TECHNICAL MEMORANDUM

820

Evaluation and diagnostics of the CERA-20C climate reanalysis ensemble

P. Dahlgren

Research Department

January 2018

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts
Shinfield Park, Reading, RG2 9AX, England

Abstract

CERA-20C is a coupled ocean-atmosphere climate reanalysis produced by the European Centre for Medium Range Weather Forecasts (ECMWF) and it covers the years 1901 to 2010. The atmospheric data assimilation is produced with a 4D-Var technique and assimilates surface pressure observations and wind speed at 10 metres over sea. It runs as an Ensemble of Data Assimilations (EDA) with 10 ensemble members. This report investigates different aspects of the behaviour of the atmospheric ensemble throughout the century.

By looking at how the analysis ensemble spread evolves with maps and time series it can be seen that the spread gradually decreases during the century as more observations are being assimilated into the system. It also shows that the surface observations have a substantial impact on the atmospheric state up to around 300hPa. This is especially noticeable in the northern hemisphere in the later part of the century.

Ensemble reliability studies have been performed for 2010 using observations of surface pressure as a reference for the truth. These show that the ensemble spread can indeed predict the error in the analysis ensemble mean state. It captures the spatial and temporal variations of the error but it is under dispersive (spread too small) in the extra-tropics. The larger the error, the larger the mismatch between spread and error; there is a conditional bias between the spread and the error. Compared to the operational EDA at ECMWF, the spread and error in CERA-20C are much larger and the operational EDA (which has 25 members) is more reliable.

Lastly, the ensemble mean and spread were compared to the climatological spread and variability by using a metric called Relative Entropy (RE). RE takes into account how both the ensemble mean and spread compares to the climate distribution and is used to assess the information content in the CERA-20C reanalysis. Maps from the beginning and the end of the century show that the information content increase with time and that the analyses in southern hemisphere in the beginning of century are close to the climate distribution.

1 Introduction

Within the ERA-CLIM2 project the European Centre for Medium Range Weather Forecasts (ECMWF) has produced a century long climate reanalysis, CERA-20C (Laloyaux et al. (2017)), covering the years 1901 to 2010. The reanalysis is produced with the ECMWF coupled data assimilation system described in Laloyaux et al. (2016). All non linear integrations are done with an ocean atmosphere coupled model and the analyses in the ocean and in the atmosphere share the same time window, but the inner loop minimisations are done separately. Increments produced in the ocean (atmosphere) influence the atmosphere (ocean) in the outer loop making the data assimilation itself coupled. The upper-air analysis (4D-Var) assimilates surface pressure observations and wind speed at 10 metres over sea, while the ocean assimilates subsurface salinity and temperature profiles.

CERA-20C is produced with an Ensemble of Data Assimilations (EDA) technique with 10 ensemble members, (Bonavita et al. (2012); Fisher (2003)), where each member is produced in a separate data assimilation cycle. The analysis time window is 24h long and the horizontal resolution in the non linear atmospheric model is T159 which is $\sim$125km grid spacing. Each individual member assimilates perturbed observations, except the control member. Different realisations of the sea surface temperature (SST) from the HadISST2 product was also used by the different members. Model uncertainties are accounted for by adding stochastic perturbations to the physical tendencies in the non linear model integrations. These components keeps the different ensemble members separated.

This work will study the behaviour of this ensemble as it is unique in many ways compared to how the

EDA technique is used for Numerical Weather Prediction applications ECMWF, e.g. the time frame, 110 years, during which the number of observations changes dramatically. It will be investigated what the ensemble spread looks like and how it changed during the century. The reliability of the CERA-20C ensemble will also assessed and compared to the operational EDA at ECMWF. Reliability here refers to how well the CERA-20C ensemble manage to predict the spatial and temporal variations in the error of the ensemble mean.

In the next section the ensemble spread is displayed as maps and time series. Then, in section 3, the reliability of the ensemble is assessed where the error is computed by comparing to observations in the year 2010. These results are also compared to the reliability of the operational EDA at ECMWF. In section 4 the long term behaviour of the ensemble will be studied using different statistical metrics that relate the ensemble mean and spread to the climatological mean and variability. Section 5 gives a summary and conclusions.

2 Displaying the ensemble spread

In this section the behaviour of the ensemble spread throughout the century is presented.

In figure 1 the ensemble spread of analysed surface pressure valid at 00 UTC is plotted for different years throughout the century. The ensemble spread is averaged over three months to better illustrate the characteristics of the geographical distribution. A snapshot of a single date would be more patchy and influenced by single events going on at that particular day. Overall, comparing the different years, it is clear that the spread gradually decreases during the century as the number of observations increases and the observations themselves are improved. The largest decrease in the spread can be seen in the northern hemisphere extra-tropics which is a result of more observations. In the southern hemisphere extra-tropics the spread also decrease but not as much since it is still quite poorly observed compared to the northern hemisphere. Overall, looking at the geographical distribution over the globe, the southern hemisphere extra-tropics has the highest ensemble spread and the extra-tropics in general has larger spread that than the tropics. However, the spread over northern hemisphere extra-tropics over land, in modern times, is lower than the spread over the tropics in general. These finding are all reasonable. Spread in an ensemble system can be high in areas which are poorly observed as the model will be less constrained. The spread will also be higher in areas which are more dynamically active, such as the extra-tropical westerlies where baroclinic instability triggers low pressure systems and many non linear wave interactions occur.

It is also informative to see how the spread evolves in the whole atmosphere. CERA-20C only assimilates data at surface level but we expect to get a fairly good representation of the large scale flow in the troposphere. To illustrate this the spread of CERA-20C analyses valid at 00 UTC, over the northern and southern hemisphere extra-tropics, have been averaged on different pressure levels in the vertical. The time evolution of these averages are shown in figure 2. It shows that in the northern hemisphere, in the later parts of the century, the surface pressure observations have a substantial influence on the atmospheric state up to around 300hPa. Overall, figure 2, also shows that the spread gradually decreases with time and that the largest decrease is around 1950 in the northern hemisphere. The spread in the zonal wind is dominated by activities at the jet stream level and seems to keep that characteristic throughout the century, in both hemispheres.

