Seventh Framework Programme
Theme 6 [SPACE]

Project: 607193 UERRA

Full project title:
Uncertainties in **E**nsembles of **R**egional **Re-A**nalyses

Deliverable D2.14
Reanalysis Uncertainty Evaluation

WP no:	2
WP leader:	MO
Lead beneficiary for deliverable :	MO
Name of author/contributors:	Peter Jermey, Jelena Bojarova, Maarit Lockhoff, Richard Renshaw, Martin Ridal, Per Unden
Nature:	Report
Dissemination level:	PU
Deliverable month:	45
Submission date: October 31st 2017	Version nr: 1

Reanalysis Uncertainty Evaluation

Peter Jermey, Jelena Bojarova, Maarit Lockhoff, Richard Renshaw,
Martin Ridal and Per Unden

October 30, 2017

Reanalysis	Model	Assimilation	M	Grid	Doc.
UERRA-UB	COSMO	Nudging	20	0.11°	[Bach, 2015]
UERRA-MO	Unified Model	3DVAR	20	0.33°	[Jermey et al., 2015]
CERA-20C	ECMWF Coupled Model	4DVAR	10	1.1°	[Laloyaux et al., 2016]

Table 1: Summary of ensemble reanalyses. The ensemble size is given by M. COSMO is the Consortium for Small Scale Modelling and HIRLAM is the High Resolution Limited Area Model. Assimilation is carried out either by three or four dimensional variational assimilation (3DVAR/4DVAR) or by the ensemble Kalman filter (EnKF).

1 Introduction

The Uncertainties of Ensembles in Regional Reanalyses (UERRA), [Unden et al., 2014], includes two 20 member ensemble reanalyses covering the entire European domain. The first of these is produced by the University of Bonn in collaboration with Deutscher Wetterdienst (UERRA-UB) covering 2006-2010. The second is produced by the Met Office (UERRA-MO) covering 1979-2016. The spread of these ensembles is intended to be a useful metric of reanalysis uncertainty, which is a focus of the UERRA project. This document aims to evaluate the UERRA reanalyses, with an emphasis on uncertainty estimation, by comparing them to the 20th century ensemble reanalysis (CERA-20C) from the European Centre for Medium Range Weather Forecasts (ECMWF). A summary of the three reanalyses is given in table 1, which also references documentation describing their development.

The reanalyses are evaluated in this document using standard metrics against daily observations from the European Climate Assessment & Dataset (ECAD), [Klein Tank et al., 2002]. For 2m temperature and total precipitation, these observations are independent of the reanalyses. Use of the observations and metric calculations are detailed in appendices A and B, respectively. Results are shown and discussed for daily mean 2m temperature (TG), daily mean 10m wind speed (FG) and daily total precipitation (RR) in sections 2, 3 and 4, respectively. Finally, conclusions are given in section 5.

2 Daily Mean Temperature at 2m

Multiplicative inflation is used on each ensemble before comparison with observations to take account of representivity error and observation error, [Saetra et al., 2004]. These factors and the implied combined error variance that they represent is given in table 2 for daily mean 2m temperature.

The RMSE of the control and the RMSE of the ensemble mean for each reanalysis against observations of daily mean 2m temperature are given in figures 1 and 2, for June to August 2007 (summer) and December 2006 to February 2007 (winter), respectively. A random member of

Reanalysis	Spread Inflation factor (JJA/DJF)	Error Variance (JJA/DJF)
UERRA-UB	2.4/3.5	0.72/1.30
UERRA-MO	1.8/2.1	0.86/1.49
CERA-20C	4.6/4.5	1.47/1.73

Table 2: Inflation factors for daily mean 2m temperature.

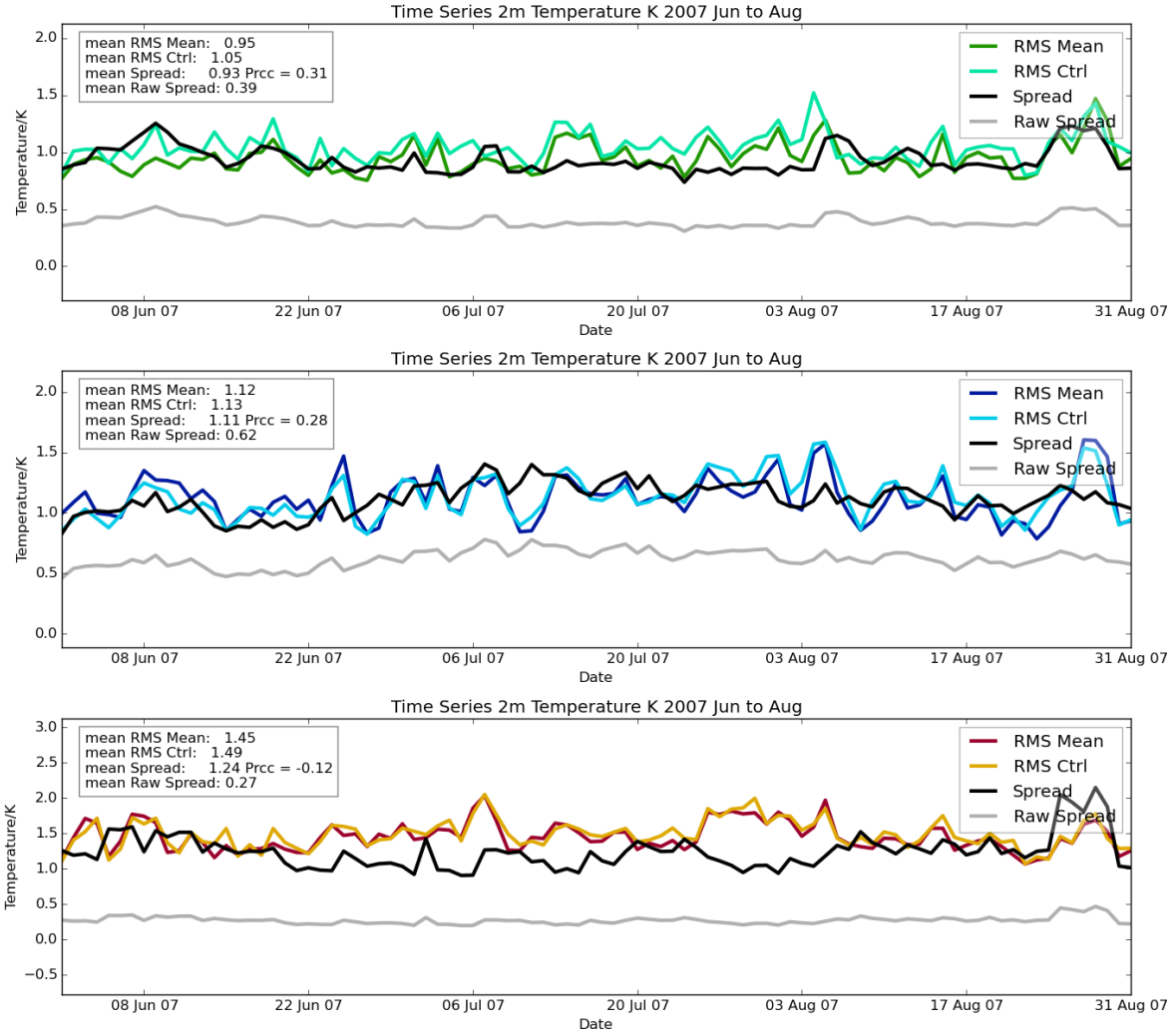

Figure 1: RMSE and spread in summer months (JJA) for daily mean temperature at 2m in 2007. The plots show RMS difference of Ensemble mean with observations, RMS difference of control with observations, raw spread of ensemble and spread of ensemble inflated to take account of representivity and observation errors. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. A random member of CERA-20C is used as a proxy for a control member.

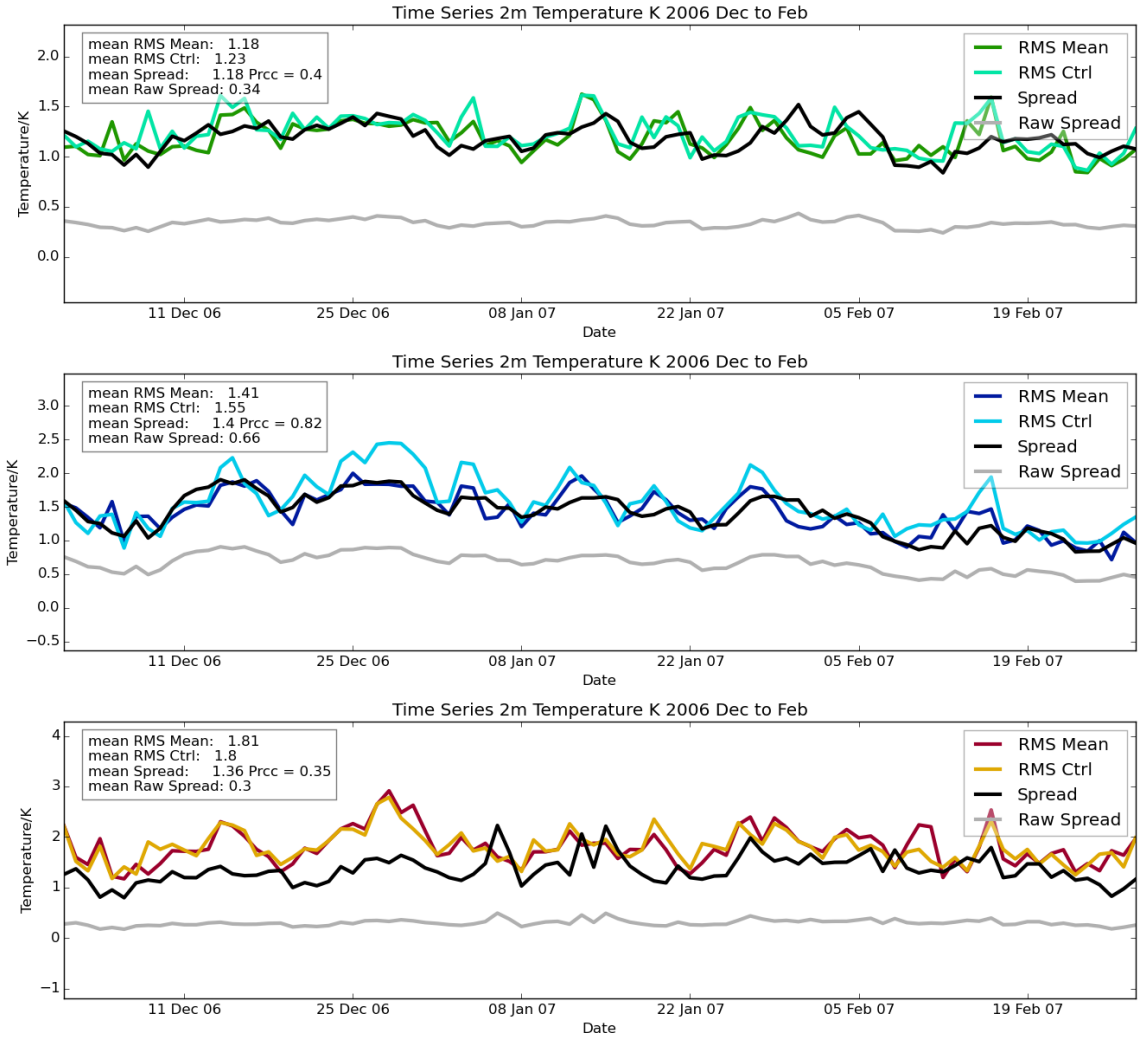

Figure 2: RMSE and spread in winter months (DJF) for daily mean temperature at 2m. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 1.

Summer	RMSE	Bias	ERR	IRR	CRPS	Prcc
UERRA-UB	0.80 (0.95)	0.13 (-0.04)	1.86 (1.86)	1.0 (1.17)	0.05 (0.06)	0.25 (0.31)
UERRA-MO	1.00 (1.12)	0.56 (0.43)	1.52 (1.76)	0.57 (0.64)	0.07 (0.07)	0.36 (0.28)
CERA-20C	1.12 (1.45)	-0.48 (-0.12)	1.38 (1.75)	1.77 (1.31)	0.14 (0.18)	-0.10 (-0.12)
Winter	RMSE	Bias	ERR	IRR	CRPS	Prcc
UERRA-UB	1.00 (1.18)	0.03 (0.06)	1.85 (1.73)	1.10 (0.97)	0.06 (0.08)	0.49 (0.40)
UERRA-MO	1.15 (1.41)	0.69 (-0.06)	1.74 (1.69)	0.37 (0.83)	0.07 (0.09)	0.67 (0.82)
CERA-20C	1.57 (1.81)	-0.94 (-1.13)	1.97 (2.02)	1.97 (2.09)	0.18 (0.21)	0.28 (0.35)

Table 3: Table comparing daily mean 2m temperature ensemble performance over France. Results for the entire domain are given in brackets. RMSE and bias of ensemble mean are shown. ERR and IRR are the external and internal rank ratios, respectively, from the rank histograms. Prcc is the Pearson's rank correlation coefficient between ensemble spread and RMSE of the ensemble mean.

the CERA-20C ensemble is used as a proxy for a control. As expected the UERRA reanalyses improve on CERA-20C, with the higher resolution UERRA-UB also consistently out-performing UERRA-MO. For an ensemble to be well formed, the error of the mean should be smaller than the error of the control, [Leith, 2007]. In this case all the ensemble means have lower errors than the control or are of a similar size.

Figures 1 and 2 also show ensemble spread (both raw and inflated) for the two seasons, summer and winter, respectively. Since the ensemble spread is intended as a measure of uncertainty in the ensemble, ideally the ensemble spread should match the RMSE of the mean, once observation and respresentivity errors are accounted for, [Grimit and Mass, 2007]. The relative quality of the spread of the ensembles is here compared using the Pearson's rank correlation coefficient (Prcc), which shows that both UERRA ensembles spreads are better estimators of the uncertainty than the spread of CERA-20C. In summer the spread of UERRA-UB is a better estimator of uncertainty and in winter the spread of UERRA-MO is best.

Similar results have been calculated for each reanalysis over France, instead of the whole domain and these are summarised in table 3. The results show that the ensemble means all have lower RMSEs over France than over the whole domain, with both regional reanalyses showing improvement over CERA-20C. In summer, correlation between spread and RMSE of ensemble mean is improved for UERRA-MO and that of UERRA-UB sees a reduction, suggesting that the spread of UERRA-MO is the better estimate of uncertainty in this region. In winter correlation between spread and RMSE of ensemble mean is an improvement over that of the whole domain for UERRA-UB, but the spread correlation of UERRA-MO remains the largest of the three reanalyses.
Figures 3 and 4 show mean error of the controls and ensemble means for the three reanalyses. These suggest that bias is generally smaller in the regional ensembles than in the global ensemble, except for UERRA-MO in summer. These figures also demonstrate that for both CERA-20C

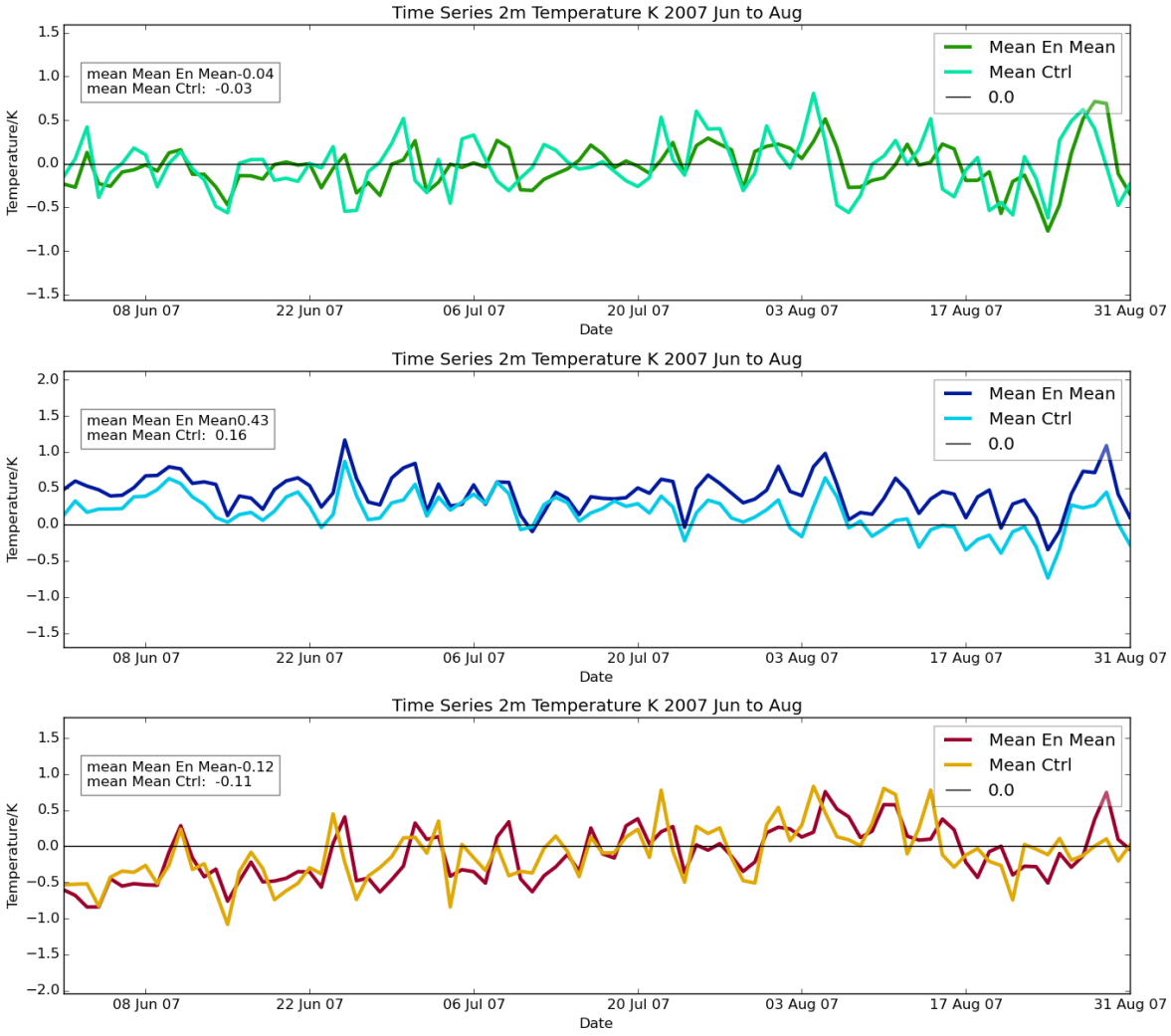

Figure 3: Bias in summer for daily mean temperature at 2m. The plots show mean difference of Ensemble mean with observations and mean difference of control with observations. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. A random member of CERA-20C is used as a proxy for a control member.

