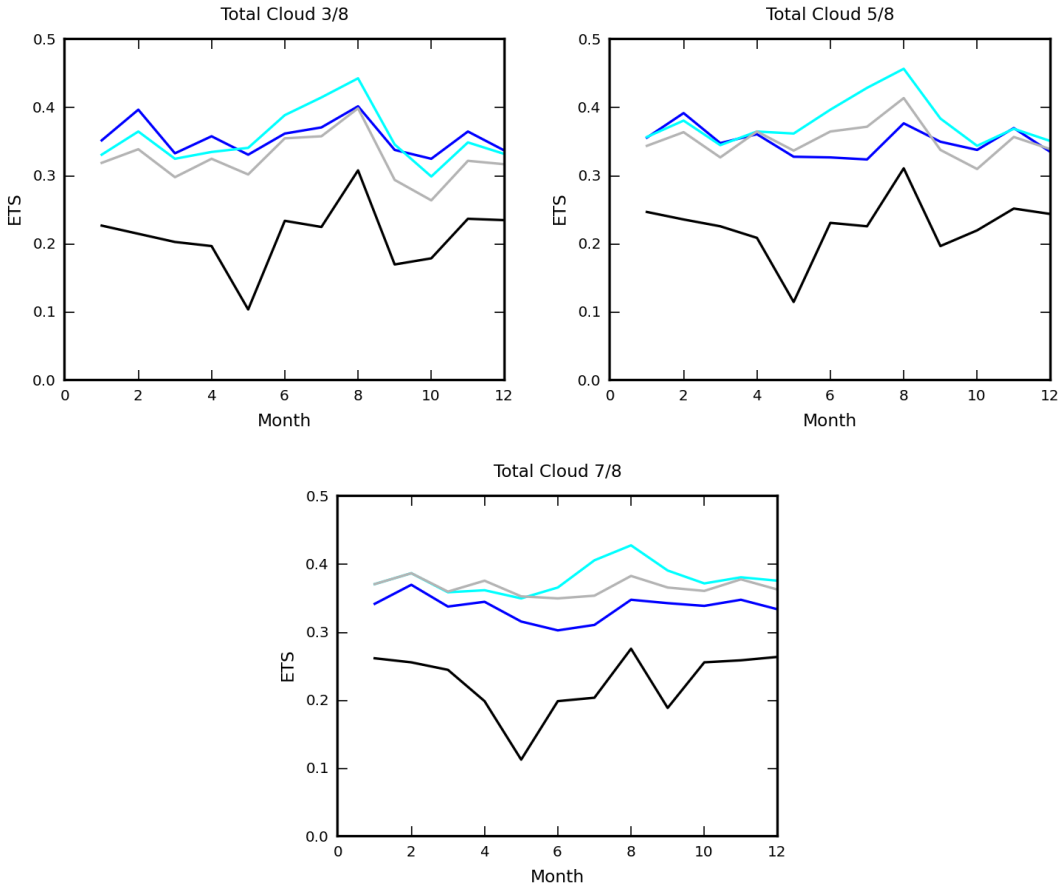


Figure 41: *Monthly cloud equitable threat scores 2008-2009 for ERA-Interim (darker blue), MO (lighter blue), MO-C (black) and MO-D (grey). 3 oktas (top L), 5 oktas (top R) and 7 oktas (bottom).*



References

Akhtar M et al. 2009. Use of regional climate model simulations as input for hydrological models for the Hindukush-Karakorum-Himalaya region. *Hydrol. Earth Syst. Sci.* **13**: 1075–1089

Akhtar M et al. 2011. https://code.zmaw.de/projects/cdo. *Max-Planck-Institut* **1.5.3**

Andersson A et al. 2010. The Hamburg ocean atmosphere parameters and fluxes from satellite data - HOAPS-3. *Earth System Science Data* **2**: 215–234

Becker A et al. 2013. A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901-present. *Earth Syst. Sci. Data* **5**: 71–99

Betts AK et al. 2006. Comparison of ERA40 and NCEP/DOE near-surface data sets with other ISLSCP-II data sets. *J. Geophys. Res.* **111**: 1–20

Chan SC et al. 2013. Does increasing the spatial resolution of a regional climate model improve the simulated daily precipitation? *Climate Dynamics* **41**: 5–6

Dahlgren P and Gustafsson N. 2012. Assimilating host model information into a limited area model. *J. Geophys. Res (Atmospheres)* **64**: 1–17