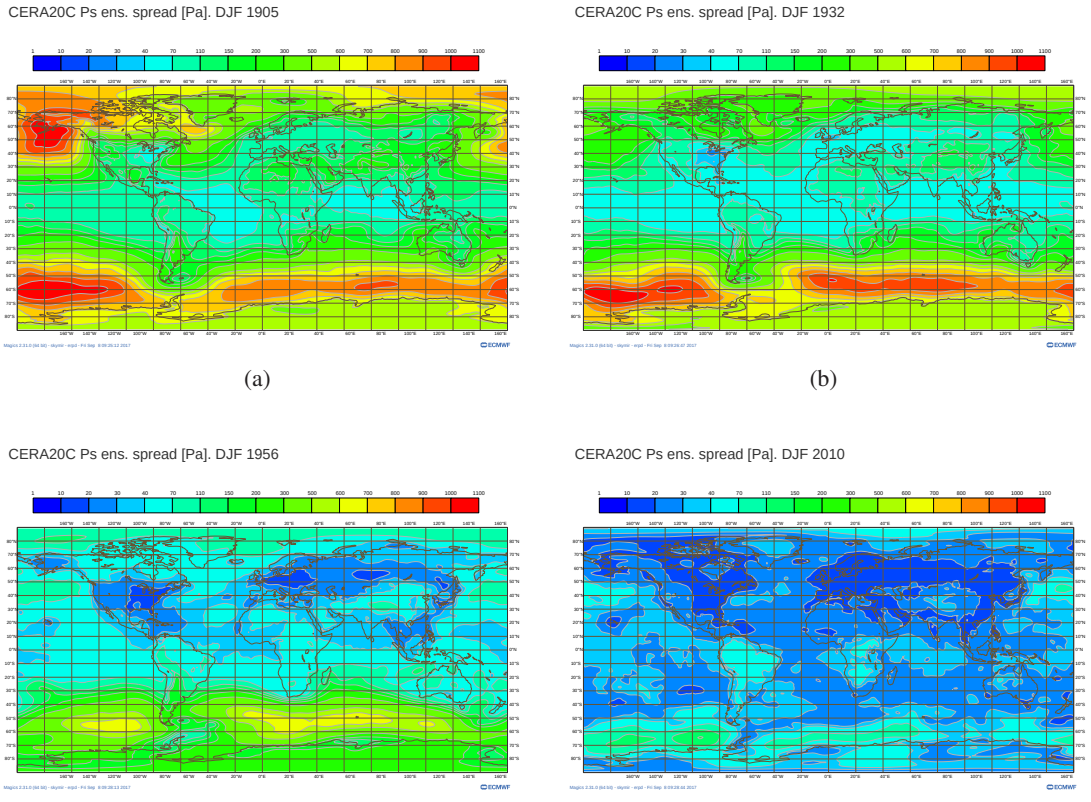
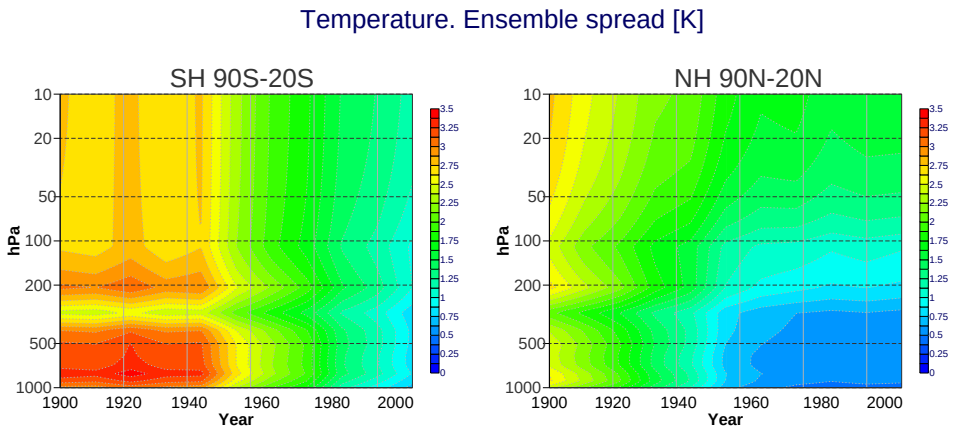
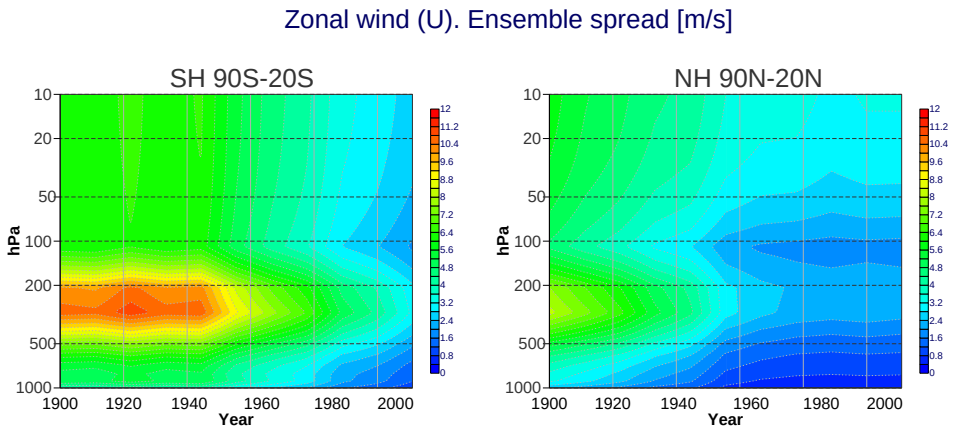


*Figure 1: CERA-20C analysis ensemble spread of surface pressure at 00 UTC. Average over three months. (**a**): 1905 (1 Dec 1904 to 28 Feb 1905) (**b**): 1932 (1 Dec 1931 to 28 Feb 1932) (**c**): 1956 (1 Dec 1955 to 28 Feb 1956) (**d**): 2010 (1 Dec 2009 to 28 Feb 2010)*



(a)



(b)

*Figure 2: Time evolution of the CERA-20C analysis ensemble spread at 00 UTC in the extra-tropical regions. Southern hemisphere (90S-20S) to the left and northern hemisphere (20N-90N) to the right. y-axis is the vertical level in hPa and the x-axis shows the time, the colour scale shows the ensemble spread. (**a**): Temperature (**b**): Zonal wind*

3 Ensemble reliability diagnostics

In the previous section the ensemble spread was presented as maps and time-series. These show sensible qualitative results with decreased spread for periods and regions for which the model is constrained by a relatively large number of assimilated observations. In this section the reliability is quantitatively assessed by looking at diagnostics that relates the ensemble spread to the ensemble mean error, which are explained in detail in Leutbecher (2010).