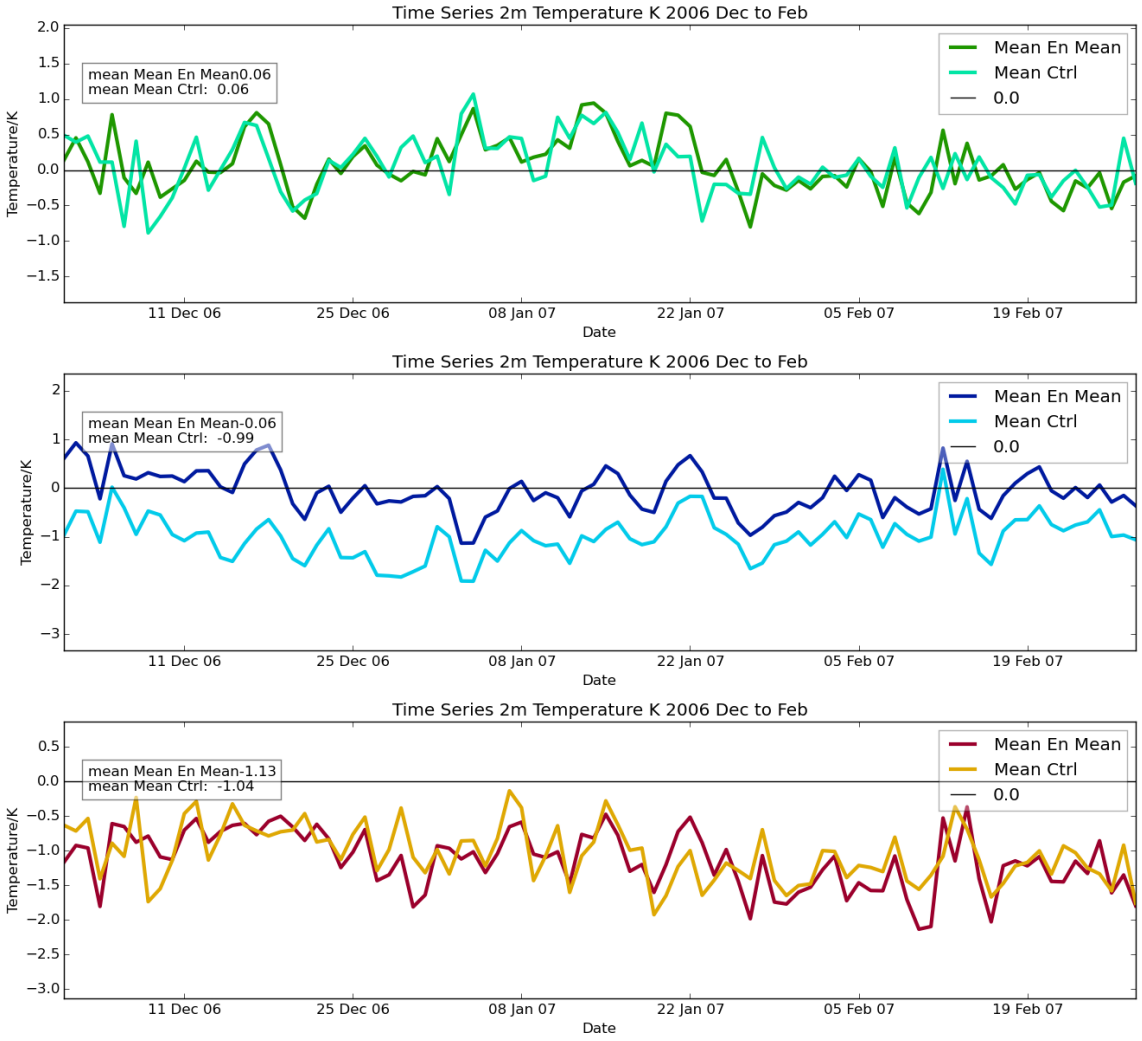

Figure 4: Bias in winter for daily mean temperature at 2m. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 3.

and UERRA-UB the mean bias and the control bias are similar, but UERRA-MO has a warmer ensemble mean than control. In summer, the control of UERRA-MO is also too warm so the bias is increased in the ensemble mean, but in winter the control is too cold so the bias is decreased in the mean. Results over France also show that the UERRA reanalyses have smaller biases than CERA-20C, except UERRA-MO in summer, see table 3. In summer, UERRA-UB and CERA-20C both see an increase in bias in this region compared with the entire domain, and in winter, they both see a decrease. UERRA-MO sees an increase in bias during both periods.

Figure 5 shows rank histograms for the three ensembles for each of the two periods, summer and winter. The external rank ratio is the ratio between the average proportion of the two external ranks to the average proportion of the internal ranks. Perfect spread would result in a score of 1.0. The internal rank ratio is the ratio between the higher half of internal ranks with that of the lower half of internal ranks. An unbiased ensemble would also result in a score of 1.0. For both seasons the spread of the UERRA-UB ensemble is of superior quality to that of UERRA-MO, both being slightly underspread, with UERRA-UB ensemble members being somewhat too cool in summer and too warm in winter. UERRA-MO members are also too warm in both seasons. CERA-20C exhibits a large cool bias in both periods. Restricting the validation region to France yields similar results to that of the whole domain, as shown in table 3.

Figures 6 and 7 show the continuous ranked probability scores (CRPS) for the three ensembles across summer and winter, respectively. The CRPS is a measure of the accuracy of the ensemble and is a difference between the cumulative distribution function of the ensemble and of that implied by the observations. In both months, both the regional ensembles show a much reduced CRPS compared to CERA-20C, indicating that temperature probabilities derived from these regional ensembles are more accurate than those of the global ensemble. In both seasons, UERRA-UB is a slight improvement on UERRA-MO. Table 3 summarises CRPS over France. These are similar to the results over the whole domain.

The Brier score, as shown in figure 8, is a measure of the distance between the modelled probability of an event and the observed outcome. A perfect score is zero. Uncertainty varies from zero (an event always or never occurs) to 0.25 (an event occurs exactly half of the time). Reliability is a measure of how well the modelled probabilities match observed frequencies, again with a perfect score of zero. The resolution is a measure of how much the observations vary from climatology. A resolution of zero means the event always or never occurs. The Brier Score, together with its components is displayed for the three ensembles for both periods in figure 8. The Brier Score for various events (e.g. $TG < 293.15K$) is plotted against an x axis of uncertainty. The uncertainty is also plotted as a diagonal line. The third line is reliability-plus-uncertainty. The distance between this line and that of the uncertainty gives the reliability of each event and the distance between this line and that of the Brier score gives the resolution of each event.

Figure 8 shows that the Brier score for the ensembles decreases with increasing resolution. The lowest (best) scores are achieved by UERRA-UB, then UERRA-MO, and then CERA-20C.

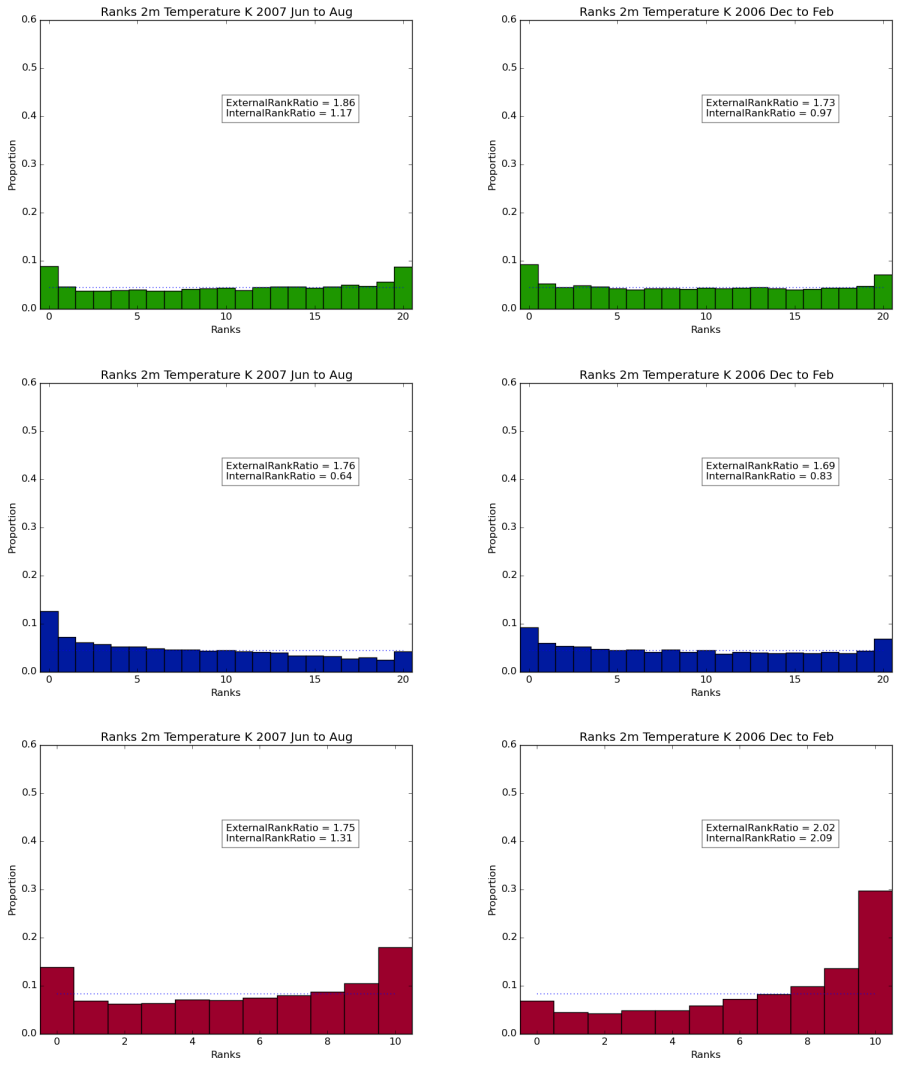

Figure 5: Rank histograms for daily mean temperature at 2m. The plots show histograms of ranks of the observations with respect to the ensemble members. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

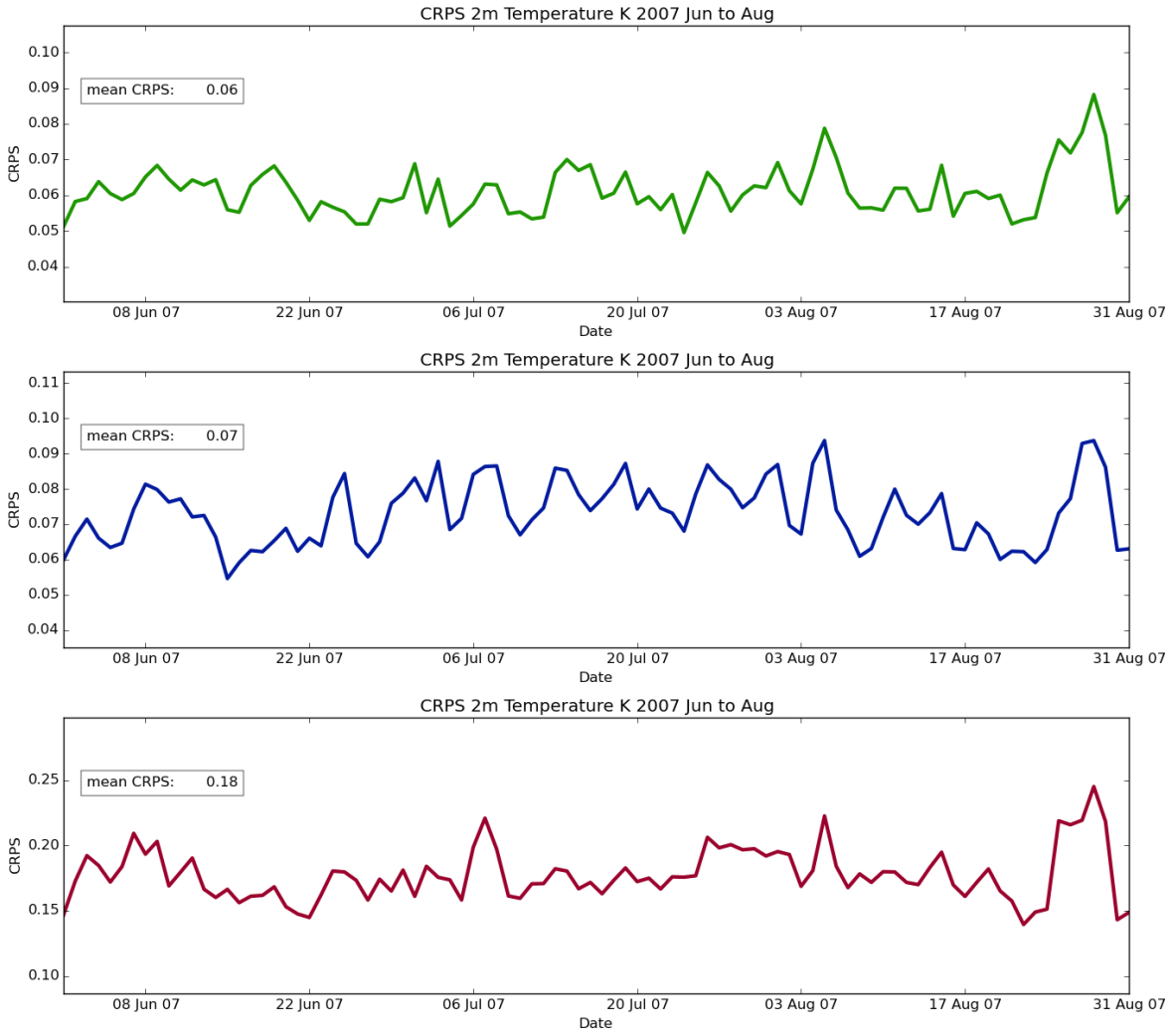

Figure 6: Continuous ranked probability score (CRPS) for daily mean temperature at 2m in summer. The plots show CRPS for the ensembles. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. Scores larger than zero represent discrepancies between the probability distributions of the ensemble and the observation.

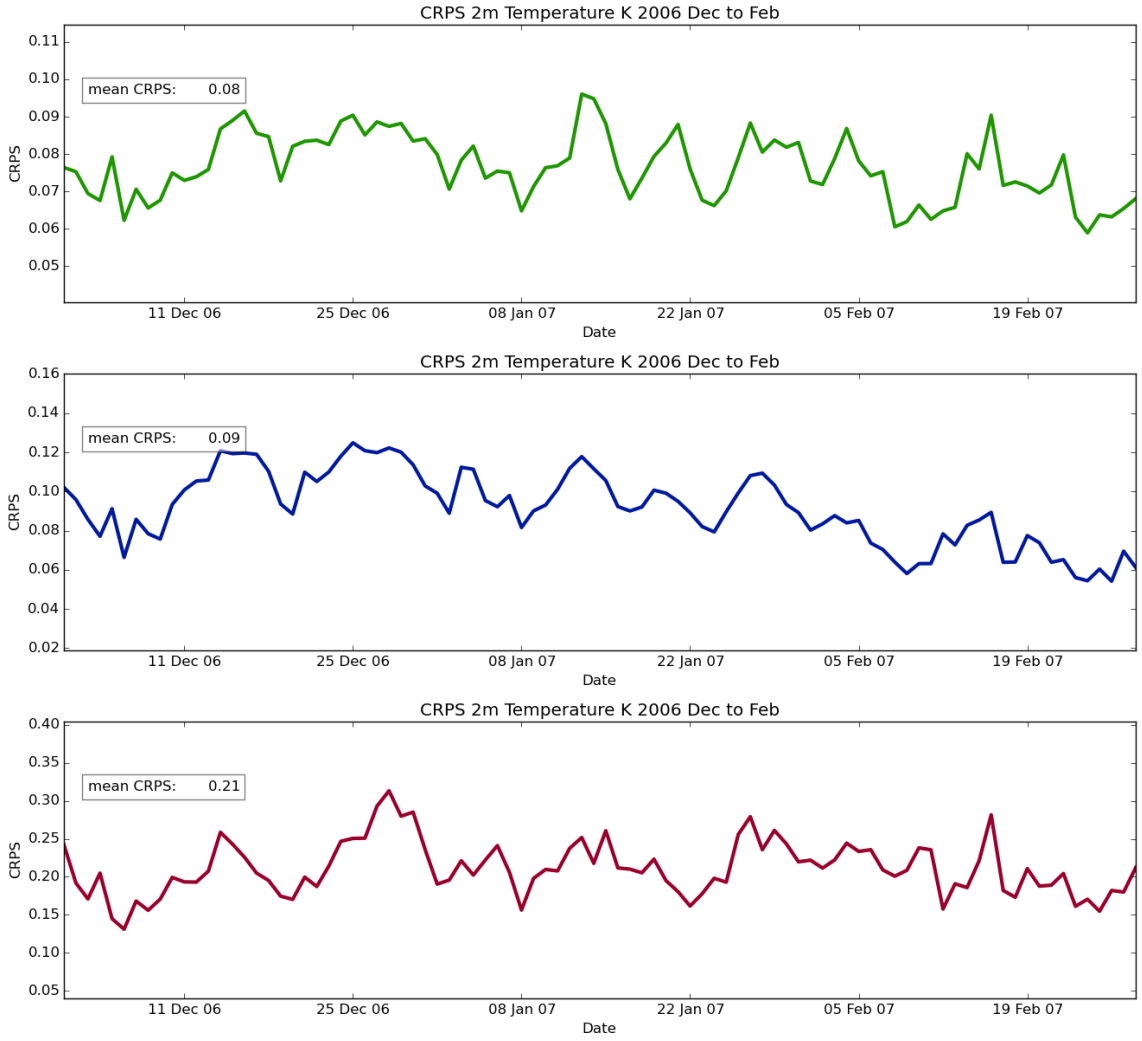

Figure 7: CRPS for daily mean temperature at 2m in winter. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 6.