Dee DP et al. 2011. The ERA-Interim reanalysis. *Q. J. R. Meteorol. Soc.* **137**: 553–597

Done J et al. 2004. The next generation of NWP: Explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmos. Sci. Lett* **5**: 110–117

Ghelli A and Lalaurette F. 2000. Verifying precipitation forecasts using upscaled observations. *ECMWF Newsletter* **87**: 9–17

Gilbert GK. 1884. Finley's tornado predictions. *Amer. Meteor. J.* **1**: 166–172

Haggmark L et al. 2000. MESAN, an operational mesoscale analysis system. *Tellus* **52A**: 2–20

Harris I et al. 2013. Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 dataset. *Int. J. Climatol.*

Kallberg P. 2001. An overview of the ERA-40 analyses. *ERA-40 Project Report Series* **3**: 31–40

Kallberg P. 2010. The use of reanalysis data for monitoring the state of the climate in The State of the Climate in 2010. *Bull. Amer. Meteor. Soc.* **92**

Kallberg P. 2011. Forecast drift in ERA-Interim. *ERA report series* **10**

King AD et al. 2012. The efficacy of using gridded data to examine extreme rainfall characteristics: a case study for Australia. *Int. J. Climatol.* **33.10**: 2376–2387

Lanciani A et al. 2008. A multiscale approach for precipitation verification applied to the FORALPS case studies. *Adv. Geosci* **16**: 3–9

Lopez P. 2013. Experimental 4D-Var assimilation of SYNOP rain gauge data at ECMWF. *Bull. Amer. Meteor. Soc.* **141**: 1527–1544

Mass C et al. 2002. Does increasing horizontal resolution produce more skillful forecasts? the results of two years of real-time numerical weather prediction over the pacific northwest. *Bull. Amer. Meteor. Soc.* **83**: 407–430

Mittermaier M and Roberts N. 2010. Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Q. J. R. Meteorol. Soc.* **25**: 343–354

Pena-Arancibia JL et al. 2013. Evaluation of precipitation estimation accuracy in reanalyses, satellite products and an ensemble method for regions in Australia and South and East Asia. *J Hydrol.* **14**: 1323–1332

Pfeifroth U et al. 2013. Evaluation of satellite-based and reanalysis precipitation data in the tropical pacific. *J. Appl. Meteor. Climatol.* **52**: 634–644

Rawlins F et al. 2007. The Met Office global four-dimensional variational data assimilation scheme. *Q. J. R. Meteorol. Soc.* **133**: 347–362

Roberts NM and Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.* **136**: 78–97

Rodwell MJ et al. 2010. New equitable score for precipitation in NWP. *Q. J. R. Meteorol. Soc.* **136**: 1344–1363

Salamon AL et al. 2013. Advances in pan-European flood hazard mapping. *Hydrol. Process.* **27**

Stratman R et al. 2013. Use of multiple verification methods to evaluate forecasts of convection from hot- and cold-start convection-allowing models. *Bull. Amer. Meteor. Soc.* **28**: 119–138

Thieblemont RYJ et al. 2013. A climatology of frozen-in anticyclones in the spring arctic stratosphere over the period 1960-2011. *J. Geophys. Res. Atmos.* **118**: 1299–1311

van den Besselaar EJ et al. 2011. A European daily high-resolution observational gridded data set of sea level pressure. *J. Geophys. Res.* **116**: 1–11

van Engelen A et al. 2013. http://eca.knmi.nl. *European Climate Assessment & Dataset*

Vila D et al. 2009. Improved global rainfall retrieval using the special sensor microwave imager (SSM/I). *J. Appl. Meteor. Climatol.* **49**: 1032–1043

Vormoor K and Skaugen T. 2013. Temporal disaggregation of daily temperature and precipitation grid data for norway. *Bull. Amer. Meteor. Soc.* **14**: 989–999

Yeh PJ-F and Famiglietti JS. 2008. Regional terrestrial water storage change and evapotranspiration from terrestrial and atmospheric water balance computations. *J. Geophys. Res.* **113**: 1–13

Zhang X. 2013. http://etccdi.pacificclimate.org. *ETCCDI/CRD Climate Change Indices*

Zolina O et al. 2004. Analysis of extreme precipitation over Europe from different reanalyses: a comparative assessment. *Global and Planetary Change* **44**: 129–161

Met Office
FitzRoy Road, Exeter
Devon, EX1 3PB
UK

Tel: 0870 900 0100
Fax: 0870 900 5050
enquiries@metoffice.gov.uk
www.metoffice.gov.uk