In a perfectly reliable ensemble of size M the members x_j are independent, identically distributed random variables with mean μ and variance σ_p. Subscript p refers to *perfect* to emphasise that σ_p is *the variance we would get if the ensemble was perfectly reliable*. To be able to estimate the error in the ensemble mean, a reference *true* state, y, needs to be defined. This can be either a model that is believed to be better than the one used to produce the ensemble, or observations. With the addition of the below notations,

$$
\begin{aligned}
\mathbb{E} &= \text{Expectation operator} \\
\bar{x}_m &= \text{Ensemble mean} \\
\varepsilon_e &= \text{Estimated error of the ensemble mean} \\
\varepsilon_e^2 &= (y - \bar{x}_m)^2 \\
s_e &= \text{Estimated spread of the ensemble} \\
s_e^2 &= \frac{1}{M-1} \sum_{j=1}^{M} (x_j - \bar{x}_m)^2
\end{aligned}
$$

it is now possible to write down , following Leutbecher (2010), how the estimated error and spread depend on ensemble size:

$$\mathbb{E}\varepsilon_e^2 = \frac{M+1}{M}\sigma_p^2 \tag{1}$$

$$\mathbb{E}s_e^2 = \frac{M-1}{M}\sigma_p^2 \tag{2}$$

Eq. 1 does not take into account that the reference truth can also have uncertainties. If we have an estimate of that uncertainty, σ_o, then Eq. 1 can be expanded to account for this by subtracting it from the estimated error (see Yamaguchi et al. (2016) for details):

$$\mathbb{E}\left[\varepsilon_e^2 - \sigma_o^2\right] = \frac{M+1}{M}\sigma_p^2 \tag{3}$$

Now equations 2 and 3 can be combined to obtain the following relation:

$$\mathbb{E}\varepsilon_e^2 = \mathbb{E}\left[\sigma_o^2 + \frac{M+1}{M-1}s_e^2\right] \tag{4}$$

which is what will be used for the ensemble reliability diagnostics. Equation 4 implies that if the ensemble is reliable, then for a large enough sample the error in the ensemble mean equals the ensemble spread

times a correction factor that accounts for the finite ensemble size, plus the error of the reference truth used to assess the ensemble mean error.

To be able to compute the error in equation 4 a reference truth needs to be defined as mentioned earlier. Here we are studying the CERA-20C reanalysis which only assimilates surface observations and it is therefore appropriate to look at surface pressure or mean sea level pressure as this should be the best described variable with the best skill.

The diagnostics presented here are calculated with observations as a reference for the truth. One reason is that much of what is needed to calculate the statistics in equation 4 can be obtained from the output of the assimilation system; the analysis feedback files. Another advantage of calculating the diagnostics in observation space is the importance of removing the systematic part of the error assessment so that the random component of the error is used. Since surface pressure observations undergo variational bias correction (Dee (2004)) some of the bias between observation and model is removed by the assimilation system. Accounting for the errors in the reference used for verification is also important, (Yamaguchi et al. (2016)), especially for short lead times, and here we can use the observation errors. These are not constant for surface pressure data but vary with observation type and station altitude, so they should be treated as unique to each observation. A disadvantage of doing the statistics in observation space is that observations are inhomogeniously distributed, i.e. we only get statistics from observation locations.

To evaluate the left hand side of equation 4, the error should be calculated with respect to the ensemble mean. The error in the reference truth, σ_o, is the observation error used in the assimilation and the spread s_e^2 is the spread of the ensemble, interpolated to the observation location. It was first tested to use the ODB column *eda_spread* for this but it turned out to only be valid at the beginning of the time window which created a big mismatch between the error and spread for data later in the window. It turned out to be much more appropriate to calculate the spread from the ensemble members directly by:

$$\overline{H(x)} = \frac{1}{M}\sum_{j=1}^{M} H(x_j)$$

and then

$$s_e^2 = \frac{1}{M-1}\sum_{j=1}^{M}(H(x_j) - \overline{H(x)})^2$$

Eq. 4. is here evaluated by taking the square root of both sides:

$$\sqrt{\mathbb{E}(y - \overline{H(x)})^2} = \sqrt{\mathbb{E}\left[\sigma_o^2 + \frac{M+1}{M-1}s_e^2\right]} \tag{5}$$

where y is the bias corrected observation.

To quantify the reliability of the variations of the CERA-20C ensemble spread, pairs of error versus obs error plus ensemble spread were collected over 3 months: Dec 2009 to Feb 2010. The sample was then stratified by the spread and partitioned into 10 equally populated bins. Then, equation 5 was evaluated in each bin, i.e. the average error versus the average of obs error plus ensemble spread. To illustrate seasonal variations, the same was also done for Jun 2010 to Aug 2010. If the bins are plotted in a

diagram with the error (left hand side of Eq. 5) on the y-axis and the spread (right hand side of Eq. 5) on the x-axis, a perfectly reliable ensemble would lie on the diagonal. If the curve is above the diagonal the ensemble is under-dispersive and if it is below it is over-dispersive.

In figure 3 the spread-error reliability is plotted for winter and summer 2010 as described above, and for northern hemisphere extra-tropics (top figure), tropics (middle figure) and southern hemisphere extra-tropics (bottom figure). In addition, the spread-error for the operational ECMWF EDA in DJF 2017 is also plotted. The operational EDA has higher resolution, T639, 25 ensemble members and assimilates the full observing system and has been included as a reference as it should be the best calibrated, state of the art EDA available.