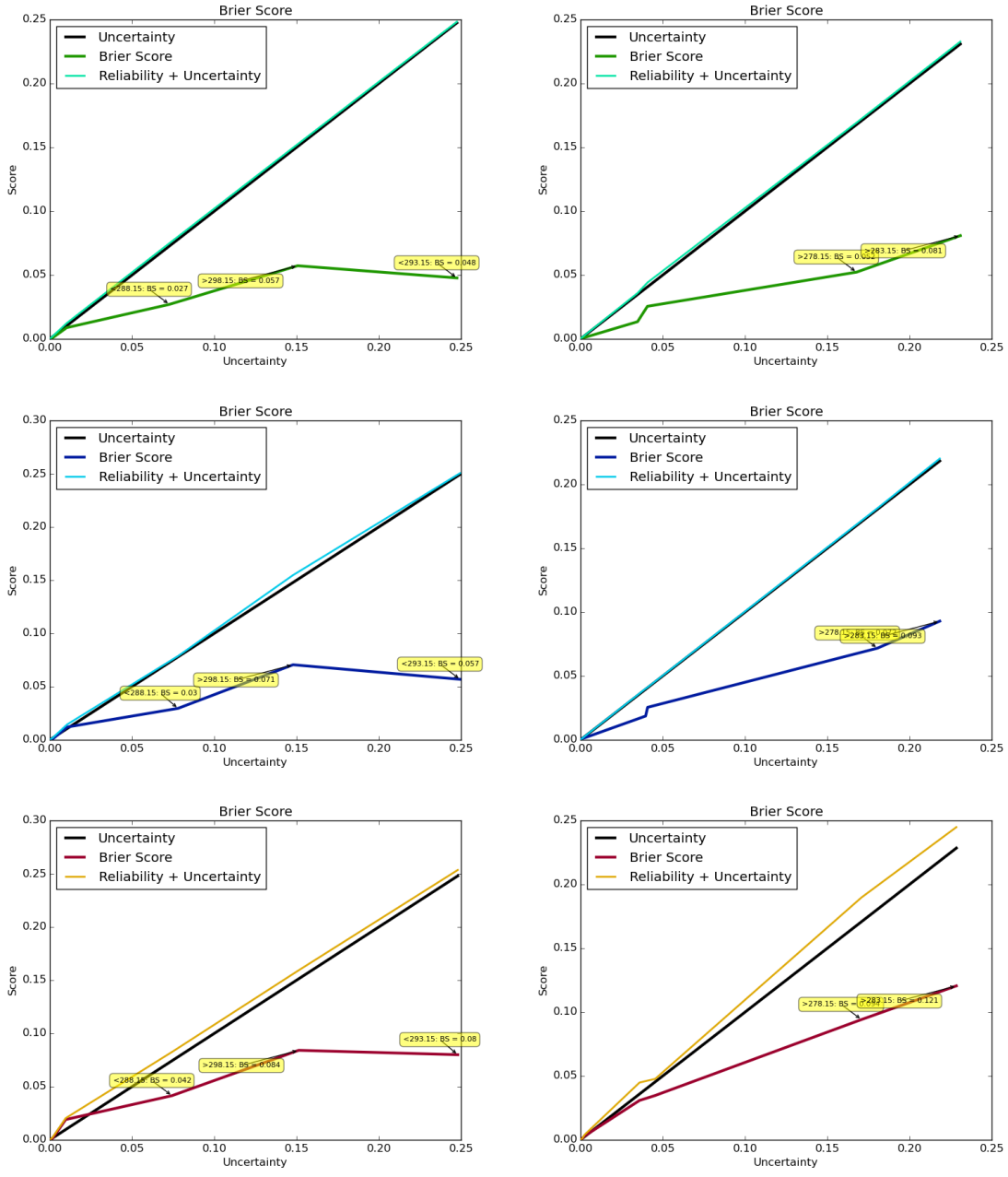

Figure 8: Brier scores and components for daily mean temperature at 2m in 2007. The plots show Brier scores of the ensemble, uncertainty and reliability for various appropriate categories (some labelled). The resolution is given by the distance between the reliability + uncertainty line and the Brier score line. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C.

Reanalysis	Spread Inflation factor (JJA/DJF)	Error Variance (JJA/DJF)
UERRA-UB	3.9/4.7	1.2/1.2
UERRA-MO	4.2/6.0	1.3/2.0
CERA-20C	8.8/12.9	1.5/2.0

Table 4: Inflation factors for daily mean 10m wind speed.

Scores are consistently higher (worse) in winter. This figure also shows that the regional reanalyses have improved reliability over CERA-20C.

The attribute diagrams for each of the ensembles when $TG > 303.15K$ ($TG > 30°C$) in summer are given in figure 9 (LHS). The equivalent receiver operating characteristic curves (ROCs) are also given (RHS). It is by far most common for none of the members to be registering these temperatures at observation locations, as shown by the frequency line of the attribute diagrams. The reliability line for UERRA-UB places 11/21 points between the 'No Skill' and 'Perfect' lines indicating that the ensemble has some skill in representing the probability of these events, especially when the modelled probability is at least 0.8. The same line for UERRA-MO places 5/21 points lying between the skill lines. For CERA-20C, none of the points lie between the lines. These results confirm those suggested by figure 8 that the best reliability is achieved with the highest resolution. The receiver operating characteristic, also shown in figure 9, also shows that the accuracy of representing these temperature events also increases with increasing resolution.

Similar attribute diagrams and ROCs are given in figure 10 when $TG < 273.15K$ ($TG < 0°C$) in winter. For these events, 16/21 points for UERRA-UB have skill, 17/21 points for UERRA-MO have skill and 4/11 points for CERA-20C have skill. The ROC curves again demonstrate that the accuracy of representing these events increases with grid resolution.

3 Daily Mean Wind Speed at 10m

For daily mean 10m wind speed, multiplicative inflation factors and the implied combined error variance is given in table 4.

Figures 11 and 12 show the RMSE of the ensemble means and controls for the three reanalyses for summer and winter, respectively. The RMSE of the mean of UERRA-UB is a substantial decrease on that of CERA-20C, but there is no such decrease shown in UERRA-MO. Additionally the RMSE of the mean of UERRA-MO is slightly higher than that of its control, suggesting that the ensemble is not centred near the observed truth. Both UERRA-UB and CERA-20C have means with smaller RMSE than their controls.

Figures 11 and 12 also show the spread of the ensembles. The improvement by UERRA-UB over CERA-20C in correlation between spread and RMSE of the ensemble mean that was seen for temperature, is not seen for wind speed. When compared to results for CERA-20C, UERRA-UB

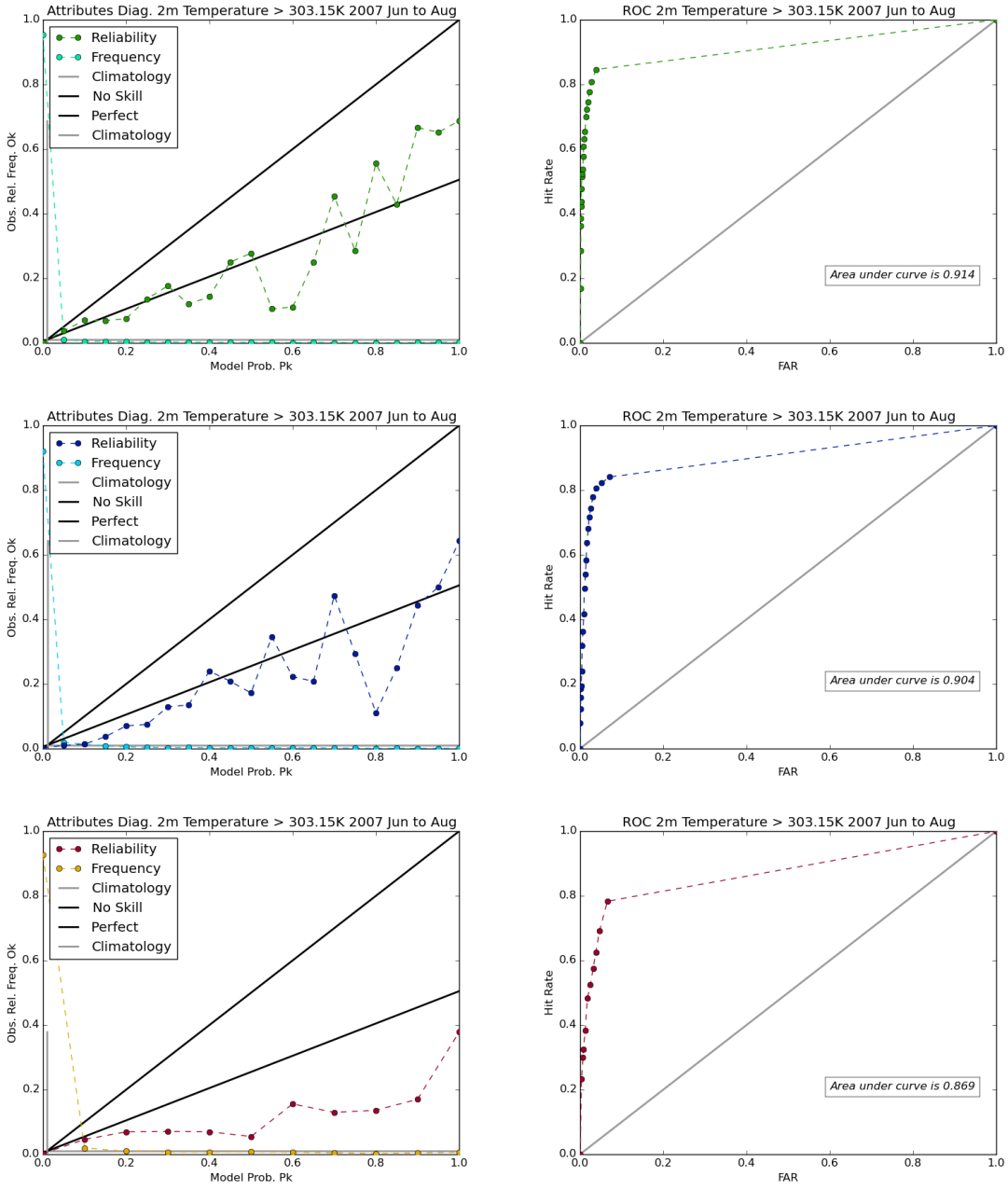

Figure 9: Attribute diagrams and receiver operating characteristic (ROC) curves in summer for daily mean temperature at 2m $TG > 30°C$. The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C. The left hand column shows reliability diagrams for the two systems. These contain plots of reliability, which here is the probability of observed events against model probability. Values for which the modelled probability has skill lie between the 'no skill' and 'perfect' lines. The frequency of each model probability is also displayed. The right hand column shows (ROC) curves for the two systems. This is a plot of hit rate against false alarm rate(FAR). The area under the curve would be one if the model is perfect.

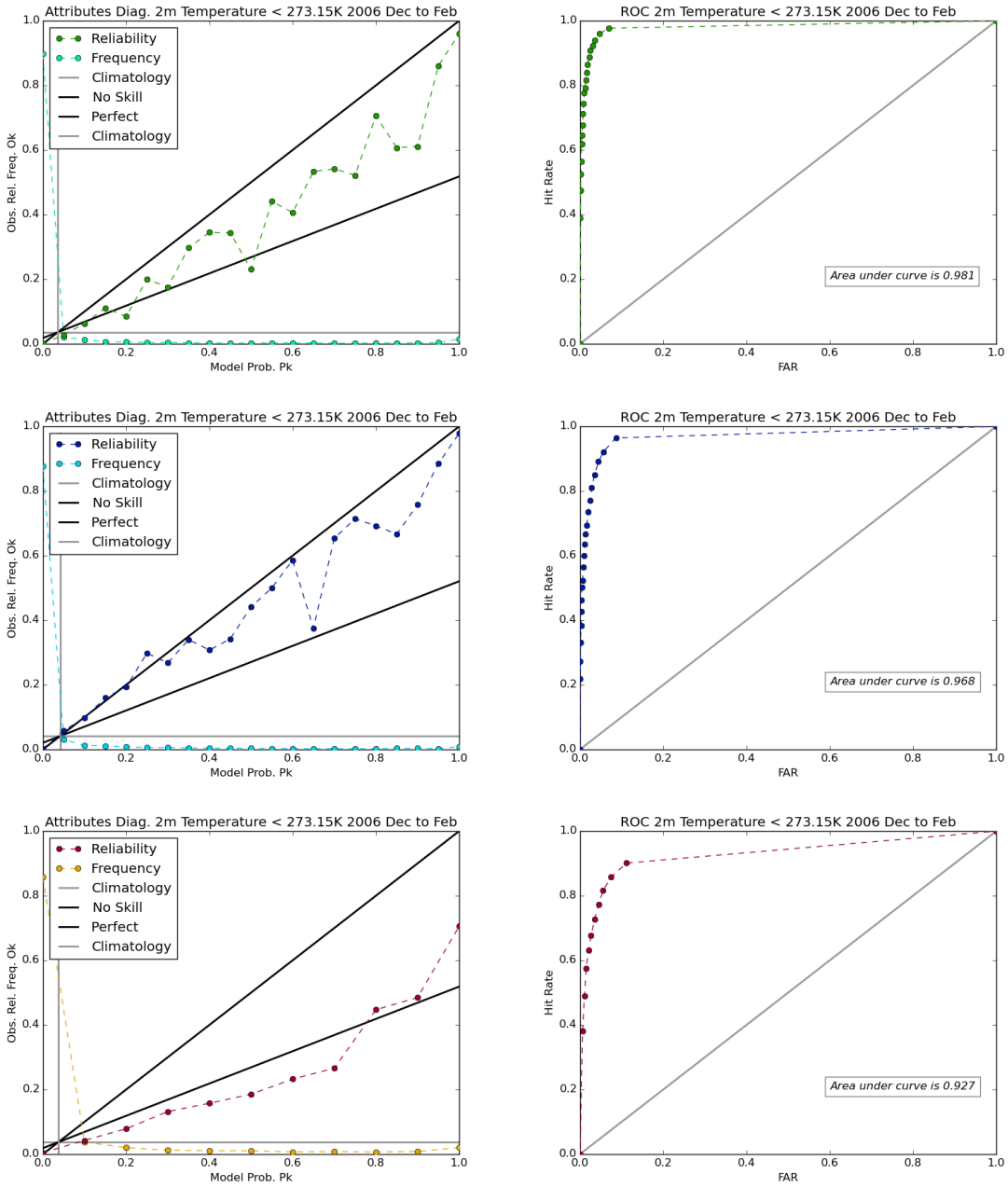

Figure 10: Attribute diagrams and receiver operating characteristic (ROC) curves in winter months (DJF) for daily mean temperature at 2m $< 0°C$. See figure 9.

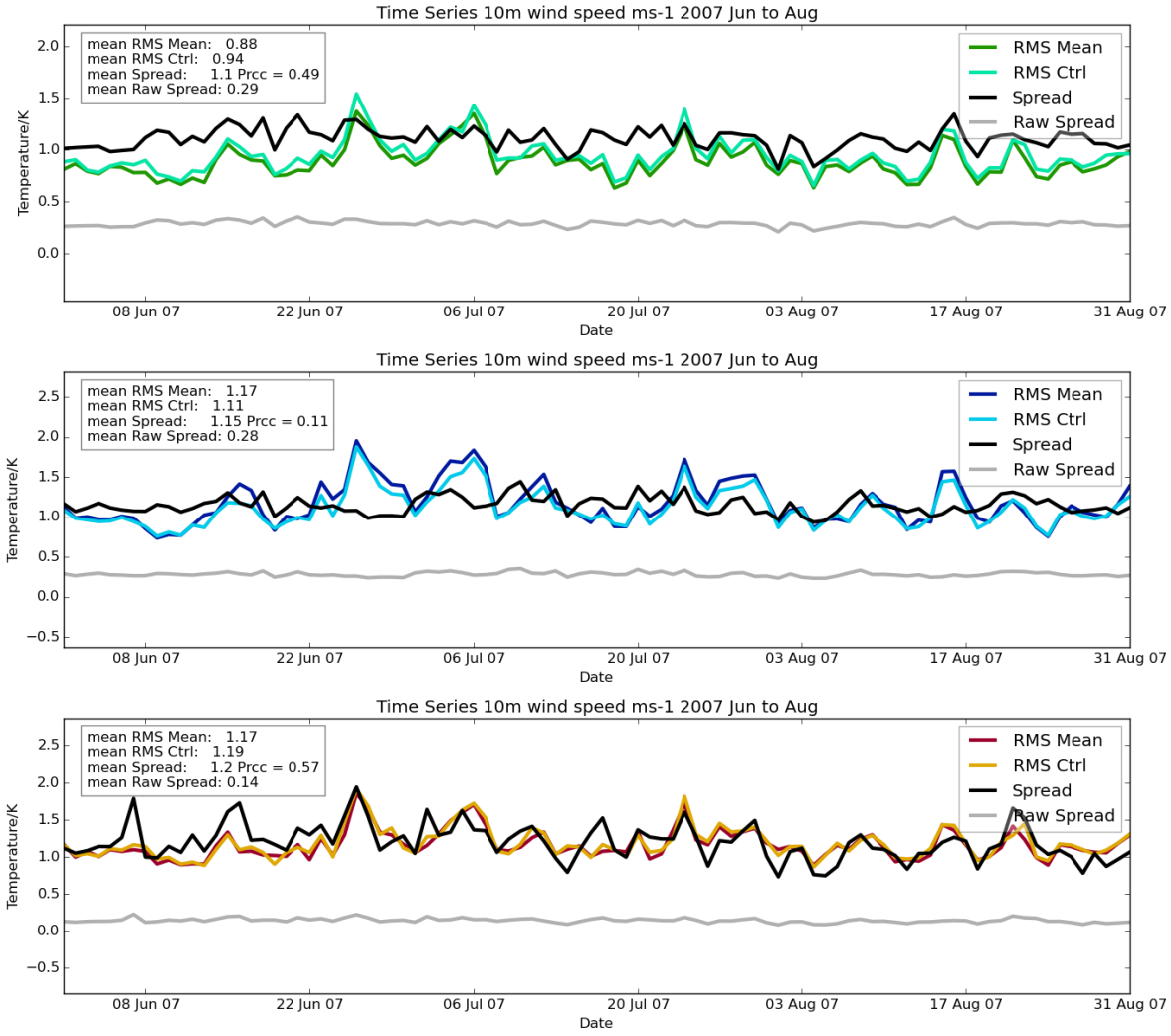

Figure 11: RMSE and spread in summer months for daily mean wind speed at 10m. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 1.

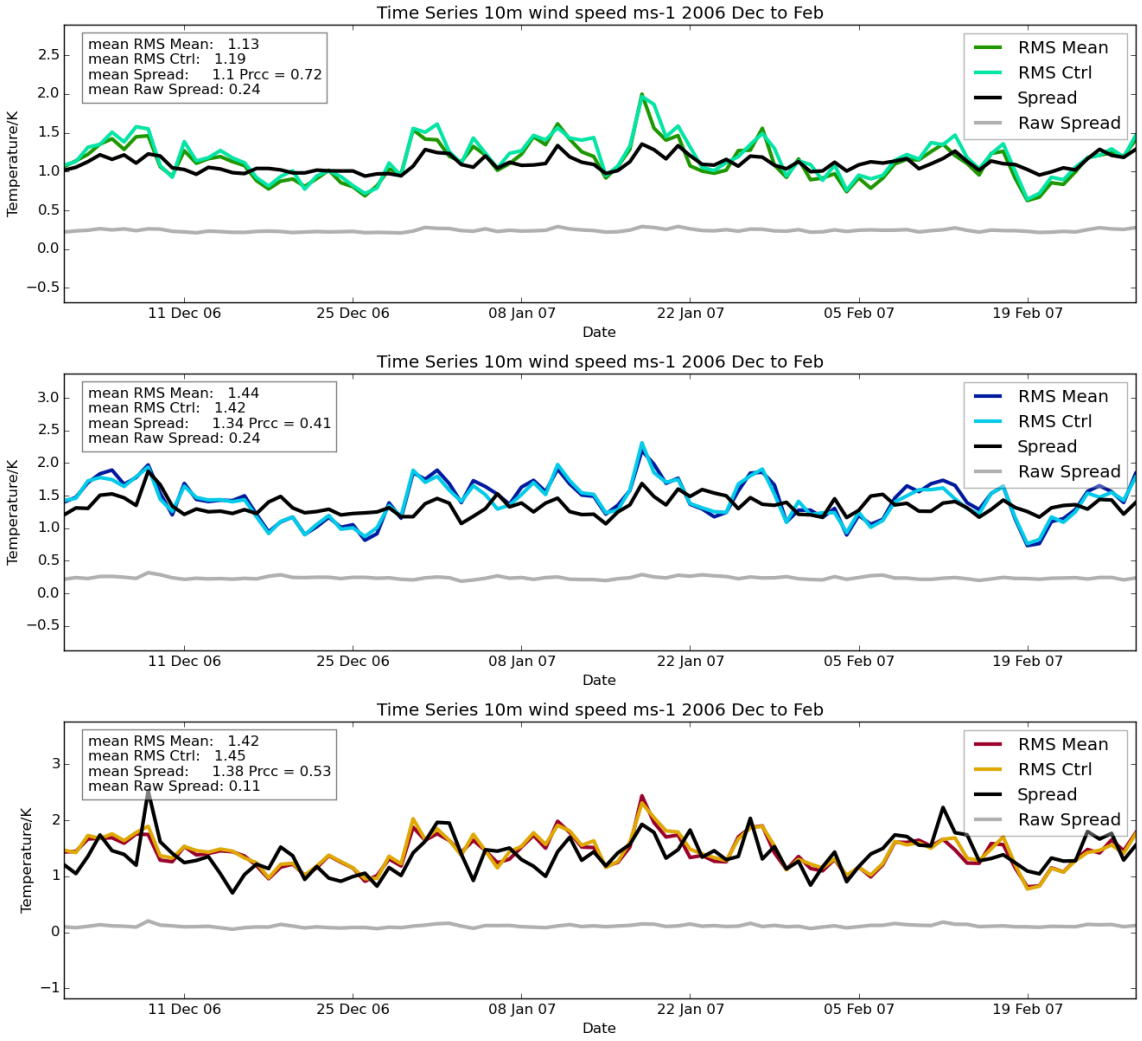

Figure 12: RMSE and spread in winter months for daily mean wind speed at 10m. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 1.

Summer	RMSE	Bias	ERR	IRR	CRPS	Prcc
UERRA-UB	0.77 (0.91)	-0.38 (-0.23)	1.00 (1.05)	1.43 (1.05)	0.06 (0.06)	0.37 (0.49)
UERRA-MO	1.19 (0.92)	-0.14 (-0.05)	1.81 (1.72)	0.77 (0.75)	0.08 (0.08)	-0.05 (0.11)
CERA-20C	1.29 (0.91)	-0.84 (-0.59)	1.50 (1.19)	1.41 (1.41)	0.17 (0.16)	0.35 (0.57)
Winter	RMSE	Bias	ERR	IRR	CRPS	Prcc
UERRA-UB	0.93 (092)	-0.59 (-0.33)	1.35 (1.57)	1.95 (1.01)	0.06 (0.07)	0.45 (0.72)
UERRA-MO	1.27 (0.87)	-0.23 (-0.08)	1.93 (1.82)	0.98 (0.77)	0.08 (0.09)	0.00 (0.41)
CERA-20C	1.49 (0.94)	-0.96 (-0.54)	1.55 (1.33)	1.67 (1.05)	0.17 (0.17)	0.15 (0.53)

Table 5: Table comparing daily mean 10m wind speed ensemble performance over the Netherlands. Results for the entire domain are given in brackets. RMSE and bias of ensemble mean are shown. ERR and IRR are the external and internal rank ratios, respectively, from the rank histograms. Prcc is the Pearson's rank correlation coefficient between ensemble spread and RMSE of the ensemble mean.

sees an improvement in winter, but a reduction in correlation in summer. UERRA-MO shows a reduction in both periods.

Similar results have been calculated for each reanalysis against observations in the Netherlands only. A summary of results comparing these with results for the entire domain are given in table 5. Over this sub-domain, the RMSE of the mean of UERRA-MO is lower than that of CERA-20C. The RMSE of the mean of UERRA-UB is lower than both. The ensemble spread correlation is worse for all three ensembles. The spread correlation of UERRA-UB is an improvement on both UERRA-MO and CERA-20C in both the full domain and the sub-domain.
Figures 13 and 14 show mean error for the controls and ensemble means for summer and winter, respectively. All three ensembles show a slow bias across both periods. The regional models show less bias than the global CERA-20C and these also have less bias in their ensemble means than in their controls. This reduction in bias is substantial for UERRA-MO, whose ensemble mean is the least biased of the three ensembles. All three ensembles see an increase in bias when moving from the full domain to the Netherlands sub-domain, see table 5.

Figure 15 shows rank histograms for the three ensembles across both periods. This suggests that, for all three systems, there is a greater than expected occurance of observations falling outside of the range of ensemble members. This is worst in UERRA-MO and least bad in CERA-20C. CERA-20C displays a bias such that most members tend to be slower than the observed values and UERRA-MO has a bias such that most members tend to be faster than the observed values. UERRA-UB has a negligible fast or slow bias of internal ranks, but appears over-spread (assuming the inflation factor is appropriate). The slow bias in the mean for each ensemble is explained by the fact that when observations fall outside of the range of ensemble members, they are most often faster than the ensemble. Similar results are seen over the Netherlands as over the entire domain, see table 5. The most significant difference between the sub-domain and the whole domain is that the internal rank ratio is larger for UERRA-UB in the sub-domain,

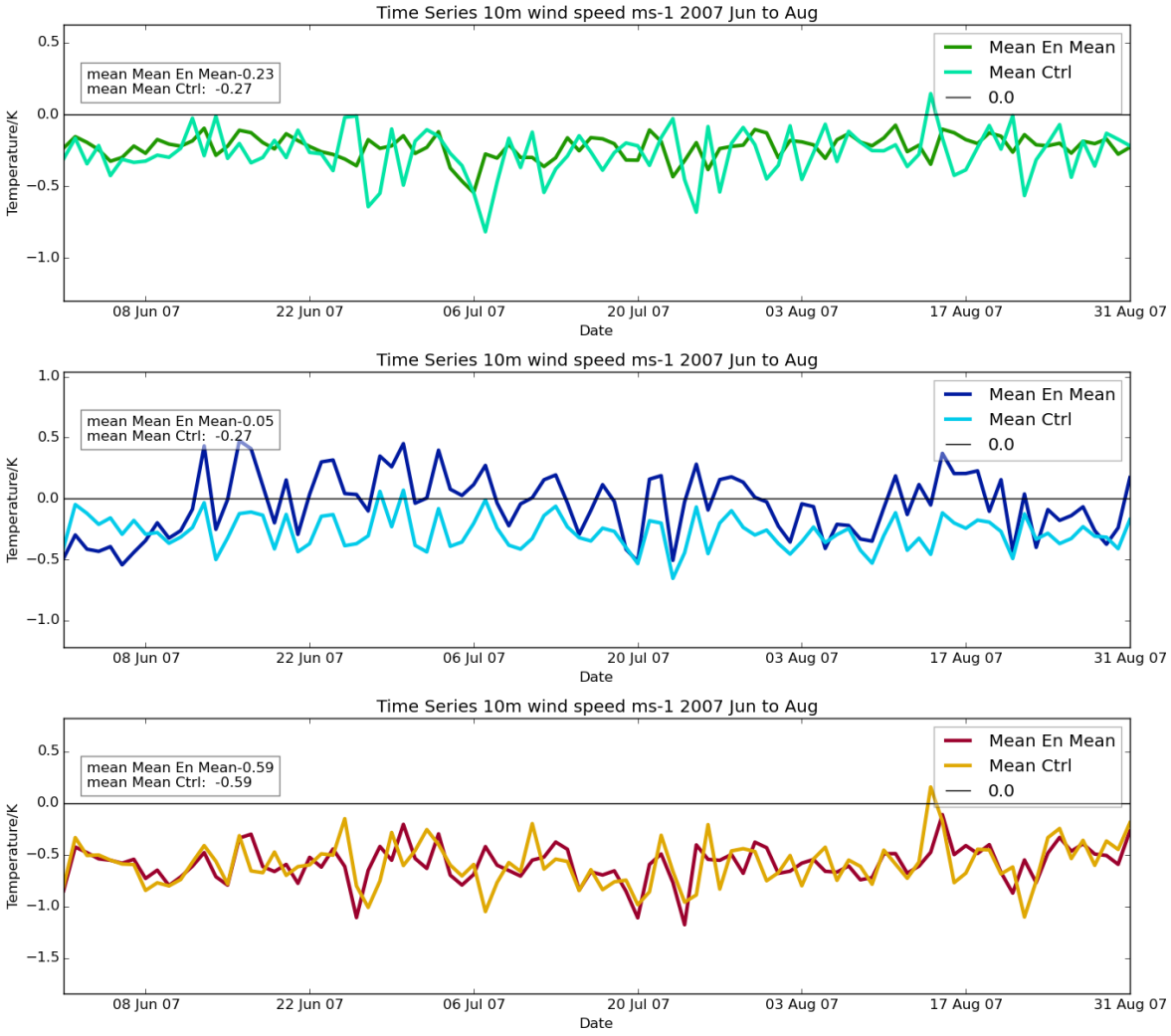

Figure 13: Bias in summer for daily mean wind speed at 10m. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 3.

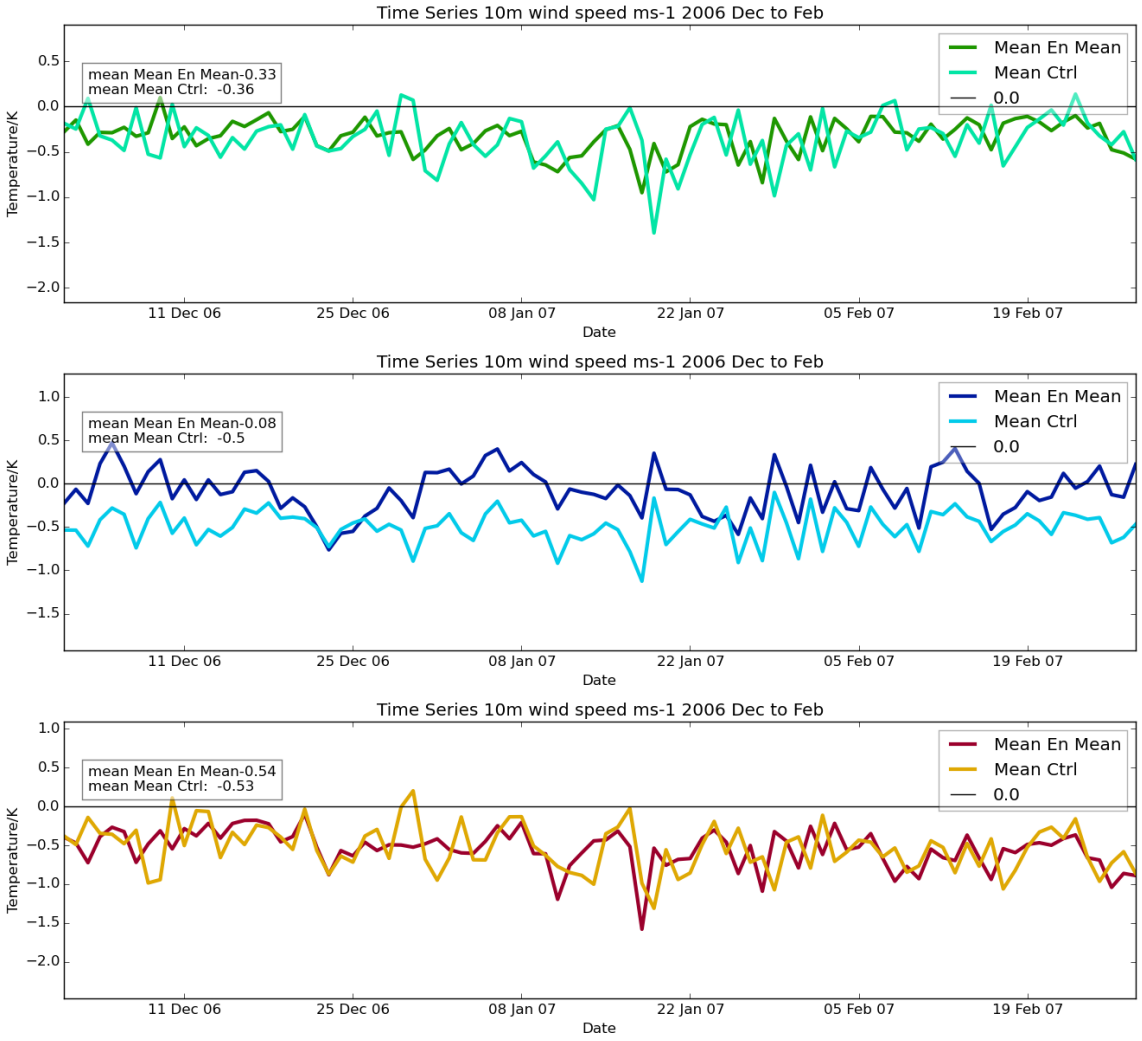

Figure 14: Bias in winter for daily mean wind speed at 10m. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 3.

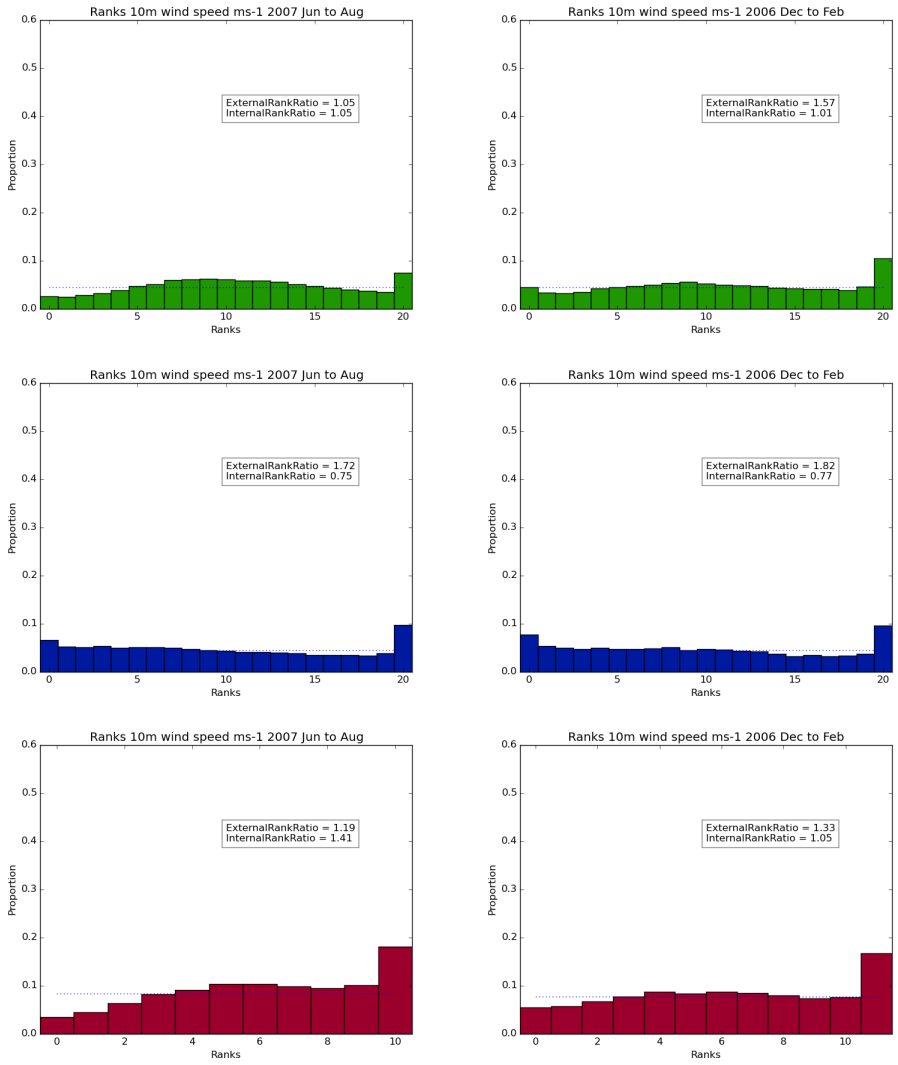

Figure 15: Rank histograms for daily mean wind speed at 10m. The plots show histograms of ranks of the observations with respect to the ensemble members. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

Reanalysis	Spread Inflation factor (JJA/DJF)	Error Variance (JJA/DJF)
UERRA-UB	1.5/3.1	3.7/3.6
UERRA-MO	1.2/2.1	3.4/3.8
CERA-20C	2.7/5.7	7.1/7.0

Table 6: Inflation factors for 24h precipitation.

indicating that the ensemble members tend to be faster than the observations in this region. This impact is also seen for CERA-20C in winter.

As with temperature, the CRPS decreases with increasing grid resolution, as shown for summer in figure 16 and winter in figure 17. This suggests that the wind speed probabilities of the regional systems are more accurate than the global system. CRPS scores over the Netherlands are very similar to those over the whole domain, see table 5.

Figure 18 shows that the Brier score for the ensembles also decreases with increasing grid resolution. The lowest (best) scores are achieved by UERRA-UB, then UERRA-MO, and then CERA-20C. These scores are consistently higher (worse) in summer. The regional reanalyses also have better reliability than CERA-20C, except for UERRA-MO in winter.

Figure 19 shows the attribute diagrams for events when $FG > 5.5ms^{-1}$ (at least a moderate breeze) for all three ensembles and for both periods. As can be see by the frequency lines, zero probability for each ensemble occurs in at least a third of cases. For UERRA-UB, 17/21 and 13/21 of the points, for summer and winter, respectively, lie between 'No Skill' and 'Perfect' lines, indicating that the ensemble is reliable for the majority of model probabilities (especially low and high probabilities). For UERRA-MO there are 7/21 summer points and 8/21 skillful points in winter. For CERA-20C all the points in summer are skillful and 5/11 are skillful in winter. This result suggests that CERA-20C has the best reliability in summer for this event.