If the ensemble system captures the variations in the ensemble mean error, a spread-error curve should have a slope that is positive. This means that as the error increase so should the spread, on average. Figure 3 shows that the CERA-20C is under-dispersive in the extra-tropics, with the largest errors in the winter season. In the extra-tropics, for the first 7 bins, the spread error curves are almost linear with a slope greater than one. This means that the errors in the bins increase faster than the spread and there is a conditional bias in the spread, most pronounced in the winter season. It is also clear from figure 3 that the errors are larger in the southern hemisphere which stems from the system being less constrained there due to the lower observation coverage. In the highest two to three bins the slope of the curves decrease and become more parallel to the diagonal. This may be a result of computing the error relative to the ensemble mean which has the effect that the largest errors are filtered out. In the tropics the ensemble seems to be quite reliable as the curves are close to the diagonal. The operational EDA is much more reliable than the CERA-20C ensemble as it is much closer to the diagonal and the errors and corresponding spread are on a different, lower, level.

Comparing spread and error at analysis time requires careful handling of data as the uncertainty in many of the components in Eq. 5 may be comparable in size, e.g. not taking into account the errors in the reference truth and/or the bias component of the error may change the spread-error curve quite substantially. To illustrate this, a sensitivity test has been performed where the spread-error relation (Eq. 5) has been evaluated in different ways. The reference in the sensitivity test is northern hemisphere winter (DJF 2010), i.e. the same as the thick red line in the top plot in figure 3. First, the observation error is omitted in Eq. 5. and the result is the thin dashed red line in figure 4. The curve is significantly shifted to the left, so it is definitely important to take the error in the reference into account. Next, observations are used without bias correction (the reference is always the thick red curve), and we get the solid blue line (figure 4). The error is clearly higher so this shows the importance of removing the systematic part of the error. This may be even more important when using another model as a proxy for the truth. Another way to test if there are systematic errors influencing the RMS error is to compute the standard deviation of the error, thus removing the average of $(y - \overline{H(x)})$ from the RMS. This is the thin dashed black line which is indistinguishable from the reference so it seems like the variational bias correction has removed most of the bias.

In the last part of the sensitivity test the effect of the length of the assimilation window is studied. In CERA-20C the assimilation window is 24h long which means that both the error and spread will grow within the window. As a last test, only observations from the first 3.5h of the assimilation window are used in the spread-error calculations which gives the thin dashed blue line in figure 4. For the first 8 bins the slope increased and the spread-error reliability is degraded. This means that the ensemble spread growth within the assimilation window improves the match between the spread and the error, but the effect is small.

The thin dashed blue line in figure 4 levels off and decrease between bin 7 and 8 before increasing

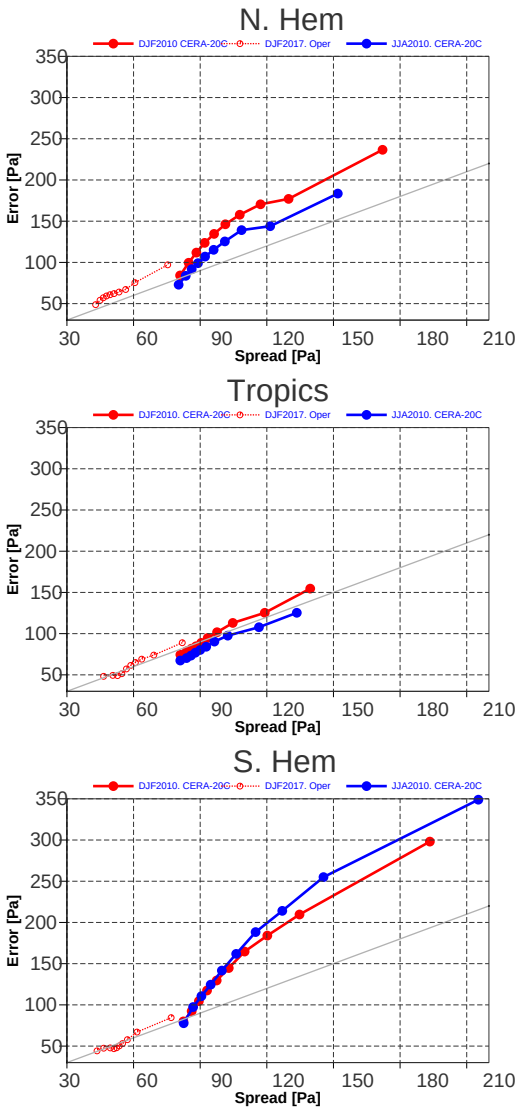


Figure 3: Spread-error reliability curves for CERA-20C. Red is CERA-20C DJF 2010, blue is CERA-20C JJA 2010, thin red line is ECMWF operational EDA in DJF 2017.

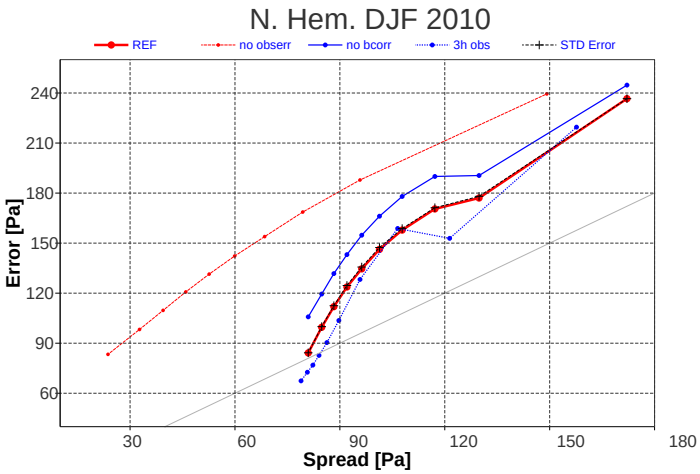


Figure 4: Sensitivity of the spread-error relation by changing the various components in Eq. 5 or calculating them in a different way. Thick red line is the reference. Thin red dashed line: omitting the observation error σ_o. Solid blue line: using observations y without bias correction. Dashed blue line: only using data from the first 3 hours of the assimilation window. Thin black dashed line: standard deviation of the error instead of root mean square.

again. Similar behaviour can be seen in all spread-error curves for the northern hemisphere extra-tropics, except the thin dashed red line in figure 4. It has not been investigated fully where this behaviour comes from but one possible explanation is that different parts of the sample have different error characteristics, e.g. stations at or close to sea level compared to stations at higher altitudes. Another thing can be that different observation types have different observation errors assigned.

With all of the above being said, there is always an offset between the spread and the error and the CERA-20C ensemble is definitely under-dispersive in the extra-tropics. The sensitivity study shows that the calculation of spread-error reliability is a delicate task, but none of the tests described above could remove the systematic, conditional offset between the spread and the error. That leaves us with the EDA spread itself; it is simply not large enough to match the error in the the extra-tropics. It captures the signal, i.e. the spatio temporal variations of the error but fails to match the magnitude as the error increase.