Figure 20 shows the attribute diagrams for the same events as figure 19 ($FG > 5.5ms^{-1}$). These show that, as with temperature, accuracy increases with resolution, with UERRA-UB performing the best and CERA-20C the least well.

4 Daily Precipitation

For 24h precipitation, multiplicative inflation factors and the implied combined error variance are given in table 6.

Figures 21 and 22 show the RMSE of the control and ensemble mean for 24h precipitation (06Z-06Z) for the three ensembles and for summer and winter, respectively. UERRA-UB improves on CERA-20C, as might be expected with improved resolution. However UERRA-UB has a larger RMSE of the ensemble mean than that of the control, which usually indicates that the ensemble

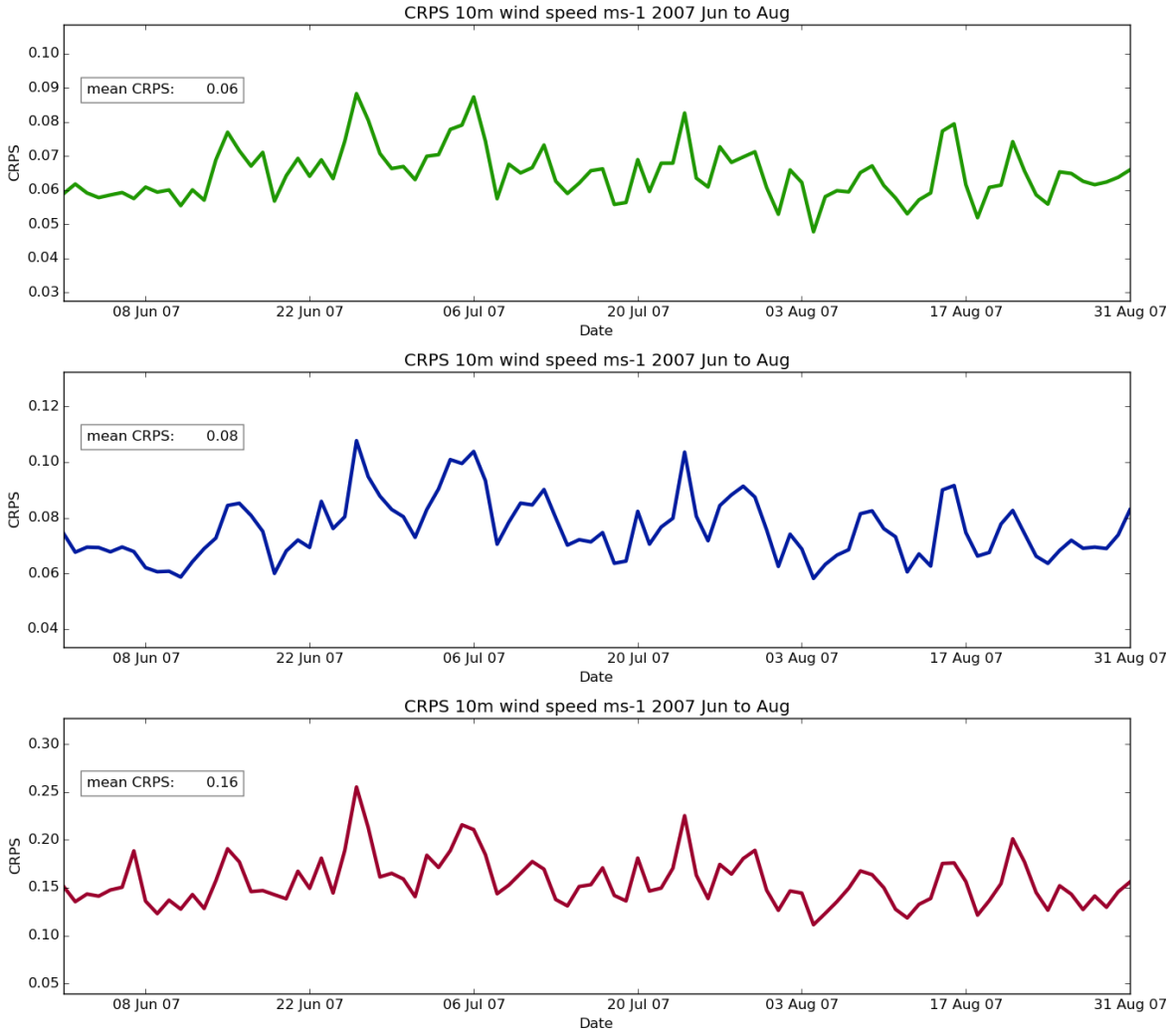

Figure 16: CRPS for daily mean wind speed at 10m in summer. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 6.

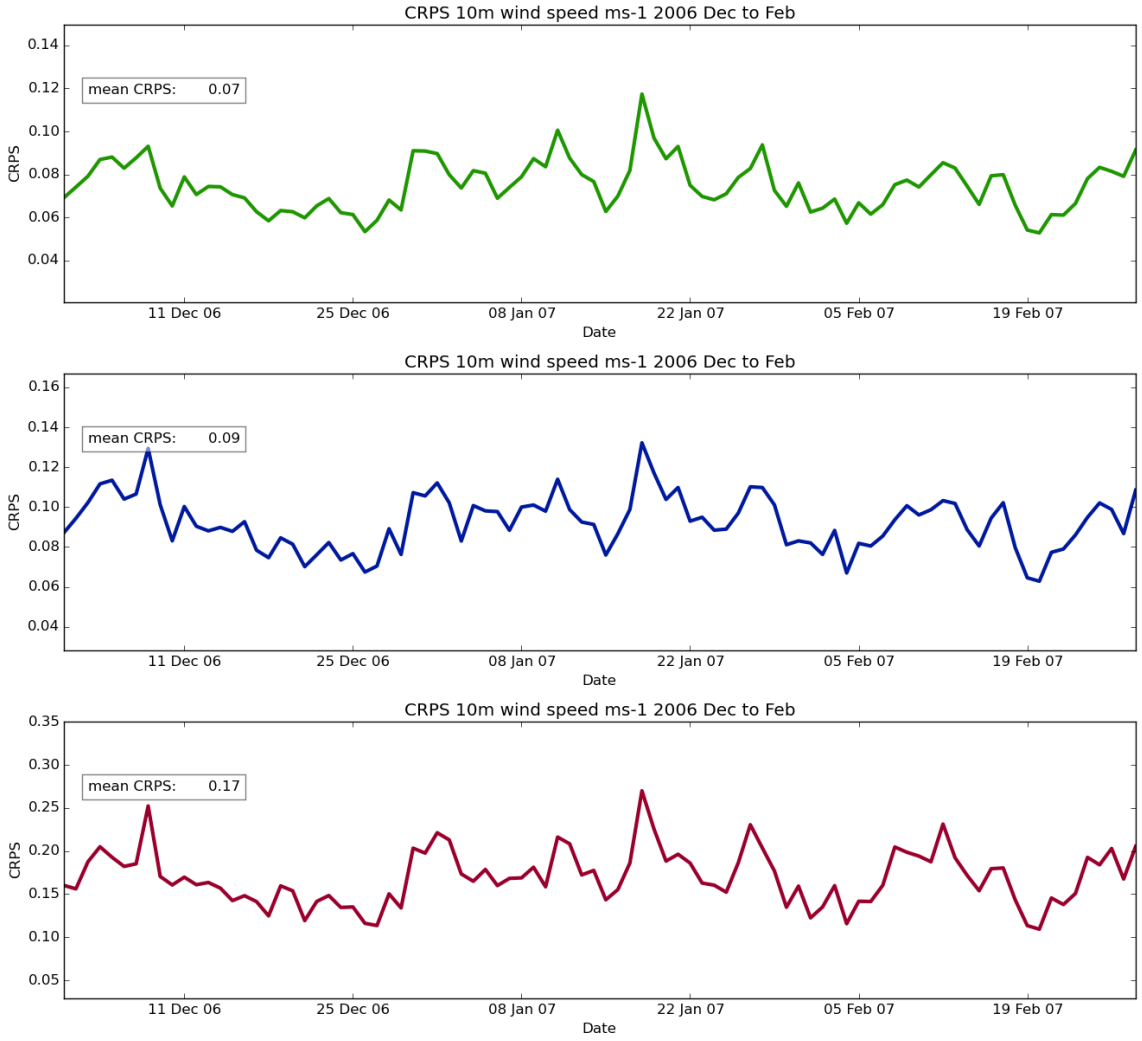

Figure 17: CRPS for daily mean wind speed at 10m in winter. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 6.

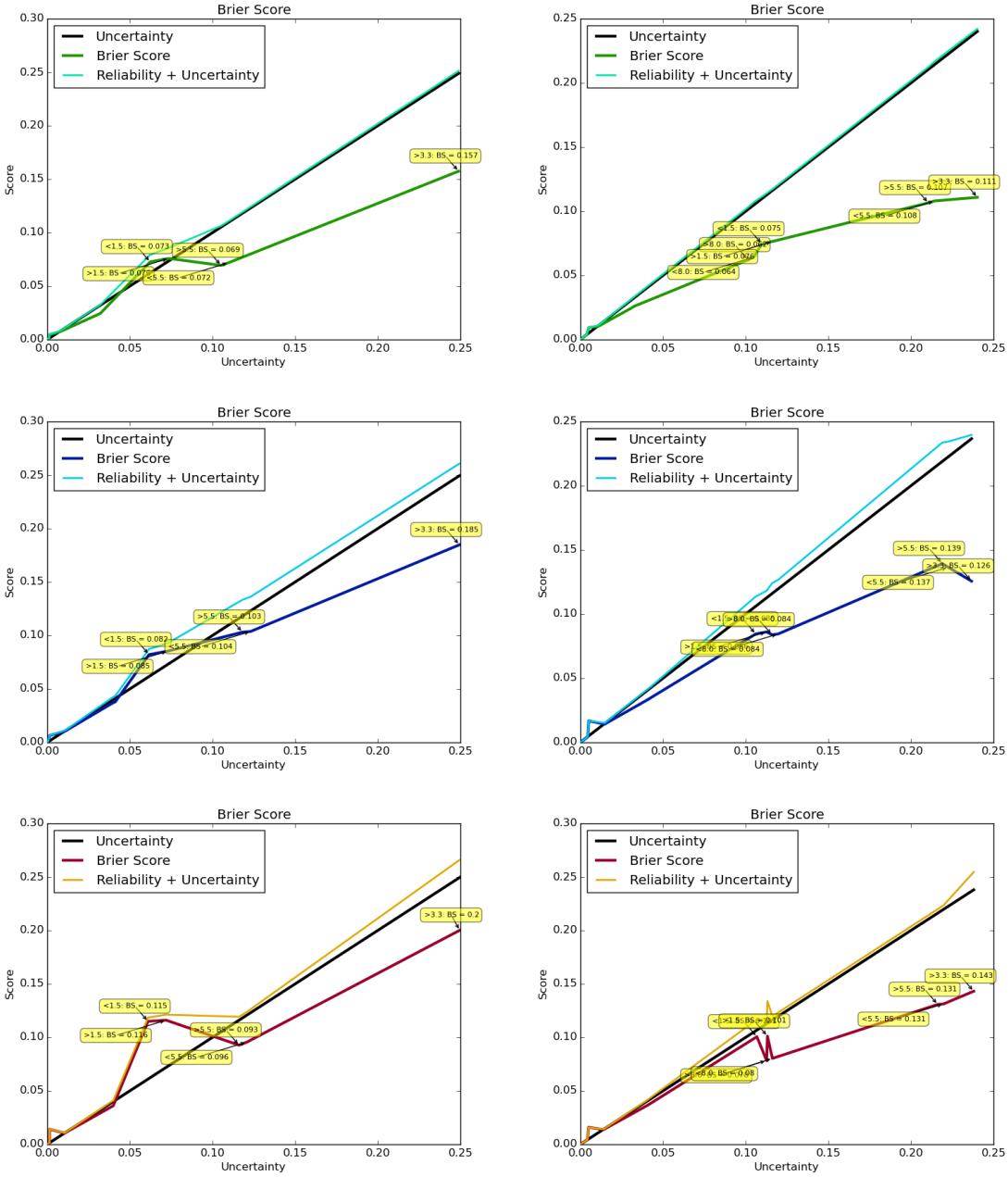

Figure 18: Brier scores and components for daily mean wind speed at 10m. See figure 8.

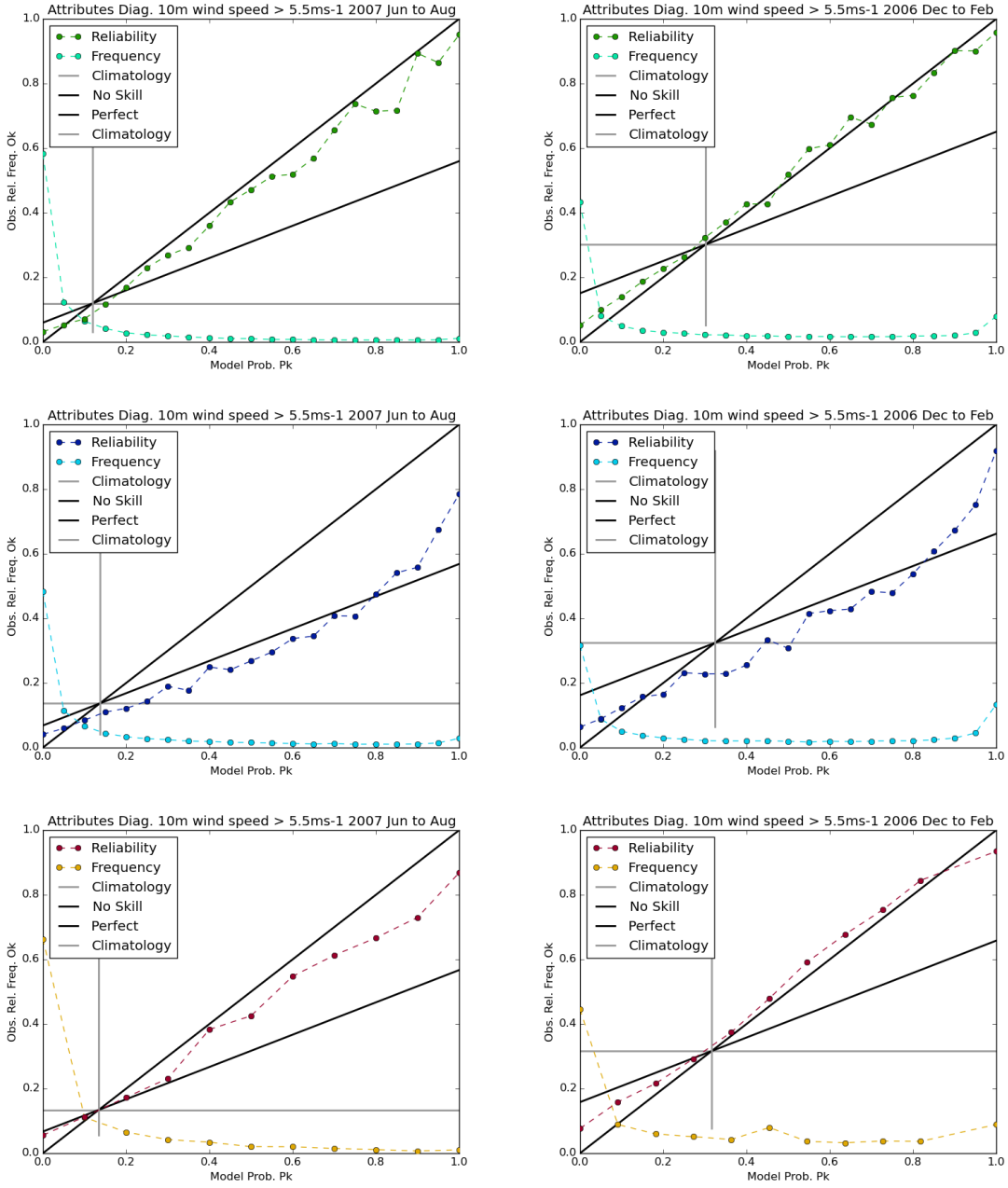

Figure 19: Daily mean wind speed at 10m in 2007 attributes diagrams, see figure 9. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

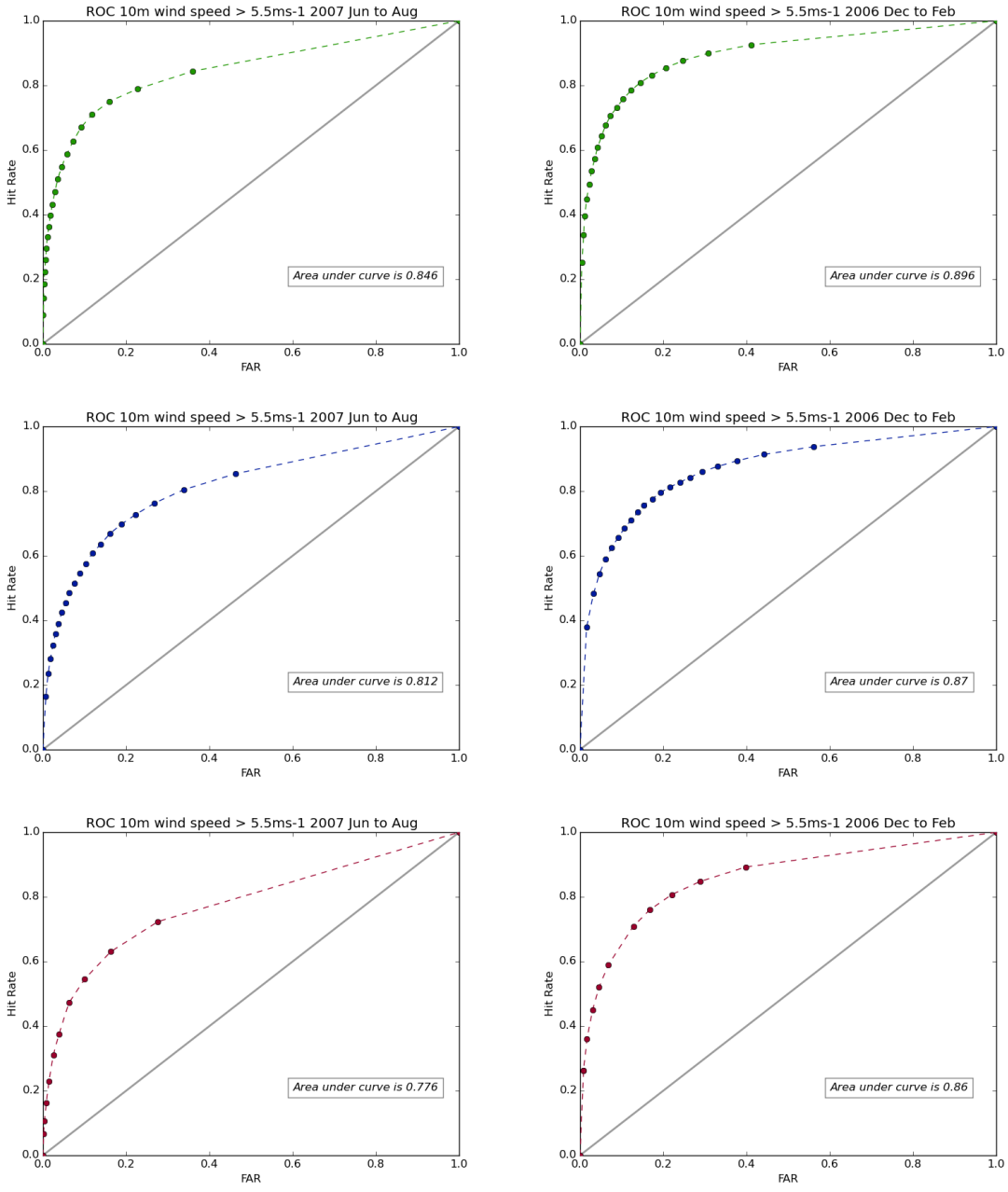

Figure 20: Daily mean wind speed at 10m in 2007 ROC curves, see figure 9. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

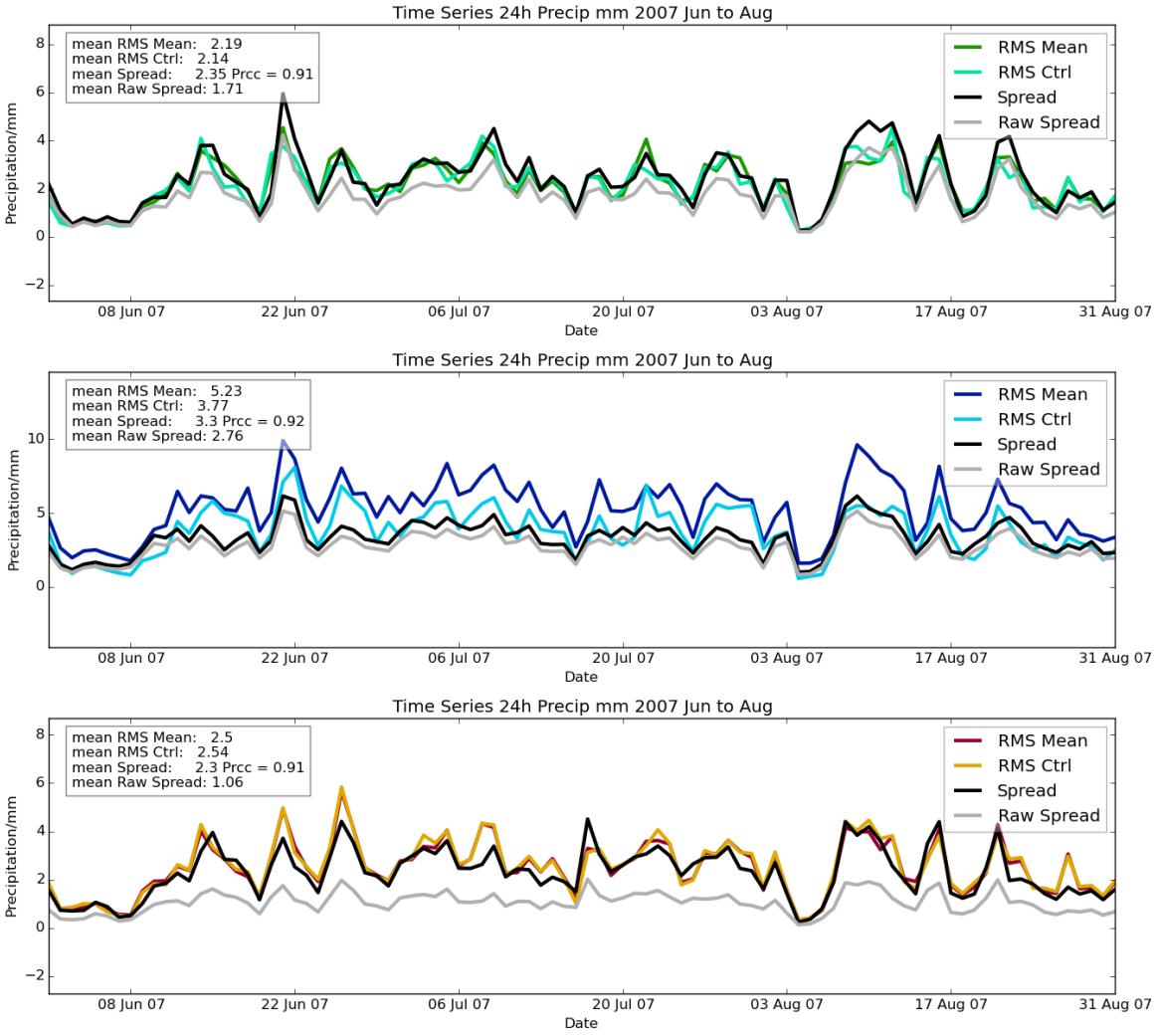

Figure 21: RMSE and spread in summer months for 24h precipitation. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 1.

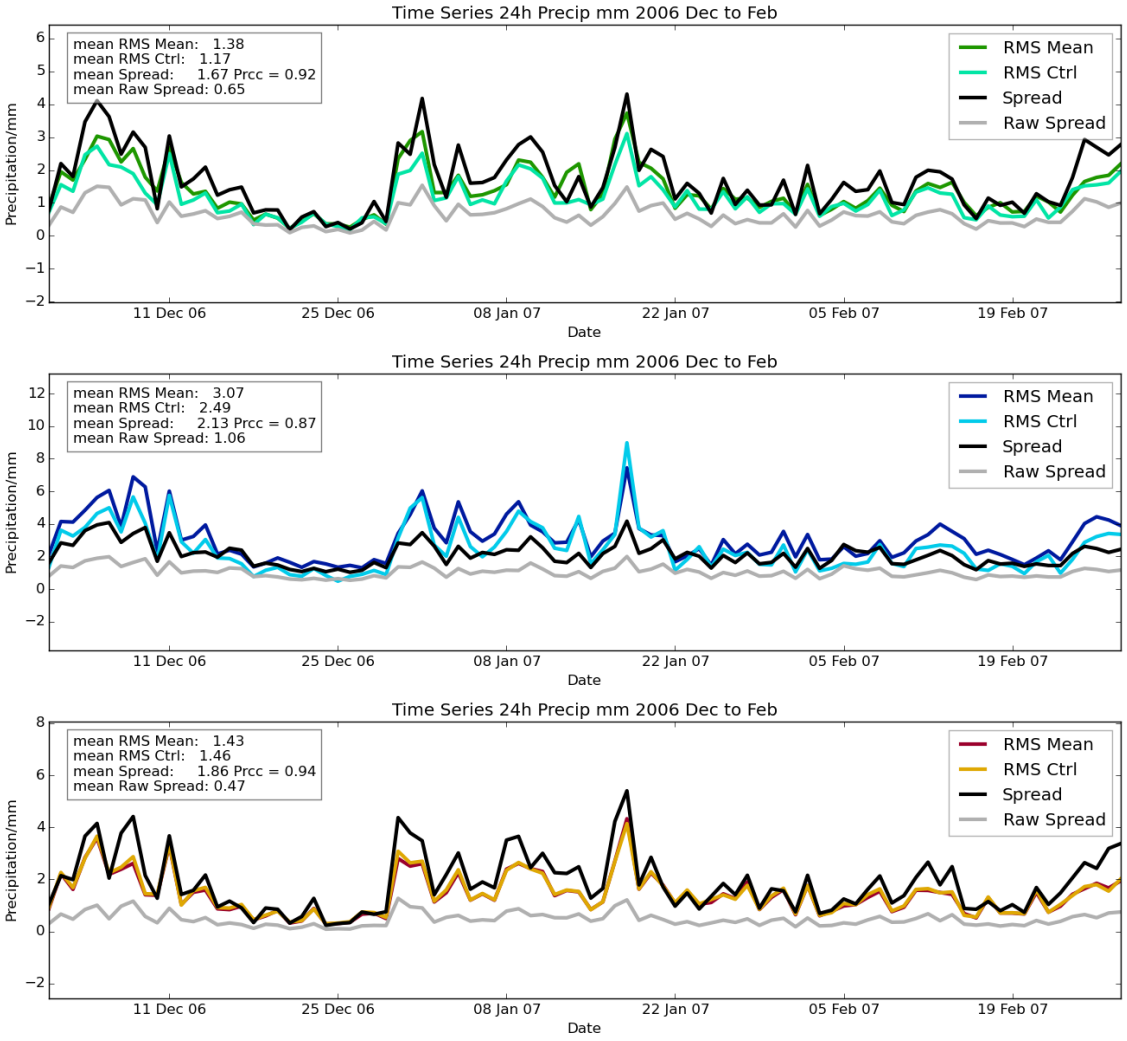

Figure 22: RMSE and spread in winter months for 24h precipitation. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 1.

Summer	RMSE	Bias	ERR	IRR	CRPS	Prcc
UERRA-UB	1.9 (2.19)	-0.20 (-0.20)	2.09 (1.68)	0.97 (0.89)	0.11 (0.14)	0.91 (0.91)
UERRA-MO	5.18 (5.23)	4.41 (4.27)	5.03 (5.69)	0.24 (0.26)	0.27 (0.27)	0.89 (0.92)
CERA-20C	2.15 (2.50)	0.15 (0.04)	1.52 (1.33)	0.55 (0.53)	0.25 (0.29)	0.92 (0.91)
Winter	RMSE	Bias	ERR	IRR	CRPS	Prcc
UERRA-UB	1.62 (1.38)	0.05 (0.17)	2.16 (1.53)	0.86 (0.72)	0.10 (0.09)	0.93 (0.92)
UERRA-MO	3.17 (3.07)	2.47 (2.51)	3.50 (3.62)	0.22 (0.18)	0.17 (0.17)	0.86 (0.87)
CERA-20C	1.70 (1.43)	0.37 (-0.11)	1.51 (0.96)	0.52 (0.75)	0.20 (0.20)	0.93 (0.94)

Table 7: Table comparing 24h precipitation ensemble performance over France. Results for the entire domain are given in brackets. RMSE and bias of ensemble mean are shown. ERR and IRR are the external and internal rank ratios, respectively, from the rank histograms. Prcc is the Pearson's rank correlation coefficient between ensemble spread and RMSE of the ensemble mean.

is not centred on the observed truth. Using grid-point metrics for precipitation can penalise 'better' models, [Jermey and Renshaw, 2016], which may account for this increase. These figures also show that the RMSE of UERRA-MO is large. The 3DVAR assimilation system, used for the ensemble, is known to suffer from spin-up for precipitation fields within the first six hours of the forecast. This results in an unrealistically wet ensemble. Use of 4DVAR or forecasting for an additional six hours would remove this issue, but is prohibitively expensive for this project. The deterministic reanalysis uses hybrid 4DVAR and is uneffected. Results over the smaller domain of France are summarised in table 7. This shows that the three reanalyses display similar quality over France as over the entire domain.

Figures 21 and 22 also show the spread of the ensembles for the same periods. For all three ensembles, the spread matches very well with the RMSE of the ensemble mean. In summer the correlation between the spread and RMSE of the ensemble mean is similar in all three ensembles and in winter the global reanalysis performs best. This shows that, even with spuriously large values of precipitation, UERRA-MO is useful for estimating precipitation uncertainty. Again, similar results are seen over France as over the entire domain, see table 7.

Figures 23 and 24 show the mean error of the control and ensemble mean for 24h precipitation (06Z-06Z) for the three reanalyses and for summer and winter, respectively. The increased precipitation due to spin-up in UERRA-MO is clearly shown in these figures. Both of the other ensembles have a smaller bias in the mean than the control in summer and a larger one in winter. CERA-20C has a smaller bias in both periods than UERRA-UB. When looking at results for the sub-domain over France, bias is increased in CERA-20C, indicating that UERRA-UB has a smaller bias than CERA-20C in summer, see table 7.

The rank histograms for all three ensembles and for both periods are shown in figure 25. Again, the increased precipitation of UERRA-MO is clear, with most ensemble members wetter than

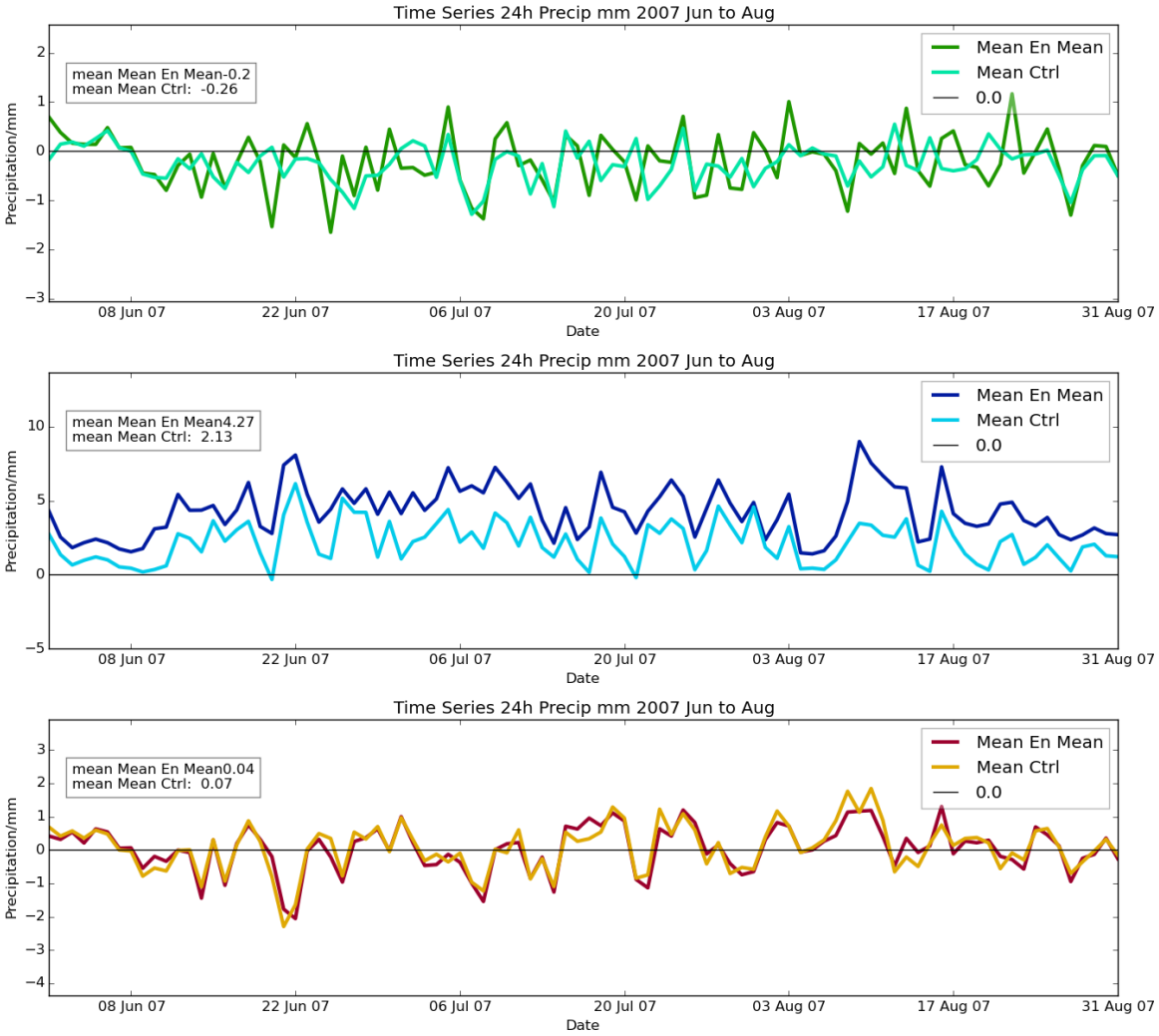

Figure 23: Bias in summer for 24h precipitation. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 3.

Figure 24: Bias in winter for 24h precipitation. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 3.

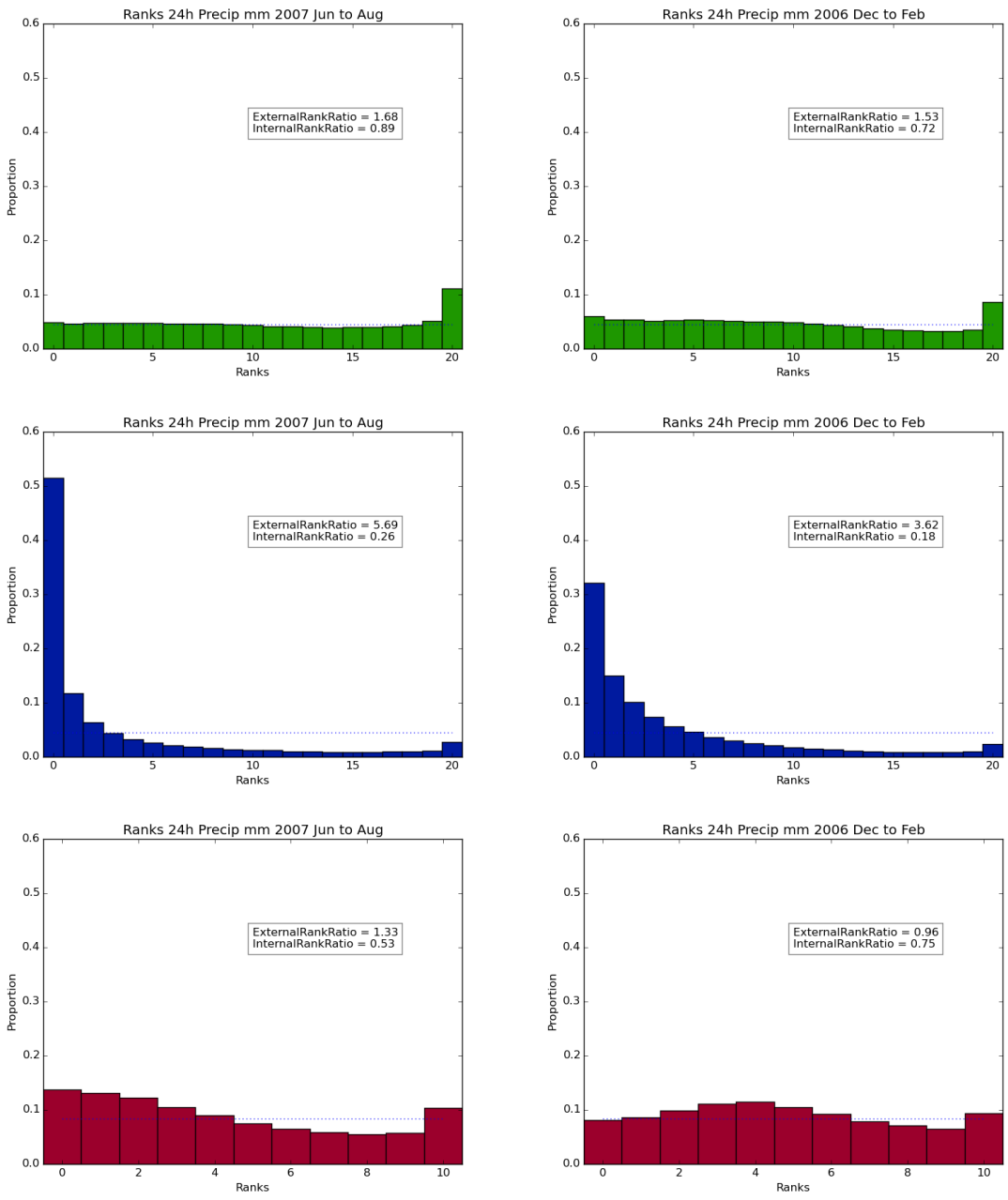

Figure 25: Rank histograms for daily precipitation. The plots show histograms of ranks of the observations with respect to the ensemble members. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

the observations and a large proportion of observations falling below the range of the ensemble. The rank histograms for CERA-20C also suggest that ensemble members tend to be wetter than the observations, but observations that fall outside of the range of the ensemble are more evenly balanced. The rank bias is greatly reduced in UERRA-UB, but the ensemble still sees a disproportionate number of observations lying outside its range. Similar results are seen in the sub-domain of France, see table 7.