4 Reanalysis information content

Here it will be investigated how the information content in the CERA-20C reanalysis changes in the long term perspective. Since CERA-20C begins in the year 1901 and ends in 2010 it goes through significant changes in the observation coverage. In the early part of the century in the southern hemisphere, the CERA-20C reanalysis is probably very close to a climate simulation and an individual analysis is probably not very informative about the details of the weather that particular day. In the later part of the century, on the northern hemisphere, the analyses are highly informative and can be used as starting points for forecast integrations that are skillful into the medium range. The problem of measuring reanalysis quality or information content will here be addressed from another angle compared to the previous section. Quality will now be assessed by comparing the ensemble information (mean and spread) to the climatology of the model. If the ensemble distribution is indistinguishable from the climate distribution, then we say that the CERA-20C reanalysis provide no information. Climatology in this context means

the model's own equilibrium state and the model's own variability around that equilibrium.

A suitable metric to use is *Relative Entropy* (RE) which has been used to visualise the information content of the the 20th century reanalysis (20CR)[1], Compo et al. (2011), and also to assess the predictability of surface temperature at seasonal time scales Tang et al. (2014). Relative entropy can be described as the information loss sustained by assuming climatology when the prediction distribution is available (Kleeman (2002)). Provided a discrete set of predicted p_i and climatological q_i states then the formal definition of RE reads:

$$RE = \sum_i p_i ln\left(\frac{p_i}{q_i}\right) \tag{6}$$

When the climatological and predicted distributions are Gaussian and of finite dimension, then for the univariate case RE becomes:

$$RE = \frac{1}{2}\left[\underbrace{ln\left(\frac{s_c^2}{s_e^2}\right)}_{T1} - \underbrace{\left(1-\frac{s_e^2}{s_c^2}\right)}_{PP} + \underbrace{\frac{(\mu_e-\mu_c)^2}{s_c^2}}_{ST}\right] \tag{7}$$

where s_c and μ_c is the climatological spread and mean while s_e and μ_e is the ensemble spread and mean. RE contains two terms (T1 and PP) that relate the ensemble spread to the climatological variability and one term (ST) that relate the ensemble mean's departure from the climatological mean to the climatological variability. T1 stands for term number 1 while PP is short for Potential Predictability (Kleeman (2002)). ST means Signal Term. The first two terms, T1 and PP, indicates that when the ensemble spread is close to the climatological variability, then T1 and PP become close to 0 and and CERA-20C provide little or no information. When the ensemble spread is small compared to the climatological variability, then T1 becomes large and PP close to 1 and CERA-20C provide extra information over climatology. In figure 5 T1 and PP have been plotted with $s_c^2 = 1.0$Pa. From there it can be seen that when the ensemble spread is small T1 is much larger than PP and will dominate the spread related part of the total RE measure. Also, the slope of the T1 curve is steep while the PP curve is very flat. This means that T1 will able to resolve small changes in ensemble spread. When the ensemble spread is high, PP will be able to resolve changes in ensemble spread as the slope of curve is steeper. T1 and PP thus complement each other having different properties in different parts of the spectrum of possible values of s_e in relation to s_c.

For this study Eq. 7 has been applied to CERA-20C analyses of Mean Sea Level Pressure (MSLP) which is one of the most accurate CERA-20C variables as it benefits directly from observation data assimilation. Climatological mean values μ_c have been calculated from CERA-20C analyses with a 30 year centred average for every day between 1915 to 1995. The climatological variability s_c has been calculated as the standard deviation of CERA-20C analyses around the climatological mean. When the climatologies are calculated it is possible to compute RE daily, in each grid point, between 1915 to 1995.

Figure 6 shows the RE for two different days in January 1915 and two other days in 1995. It illustrates very well that the RE metric is responsive to the daily conditions. Individual observations can directly influence the RE, e.g. south of New Zealand 17 Jan 1915 (figure 6(b)) on the island of Macquari. Overall, by looking at the patterns in figures 6(a) and 6(b) and then 6(c) and 6(d), it can be seen that the information content in the analysis increases with time. In 1915, the RE values in the southern hemisphere is

[1]RE applied to 20CR is often referred to as the *fog of ignorance* and examples of its use can be found on the internet: http://reanalyses.org/atmosphere/dispersing-fog-ignorance

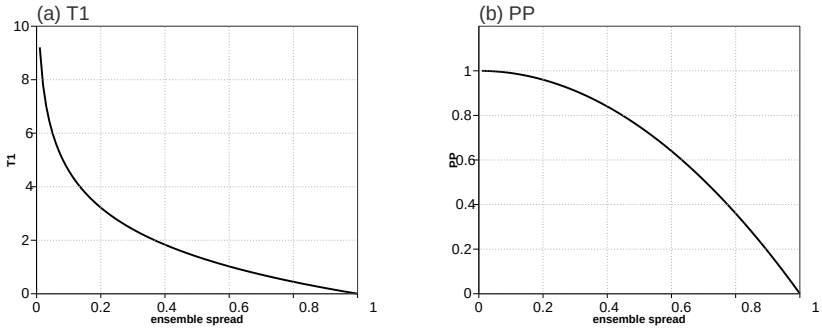


Figure 5: First two terms of Eq. 7 plotted with climatological variability, s_c^2, set to 1.0Pa.

largely close to zero and CERA-20C is almost indistinguishable from the climate distribution.

A case study will now be presented to illustrate the different behaviour of the different terms in RE. On 13 Jan 1995 there was an outstretched deep low pressure system on the leeward side (east coast), of Greenland (not shown). Ensemble members differed in the exact position of the low pressure which made the ensemble spread high. A map of RE, figure 7(a), shows higher values in the same area thus indicating there is still valuable information in the ensemble system. The high RE values comes from the ST term, figure 7(d), which means the ensemble mean is different from the climatological mean. The spread related terms, T1 and PP, have low values in the same area and therefore downgrades the reanalysis, figures 7(b) and 7(c). The same figures also shows that the PP term is unable to resolve small variations in ensemble spread as the values are very uniform except where the ensemble spread is very high. T1 on the other hand resolves a lot more structures, and the values are also much higher than the PP term. ST and T1 have values that are comparable in size. The behaviour of the T1 and PP terms can be explained from figure 5; The ensemble spread is low compared to the climatological variability and therefore we are at left part of the x-axis and T1 dominates over PP and resolves small variations in ensemble spread.

To better illustrate how the information content changes with time in the long term perspective the RE metric has been averaged over different latitude bands. The area averages can then be displayed as time series, figure 8(a). It shows an overall trend that the information content does indeed increase during the century and that the northern hemisphere is better described than the southern. It is also noticeable that the seasonal difference is bigger in NH (thick solid red line vs thin dashed red line) than the SH (thick solid blue line vs thin dashed blue line). In the winter season the weather is more dominated by baroclinic activities which gives a higher climatological variance in the surface pressure. By assimilating surface pressure observations the large scale synoptic patterns will be better described and the chance will increase of reducing the ensemble spread in areas where the climatological variance is high. This will be rewarded with a higher information content. In the northern hemisphere there are enough observations to describe the large scale synoptic patterns to get this reward, while there are simply not enough observations in the southern hemisphere to get an accurate synoptic description. This is why the seasonal difference is visible in the NH but not on the SH.

Time series of the individual terms show that the curves in figure 8(a) are completely dominated by the T1-term, figure 8(b). The PP-term on the northern hemisphere saturates very quickly to values above 0.9 and shows almost no trend after 1950, figure 8(c). Time series of the ST-term shows no trend at all in the NH throughout the whole century, figure 8(d). In the southern hemisphere, PP is quite responsive

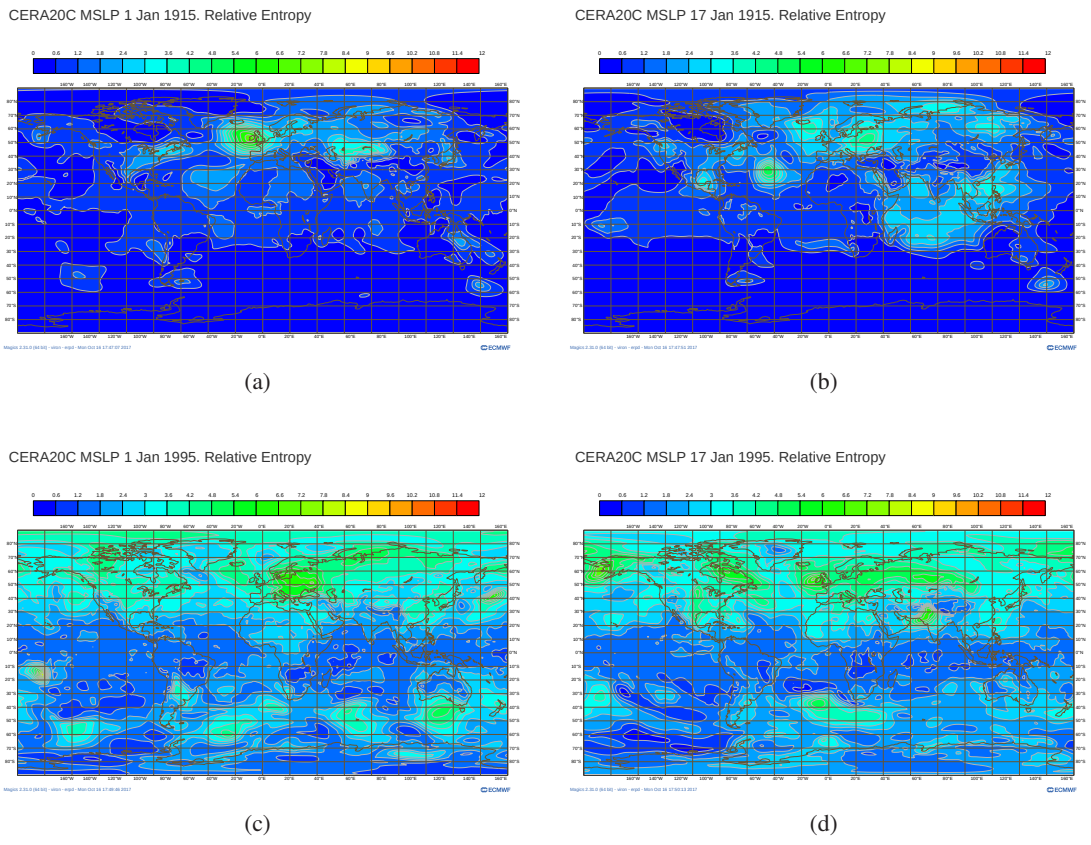


*Figure 6: Maps of relative entropy for four different dates in 1915 and 1995. (**a**): 1 Jan 1915 (**b**): 17 Jan 1915 (**c**): 1 Jan 1995 (**d**): 17 Jan 1995*