The CRPS for the three ensembles is shown in figures 26 and 27 for summer and winter, respectively. The results show that the ability of UERRA-MO to capture the cumulative density function of the observations is slightly better than that of CERA-20C, even with spuriously high precipitation values. UERRA-UB is an improvement on both of these. All three ensembles performs better in winter than in summer. Again, similar results are seen in the sub-domain of France, see table 7.

The Brier score and its components for the three ensembles is shown in figure 28 for both periods. UERRA-MO has a large Brier score and a large reliability score, indicating that it does not represent precipitation probabilities accurately and does not produce probabilities that are consistent with frequencies of observed events. Brier scores and reliability scores are better in UERRA-UB compared to CERA-20C in summer and slightly better in winter.

The attributes diagrams for 24h precipitation of at least 32mm is shown in figure 29 for both periods. Due to the increased precipitation, the attribute diagrams of UERRA-MO show few points between 'No skill' and 'Perfect', indicating that the ensemble has little skill in reliability of this event. CERA-20C has 4/11 and 3/11 skillful points for summer and winter, respectively. UERRA-UB improves on this with 13/21 and 12/21 skillful points in summer and winter, respectively.

The ROCs for 24h precipitation of at least 32mm is shown in figure 30 for both periods. These show that the hit rate of such events is higher in UERRA-MO than CERA-20C, but that this also has a higher false alarm rate (FAR). Large values of precipitation are not well represented in gridded models due both to the restrictions of the grid and to parameterisation of sub-grid processes, which tend to average out precipitation across a number of time steps [Roberts and Lean, 2008]. Although this metric appears to favour UERRA-MO, this is because of the general increased precipitation in the ensemble, rather than a desirable improvement. As expected UERRA-UB is a substantial improvement on CERA-20C.

To ameliorate precipitation spin-up problems from the ensemble of UERRA-MO, it is possible to ignore the first one or two hours of the accumulation period from each six-hour cycle, when excessive spin-up is worst. Instead of the 24h period comprising four six hour accumulations, it instead comprises four four or five hour accumulations as a proxy for the six hours. Results assessing the ensemble using this adjustment are summarised in table 8.

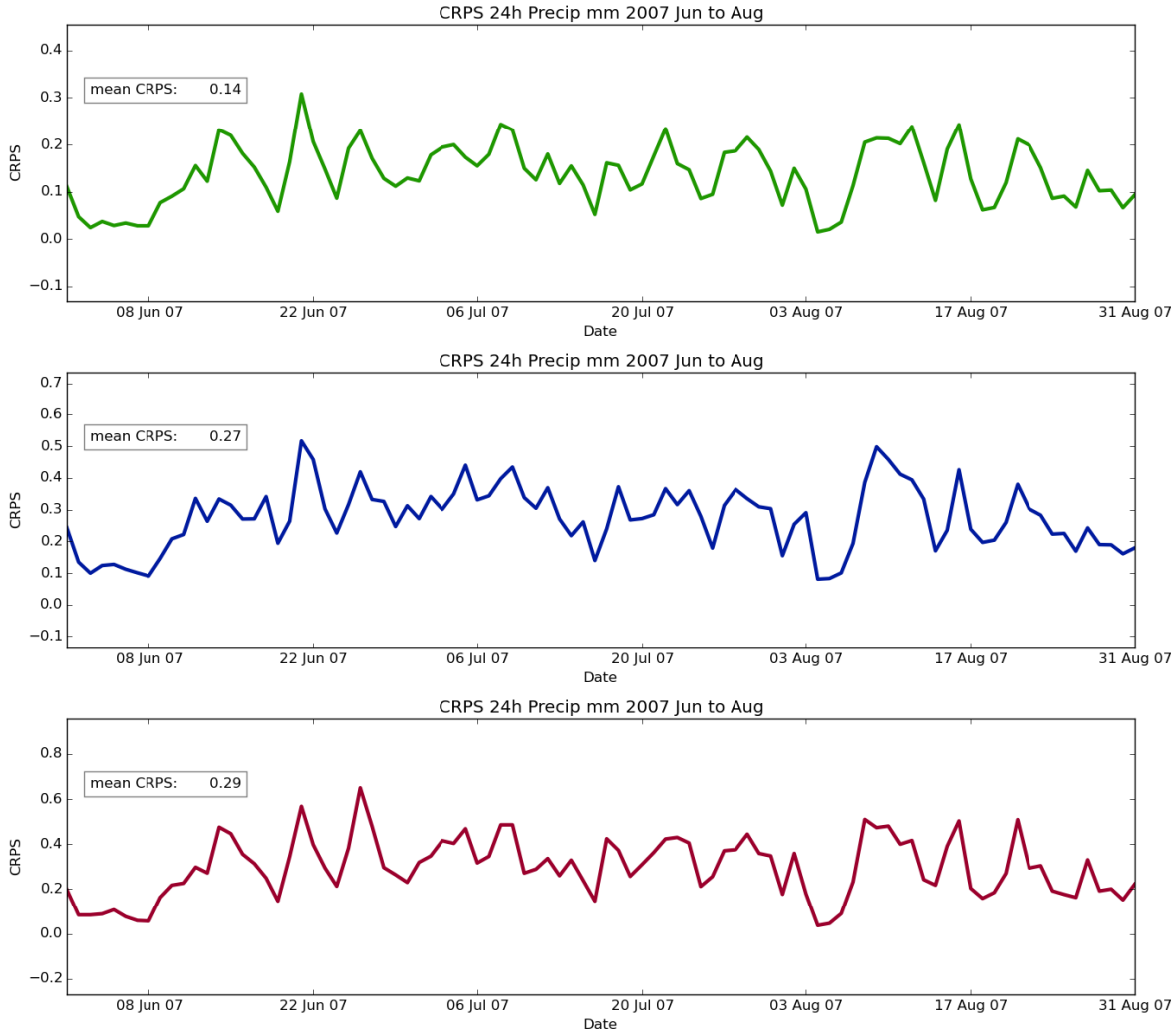

Figure 26: CRPS for 24h precipitation in summer. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 6.

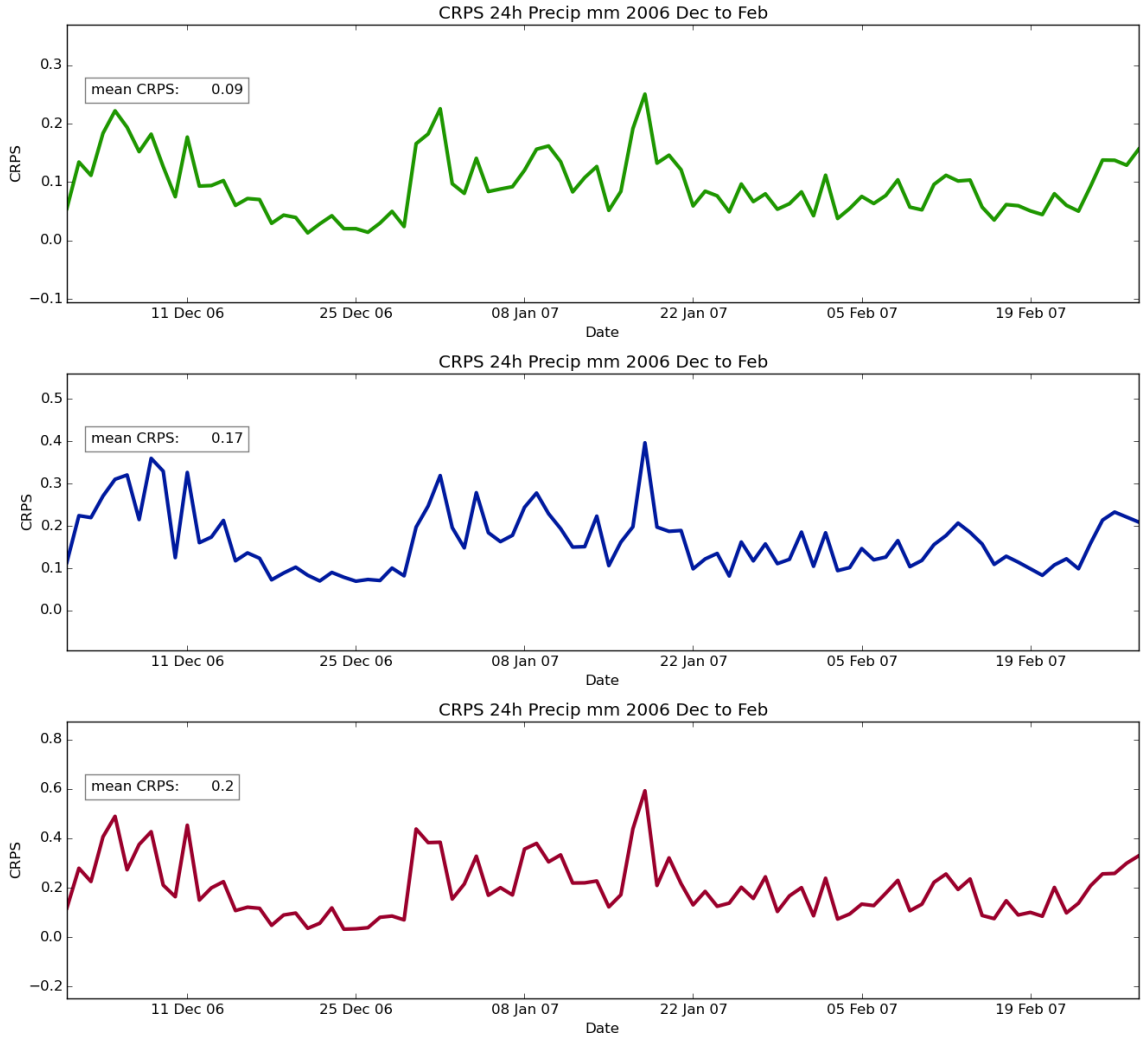

Figure 27: CRPS for 24h precipitation in winter. Top to Bottom: UERRA-UB, UERRA-MO and CERA-20C. See figure 6.

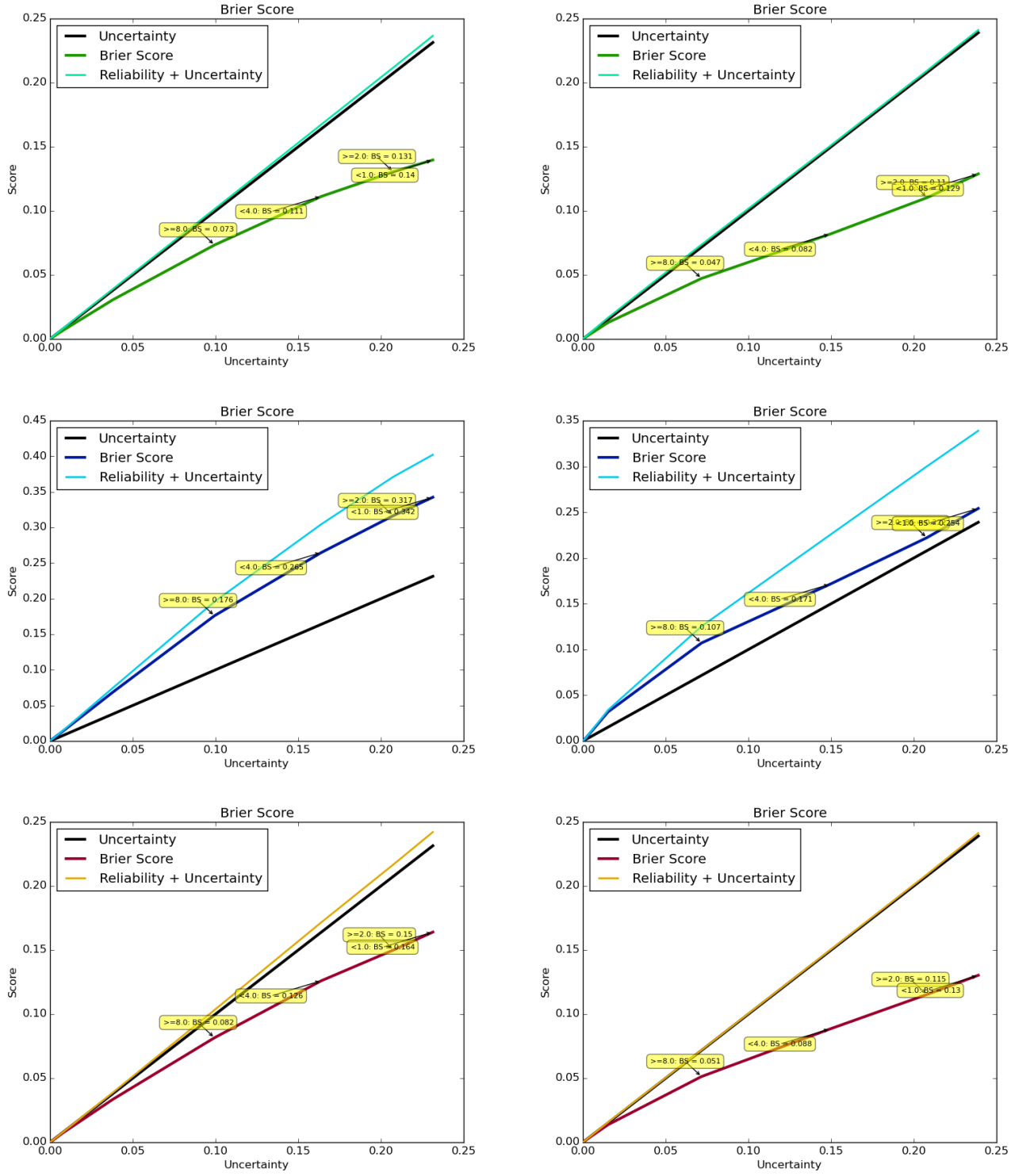

Figure 28: Brier scores and components for 24h precipitation. See figure 8.

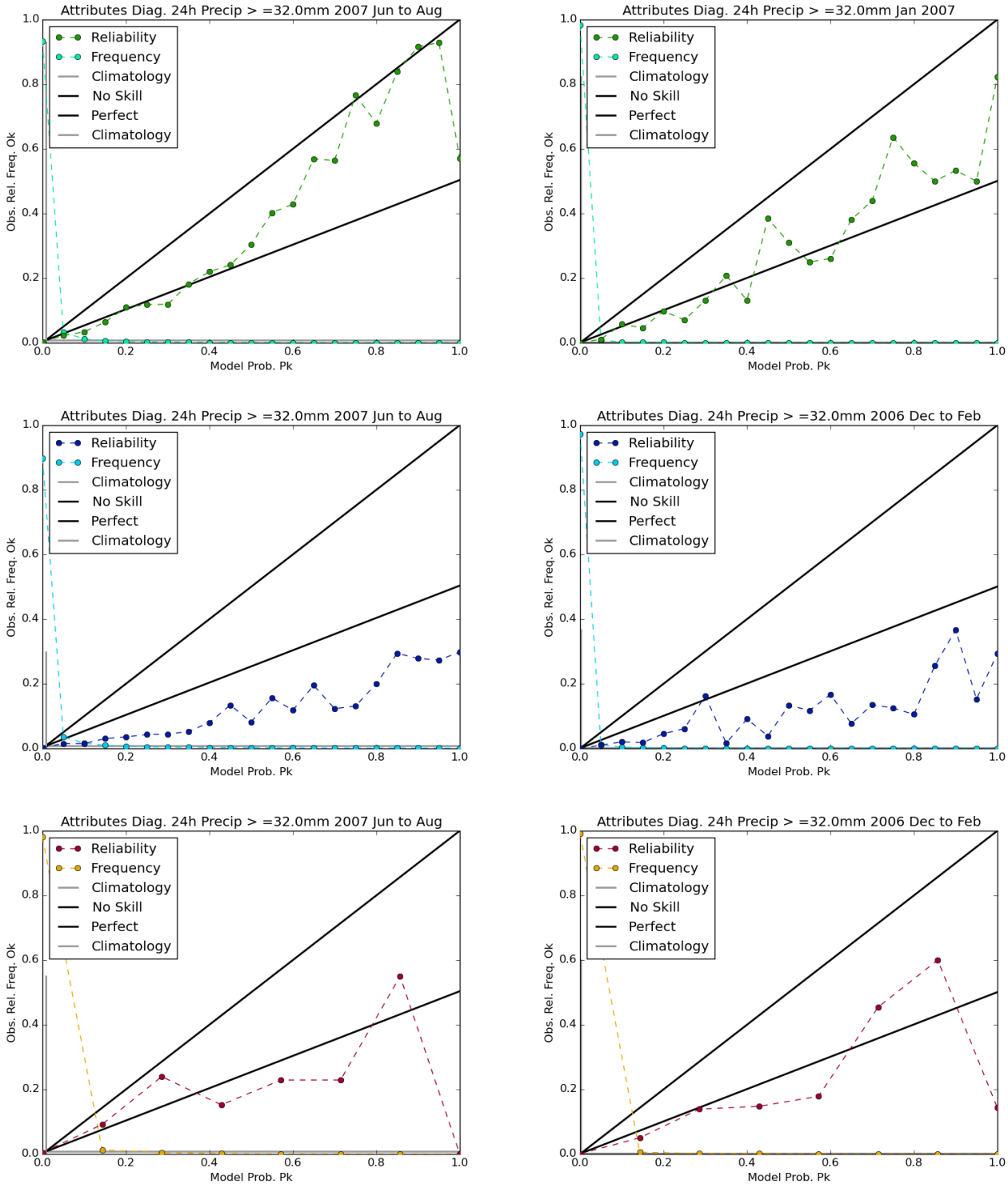

Figure 29: 24h precipitation in 2007 attributes diagrams, see figure 9. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