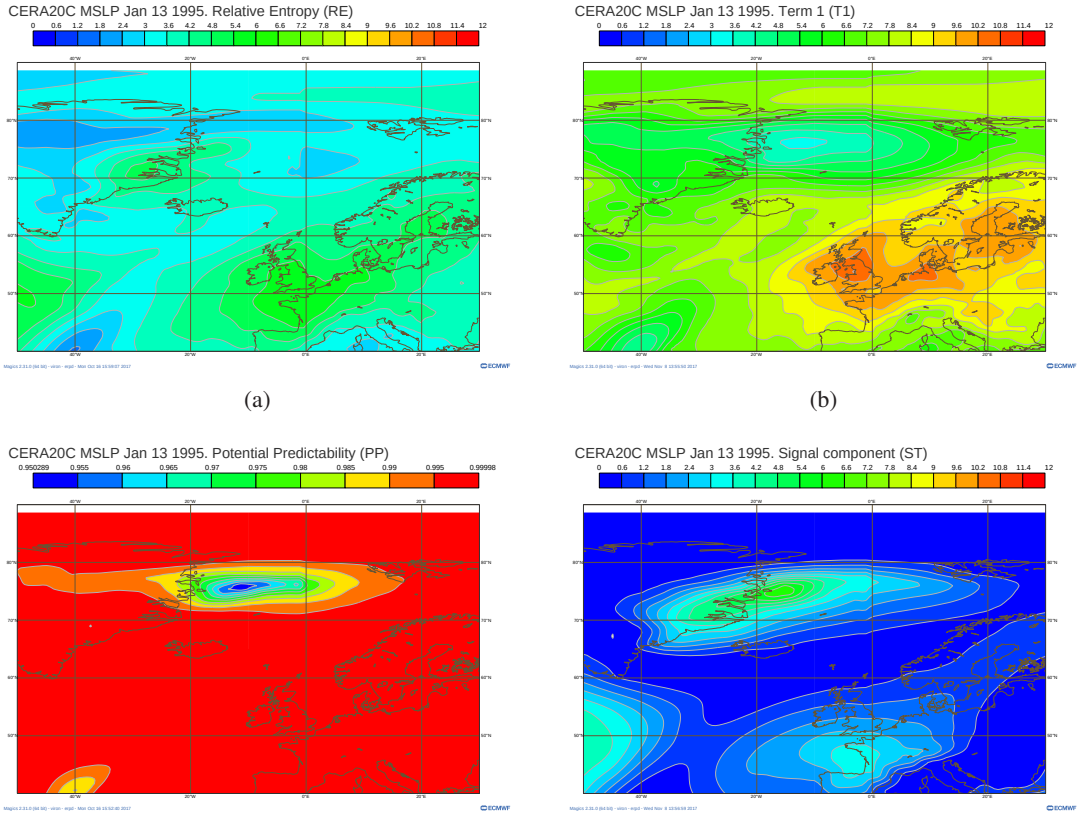


*Figure 7: Maps of Relative Entropy (RE), Potential Predictability (PP), T1 and Signal term (ST) around Greenland on 13 Jan 1995. (**a**): RE (**b**): T1 (**c**): PP (**d**): ST*

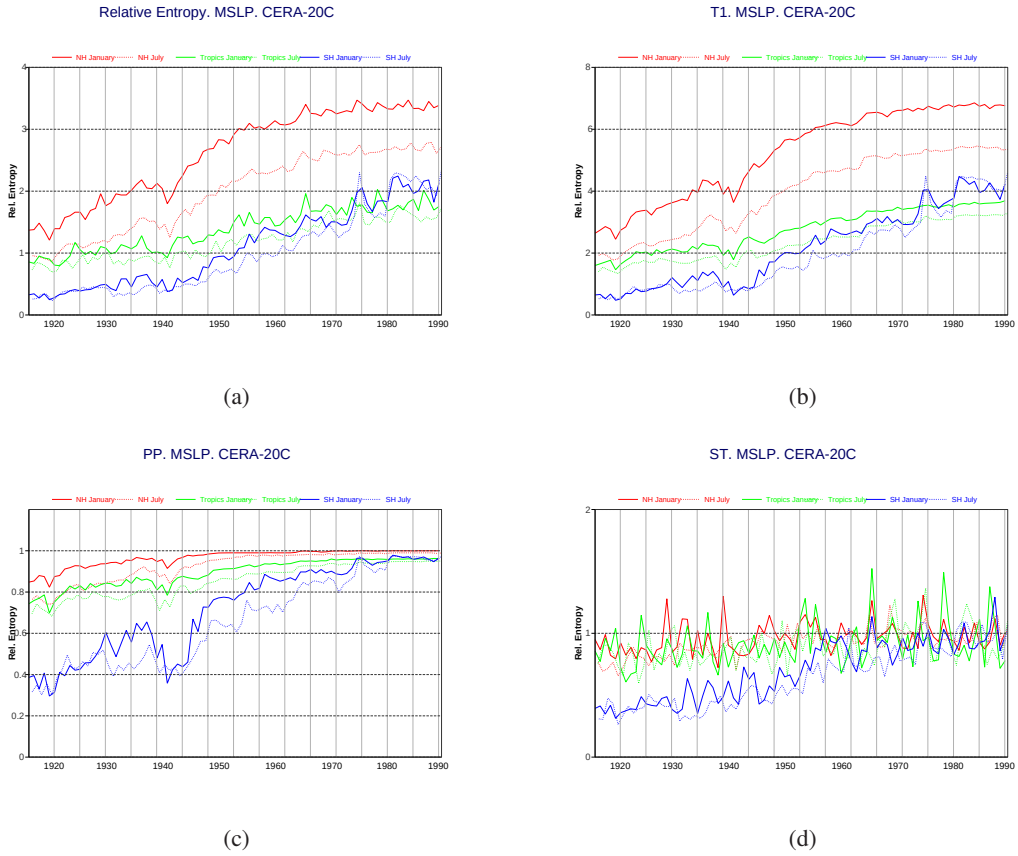


*Figure 8: Times series of Relative Entropy (RE) and the different terms T1, PP and ST for different latitude bands and months. Thick (thin) red line is Northern hemisphere extra-tropics in January (July). Thick (thin) green line is Tropics in January (July). Thick (thin) blue line is Southern hemisphere extra-tropics in January (July). (**a**): Total relative entropy (RE). (**b**): T1. (**c**): PP. (**d**): ST.*

up until 1955. As the observation coverage is very sparse in the SH in the early part of the century the ensemble spread is high in relation to the climatological variability and we are in right part of the x-axis in figure 5, so this is reasonable. The effects of world war 2 is clearly seen, blue line figure 8(c). This response is due to the loss of almost all marine observations during the war, and when the observations come back PP increases very fast. The ST did not react to these changes in the observing system, blue line figure 8(d). Instead, ST reacts to an increase in land stations over Australia and South America between 1955-1960. After that it stays almost constant. In data sparse regions it seems like PP and ST reacts to different types of changes in the observing system. A geographically well distributed increase of observations makes PP react (observations over sea), while an increase in the number of fixed land stations made ST react (observations over land).

5 Summary and Discussion

This report has investigated different aspects of the CERA-20C climate reanalysis ensemble. First, the spread was presented as maps and time series which were conceptually analysed. These showed that the behaviour of the ensemble spread looked reasonable. The overall spread decreases during the century and the spread in the northern hemisphere decrease rapidly in response to increased observation coverage and

quality around 1950. It could also be seen that by assimilating surface observations only it is possible to influence the atmospheric state up to around 300hPa. Maps of surface pressure ensemble spread showed that the spread is highest in southern hemisphere extra-tropics which is reasonable as the observation coverage is low while the dynamic activity is high. In the northern hemisphere extra-tropics, where more observations are available the spread showed more geographical differences; high spread in northern Atlantic and Pacific, lower spread over continental Europe and north America.

Reliability diagnostics showed that the CERA-20C ensemble spread can be used to describe the ensemble mean error. It is however under-dispersive which means that the spread is too small and the offset between the spread and the error is conditional, i.e. the spread fails to match the assessed error as the errors increase. It was found to be a delicate task calculating spread-error diagnostics. One can use another model as a proxy reference for the true state, or observations. In either case, an estimate of the error in the proxy truth is necessary. Ignoring, or not using a relevant assessment of the error in the reference can have a substantial impact on the the spread-error reliability diagnostics. It has also been shown that the bias between the ensemble mean and the reference truth needs to be estimated and removed from the error estimate. Again, failing to do so or not doing it accurately can give different spread-error diagnostics results. Here the spread-error diagnostics were computed in 2010 because that is the most recent year for which CERA-20C is available. The ECMWF operational analysis could have been used as reference truth and many calculations were in fact done using it in parallel with the observation based diagnostics. In the end, it was easier to account for the biases and errors in the reference with the observation based diagnostics. The biases were quite effectively removed by the variational bias-correction and the assessed observation error could be used as the error in the reference truth. One should however bear in mind that observation errors are not always a proper reflection of the errors in the observations as they are often inflated to account for e.g. representativeness.

One reason for the under-dispersive behaviour of CERA-20C can be the resolution in the non linear integrations, T159 ($\sim 125km$ grid spacing). Earlier studies have found that increasing the resolution can give more realistic estimates of the uncertainties of the data-assimilation cycle, see e.g. Bonavita et al. (2016), and thus a better spread-error relation. In general, systematic errors between spread and error reflects deficiencies in the representation of the main sources of uncertainties in the EDA. These uncertainties are represented by perturbing the observations, SST and the model physics tendencies.

In section 4, the information content of the CERA-20C was assessed by relating the ensemble mean and spread to the climatological mean and variability. A statistical measure, Relative Entropy, was calculated which includes three terms; two that relates the ensemble spread to the climatological variability and one that relate the ensemble mean departure from the climatological mean to the climatological variability. By studying the behaviour of the total RE measure and the individual terms separately it was found that the different terms are sensitive to different types of situations/conditions and to different types of changes in the observing system; the terms in RE complement each other. The RE measure indicates that the CERA-20C reanalysis ensemble is almost indistinguishable from a climate distribution over large parts of the southern hemisphere in the beginning of the 20th century, but a relatively small number of extra observations, geographically well distributed, can quickly change this situation. On the northern hemisphere the CERA-20C reanalysis provide information better than climatology throughout the whole century, which a drastic improvement from 1945.

Acknowledgments

I would like to thank the following persons for their helpful feedback and advice: Massimo Bonavita, Elias Holm, Martin Leutbecher, Simon Lang, Lars Isaksen, Patrick Laloyaux, Hans Hersbash and Andras Horanyi.
Many thanks also to Patricia de Rosnay and Stephen English for feedback on the manuscript.
The work presented in this report was supported by the ERA-CLIM2 project, funded by the European Union's Seventh Framework Programme under grant agreement number 607029.

References

Bonavita, M., , Hólm, E., Isaksen, L., and Fisher, M. (2016). The evolution of the ECMWF hybrid data assimilation system. *Quart. J. Roy. Meteorol. Soc.*, 142:287–303. doi:10.1002/qj.2652.

Bonavita, M., Isaksen, L., and Hólm, E. (2012). On the use of EDA background error variances in the ECMWF 4D-Var. *ECMWF Technical Memorandum number 664*.

Compo, G., Whitaker, P., Sardeshmukh, N., Matsui, N., Allan, R., Yin, X., Gleason, B., Vose, R., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R., Grant, A., Groisman, P., Jones, P., Kruk, M., Kruger, A., Marshall, G., Maugeri, M., Mok, H., Nordli, Ø., Ross, T., Trigo, R., Wang, X., Woodruff, S., and Worley, S. (2011). Review Article, The Twentieth Century Reanalysis Project. *Quart. J. Roy. Meteorol. Soc.*, 137:1–28. Part A.

Dee, D. (2004). Variational bias correction of radiance data in the ecmwf system. In *ECMWF workshop on assimilation of High Spectral Resolution Sounders in NWP*, pages 97–112.

Fisher, M. (2003). Background error covariance modelling. In *Proceedings of the ECMWF seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean*, pages 45–63.

Kleeman, R. (2002). Measuring Dynamical Prediction Utility Using Relative Entropy. *J. Atmos. Sci.*, 59:2057–2072. doi.org/10.1175/1520-0469(2002)059¡2057:MDPUUR¿2.0.CO;2.

Laloyaux, P., , Balmaseda, M., Dee, D., Mogensen, K., and Janssen, P. (2016). A coupled data assimilation system for climate reanalysis. *Quart. J. Roy. Meteorol. Soc.*, 142:65–78. doi:10.1002/qj.2629.

Laloyaux, P., Balmaseda, M., Broennimann, S., Buizza, R., Dahlgren, P., de Boisseson, E., Dee, D., Kosaka, Y., Haimberger, L., Hersbach, H., Martin, M., Poli, P., and Schepers, D. (2017). Cera-20c: A coupled reanalysis of the twentieth century. In *preparation*.

Leutbecher, M. (2010). Diagnosis of ensemble forecasting systems. In *Proceedings of the ECMWF annual seminar 2009 on Diagnosis of Forecasting and Data Assimilation Systems*, pages 235–266. https://www.ecmwf.int/sites/default/files/elibrary/2010/10725-diagnosis-ensemble-forecasting-systems.pdf.

Tang, Y., Chen, D., and Yan, X. (2014). Potential Predictability of North American Surface Temperature. Part I: Information-Based versus Signal-To-Noise-Based Metrics . *J. Climate.*, 27:1578–1599. DOI: 10.1175/JCLI-D-12-00654.1.

Yamaguchi, M., Lang, S., Leutbecher, M., Rodwell, M., Radnoti, G., and Bormann, N. (2016). Observation-based evaluation of ensemble reliability. *Quart. J. Roy. Meteorol. Soc.*, 142:506–514. doi:10.1002/qj.2675.