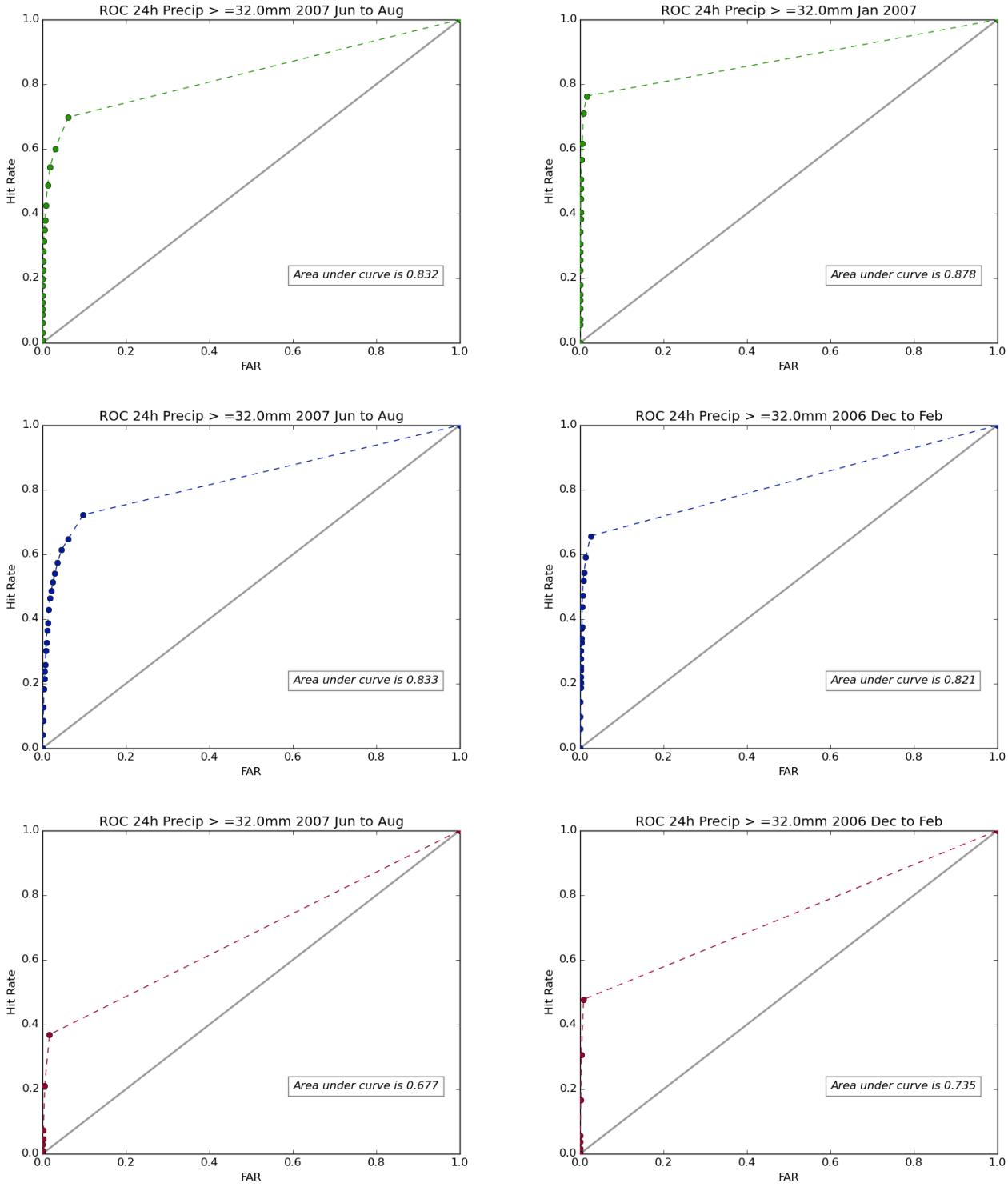

Figure 30: 24h precipitation 2007 ROC curves, see figure 9. The left hand column is for summer months (JJA) and the right hand column is for winter months (DJF). The top row shows results for UERRA-UB, the middle shows results for UERRA-MO and the bottom row shows results for CERA-20C.

Summer	RMSE	Bias	ERR	IRR	CRPS	Prcc
MO (6h)	5.23	4.27	5.69	0.26	0.27	0.92
MO (5h)	4.01	2.70	4.50	0.28	0.22	0.93
MO (4h)	3.08	1.23	3.04	0.31	0.17	0.92
DWD	2.19	-0.20	1.68	0.89	0.14	0.91
CERA-20C	2.50	0.04	1.33	0.53	0.29	0.91
Winter	RMSE	Bias	ERR	IRR	CRPS	Prcc
MO (6h)	3.07	2.51	3.62	0.18	0.17	0.87
MO (5h)	2.22	1.40	2.34	0.24	0.13	0.91
MO (4h)	1.67	0.39	1.81	0.38	0.10	0.92
DWD	1.38	0.17	1.53	0.72	0.09	0.92
CERA-20C	1.43	-0.11	0.96	0.75	0.20	0.94

Table 8: Table comparing 24h precipitation ensemble performance removing first one and two hours of accumulation period (each four hour cycle) for UERRA-MO. RMSE and bias of ensemble mean are shown. ERR and IRR are the external and internal rank ratios, respectively, from the rank histograms. Prcc is the Pearson's rank correlation coefficient between ensemble spread and RMSE of the ensemble mean. Results for

Table 8 suggests that this adjustment will greatly improve the accuracy and bias of the ensemble mean and members. The adjustment also substantially improves the accuracy of the ensemble, measured by CRPS, and improves (winter) or maintains (summer) the high correlation between the ensemble spread and the ensemble RMSE.

5 Conclusions

The results show that for all three variables (daily mean temperature, daily mean wind speed and daily total precipitation), the accuracy of the mean of UERRA-UB is an improvement over that of CERA-20C and that of UERRA-MO. The results also show that the mean of UERRA-MO is an improvement on CERA-20C for daily mean temperature and of similar quality for daily wind speed. Comparing the regional reanalyses with the global reanalysis, the bias in the mean of the ensemble is improved, except with UERRA-MO for summer temperature and with both regional reanalyses for daily total precipitation. For daily total precipitation, CERA-20C is the least biased of the three ensemble means. Other activities within the UERRA project have found different results with respect to bias, for example UERRA-UB has been found to be too warm in summer and too cold in winter over Germany, [Lockhoff, 2017]. These differing results are likely due to different observations being used as a proxy for truth and due to different treatment of the observations (interpolation method, height adjustment, etc). Users should be advised that any assessment of reanalysis quality is dependent on the accuracy and uncertainty of the 'truth'.

Similar results are shown for the accuracy of the ensemble. CRPS, Brier and ROC results all suggest that UERRA-UB has improved ensemble accuracy over both CERA-20C and UERRA-MO over all three variables, while UERRA-MO is an improvement in ensemble accuracy over CERA-20C, even in precipitation. The reliability component of the Brier score also shows an improvement for UERRA-UB over both CERA-20C and UERRA-MO for all three variables and an improvement for UERRA-MO over CERA-20C for daily mean temperature. Results from the attribute diagrams are more mixed, indicating that overall improvements do not necessarily reflect in scores representing single event categories.

Even with inflation to account for representivity and observation error, all three ensembles suffer from the observation appearing too frequently outside of the ensemble range. The rank histograms also indicate that when the observation is within the range of the ensemble, the UERRA ensembles tend to rank the observation more evenly than CERA-20C. This indicates that the truth is closer to being a member of these ensembles than of that of CERA-20C. The exception to this are the rank histograms for precipitation for which UERRA-MO behaves poorly.

The focus for the UERRA ensembles is to accurately represent the uncertainty via the ensemble spread. For daily mean temperature both regional reanalyses feature spread which correlates better to their mean RMSE than that of CERA-20C, although correlations are generally low for all three reanalyses. In summer the regional reanalyses show worse spread correlations for daily mean wind speed than CERA-20C, but in winter the situation is reversed. For daily precipitation results all three ensembles represent the uncertainty well. This indicates that even with increased values of precipitation due to spin-up, the precipitation spread is still useful from UERRA-MO.

These results show the ensembles of regional reanalyses performing well compared to the best quality global alternative, at the time of writing. The ensembles not only show improvements in representing uncertainty, the focus of the UERRA project, but also in ensemble accuracy.

Variable	Source	Description
TG	TG5	Mean calculated as average of daily minimum & maximum
FG	FG1	Av. of 24h measurements of 10-min average (0-0 UT)
	FG6	Av., mean of 4 10-min averages of 00, 07, 13 and 18 UT
	FG7	Av. of 24 hourly measurements (6 av. per hour) 0-0 UT
	FG8	Av. 10-minute from 23UT prev. day - 22UT today (24 values)
	FG9	Av. from 07h to 07h through a mechanical counter
	FG10	Av. of 24 hourly measurements of 15-min average (0-0 LT)
	FG12	Av. of 8 measurements of 10-min average (0-0 UT)
RR	RR3	AM today 06,07,08 until morning next day
	RR5	AM today 07:30 CET until morning next day
	RR9	AM today 06:00 UTC until morning next day
	RR11	AM today 05:40 UTC until morning following day

Table 9: Observation sources used for comparison of reanalyses.

A Observations

The ensemble reanalyses are compared against the public domain observations used by the European Climate Assessment & Dataset, [Klein Tank et al., 2002]. Each variable in this archive is calculated in a number of different ways dependent on location and time. Because of this, some variables have been calculated using data which is independent of those assimilated by the reanalyses. To ensure independence of observations used for validation, for 2m mean temperature (TG) and 10m mean wind speed (FG), sources were selected that have not been assimilated by the reanalyses. For daily accumulation of precipitation (RR) sources were selected which have a consistent period. These are detailed in table 9. Station positions are shown in figure 31.

B Calculations

The nearest neighbour grid point was used to compare with observation data. A correction to temperature was applied based on height difference between the model and the observation ($T = T_{raw} - 6.5(h_{ob} - h_{mod})/1000$) following the standard atmosphere lapse rate, [International Organization for Standardization, 1975].

B.1 Error & Spread

RMS error and mean error are the absolute and signed differences, respectively, between the model and observation meaned over all observations.
Spread is calculated as follows

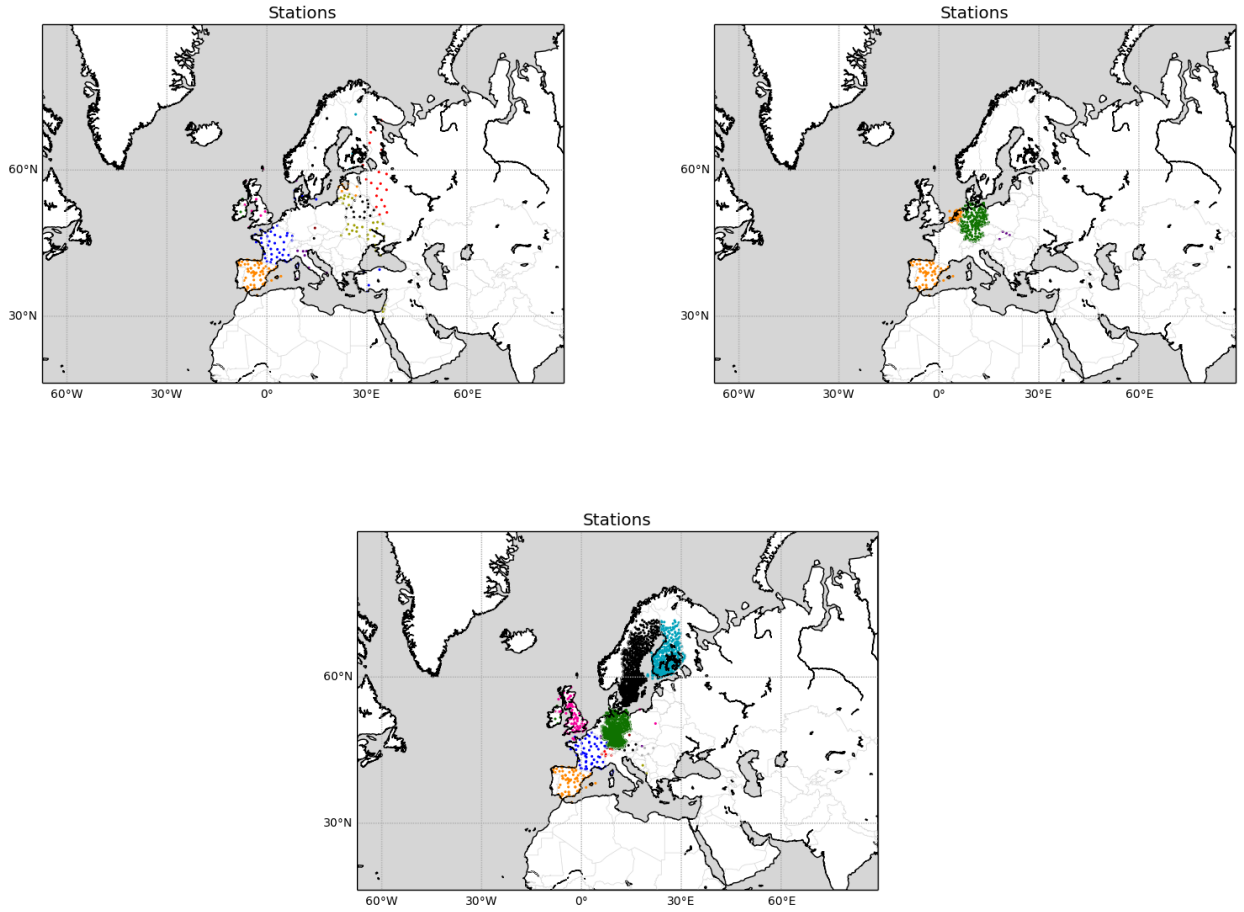

Figure 31: Station positions. Clockwise from top left: daily mean 2m temperature, daily mean 10m wind speed and 24h precipitation.

$$\text{Spread} = \frac{1}{N}\sum_{n=1}^{N}\sqrt{(\sum_{m=1}^{M} X_{mn}^2/(M-1))} \tag{1}$$

where N is the number of observations, M is the number of ensemble members and X is an ensemble error mode, such that

$$X_{mn} = x_{mn} - \bar{x_n} \tag{2}$$

for 'raw' spread, where x_{mn} is the model value for member m and observation n, and

$$X_{mn} = C\left(x_{mn} - \bar{x_n}\right) \tag{3}$$

for inflated spread, where C is a multiplication factor to account for observation and representivity errors.

B.2 Rank Histogram

The rank of an observation is defined as the number of ensemble members that it is greater than. In the case where the observation is zero and a number of ensemble members are zero the rank is given as a random integer up to the number of zero value ensemble members. The proportion of observations with each rank is displayed in the rank histogram.

B.3 CRPS

$$\text{CRPS} = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty} F_i^x(z) - F_i^y(z)dz \tag{4}$$

where F^x and F^y are cumulative density functions for the ensemble and observation, respectively.

B.4 Brier Score

$$\begin{aligned}
\text{BS} &= \frac{1}{N}\sum_{i=1}^{N}\left(p_i^x - p_i^y\right)^2 && (5)\\
\text{REL} &= \frac{1}{M+1}\sum_{m=0}^{M} n_m\left(p_m^x - p_m^{\bar{y}}\right)^2 && (6)\\
\text{RES} &= \frac{1}{M+1}\sum_{m=0}^{M} n_m\left(p_m^{\bar{y}} - \bar{p^y}\right)^2 && (7)\\
\text{UNC} &= \bar{p^y}(1-\bar{p^y}) && (8)\\
\text{BS} &= \text{REL} - \text{RES} + \text{UNC} && (9)
\end{aligned}$$

where p^x and p^y are probabilities for the ensemble and observation, respectively, and n_m are the number of events with model probability p_m^x. BS is the Brier Score, composed of reliability (REL), resolution (RES) and uncertainty (UNC).

B.5 Receiver Operating Characteristic

$$\text{Hit Rate} = \frac{\text{observed \& modelled}}{\text{all observed}} \tag{10}$$

$$\text{False Alarm Rate} = \frac{\text{modelled, but not observed}}{\text{all not observed}} \tag{11}$$

$$\tag{12}$$

References

[Bach, 2015] Bach, L. (2015). Kalman ensemble DA development. *UERRA deliverable*, 2(12).

[Grimit and Mass, 2007] Grimit, E. and Mass, C. (2007). Measuring the ensemble spread2̆013error relationship with a probabilistic approach: Stochastic ensemble results. *Mon. Wea. Rev.*, 135(1):203–221.

[International Organization for Standardization, 1975] International Organization for Standardization (1975). Standard atmosphere. *ISO*, 2533.

[Jermey et al., 2015] Jermey, P. M., Davie, J., Mahmood, S., and Renshaw, R. J. (2015). Development of ensemble variational data capability and report demonstrating ensemble uncertainty products. *UERRA deliverable*, 2(1).

[Jermey and Renshaw, 2016] Jermey, P. M. and Renshaw, R. J. (2016). Precipitation representation over a two-year period in regional reanalysis. *Q.J.R. Meteorol. Soc*, 142(696):1300–1310.

[Klein Tank et al., 2002] Klein Tank, A. M. G., Wijngaard, J. B., Knnen, G. P., Bhm, R., Demare, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Mller-Westermeier, G., Tzanakou, M., Szalai, S., Plsdttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio Lpez, J., Dahlstrm, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and Petrovic, P. (2002). Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Int. J. Climatol.*, 22(12):1441–1453.

[Laloyaux et al., 2016] Laloyaux, P., Balmaseda, M., Dee, D., Mogensen, K., and Janssen, P. (2016). A coupled data assimilation system for climate reanalysis. *Q.J.R. Meteorol. Soc*, 142(694):65–78.

[Leith, 2007] Leith, C. (2007). Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, 135(1):203–221.

[Lockhoff, 2017] Lockhoff, M. (2017). personal communication.

[Roberts and Lean, 2008] Roberts, N. and Lean, H. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, 136(1):78–97.

[Saetra et al., 2004] Saetra, O., Hersbach, H., Bidlot, J.-R., and Richardson, D. S. (2004). Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev*, 132(6):1487–1501.

[Unden et al., 2014] Unden, P. et al. (2014). Uncertainties in ensembles of regional reananlyses.