EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates

Simon A. Good,1 Matthew J. Martin,2 and Nick A. Rayner1

Received 2 May 2013; accepted 8 November 2013; published 11 December 2013.

[1] We present version 4 of the Met Office Hadley Centre “EN” series of data sets of global quality controlled ocean temperature and salinity profiles and monthly objective analyses, which covers the period 1900 to present. We briefly describe the EN4 data sources, processing, quality control procedures, and the method of generating the analyses. In particular, we highlight improvements relative to previous versions, which include a new duplicate profile removal procedure and the inclusion of three new quality control checks. We discuss in detail a novel method for providing uncertainty estimates for the objective analyses and improving the background error variance estimates used by the analysis system. These were calculated using an iterative method that is relatively robust to initial misspecification of background error variances. We also show how the method can be used to identify issues with the analyses such as those caused by misspecification of error variances and demonstrate the impact of changes in the observing system on the uncertainty in the analyses.


1. Introduction

[2] Observations of the temperature and salinity of the ocean are essential to many climate applications such as the initialization of seasonal and decadal forecasts [e.g., Arribas et al., 2011; Smith et al., 2007], for understanding the variability over time of ocean temperature, heat content, and salinity [e.g., Palmer et al., 2007; Levitus et al., 2012; Lyman et al., 2010; Boyer et al., 2005; Durack and Wijffels, 2010] and for input to ocean reanalyses [e.g., Balmaseda et al., 2008]. It is important that bad data are not allowed to impact on these applications and therefore they rely on having data sets available that have been subjected to a high level of quality control.

[3] In this paper, we present the latest version of the “EN” series of data sets, EN4, which aims to meet this need. It is a collection of ocean temperature and salinity profiles obtained across the global oceans over the period 1900 to present to which a series of quality control checks have been applied. Associated with this are monthly objective analyses with uncertainty estimates. The origin of the series of data sets (and its name) was in two European Union projects: ENACT (Enhanced Ocean Data Assimilation and Climate Prediction; http://www.ecmwf.int/research/EU_projects/ENACT/index.html) and ENSEM-BLES (http://ensembles-eu.metoffice.com/index.html) [see Ingleby and Huddleston, 2007] but, as described in this paper, the quality control procedures have been extended and uncertainty estimates are provided with the objective analyses. The latter is achieved through use of a novel method to estimate uncertainty that is relatively robust to misspecification of the variance of the errors in the analysis backgrounds compared to methods such as optimal interpolation (OI). The method is also used to improve the estimates of background error variance. Data are made freely available for scientific research and private study from the Met Office Hadley Centre observations website, http://www.metoffice.gov.uk/hadobs.

[4] The EN4 data set exists alongside a number of other similar data sets. For example, the World Ocean Database (WOD; latest version available at time of writing was WOD09, Boyer et al. [2009]), which is the main data source used for constructing EN4, and the Coriolis data set for ReAnalysis (CORA) [Cabanes et al., 2013]. Each has its own unique features that may make them most appropriate for a particular project. Features of the EN4 data set include its >100 year time span with all profiles recorded in a particular month provided in a single network common data format (NetCDF) file and the fact that it is updated to the present each month. We also apply the different bias adjustments that have been proposed by various authors [e.g., Wijffels et al., 2008; Ishii and Kimoto, 2009; Levitus et al., 2009; Gouretski and Reseghetti, 2010; Good, 2011; Gouretski, 2012; Hamon et al., 2012] to adjust for time varying biases in mechanical bathythermograph (MBT) and expendable bathythermograph (XBT) profiles and make these data available on request. This gives the user choice over which adjustments to use or the option to use...
2. The EN4 Data Set

[5] The EN4 data set is an incremental development of the previous versions EN2 and EN3 that are documented in Ingleby and Huddleston [2007] and at http://www.metoffice.gov.uk/hadobs/en3. Here, we give a brief overview of the methodology and highlight differences compared to these predecessors.

[6] The EN4 processing system is illustrated in Figure 1. Profile data are obtained from a number of sources, which are described in section 2.1. These are preprocessed (for example, to remove duplicates) and are then subjected to a series of quality control checks and quality flags are assigned (section 2.2). The quality controlled data are used to produce a monthly objective analysis of the temperature and salinity of the ocean. By persisting anomalies through the start and end of the month from the other data sources. Profiles with high vertical resolution (>400 observations within the profile) are thinned so that observations are spaced approximately 1 m apart in the top 100 m of the ocean, 10 m apart above 1500 m depth, and 50 m below that.

[10] The source data sets are highly overlapping. For example, profiles from Argo floats can be found in the Argo, GTSP, and WOD09 data sets. It is necessary to identify and remove the duplicates so that the EN4 data set only contains one copy of each profile. The method for doing this is new in EN4 and is based on that described in Gronell and Wijffels [2008]. The process has four steps:

1. The algorithm tries to find and delete the Argo profiles from WOD09 and GTSP so that only the Argo profiles from the Argo GDAC remain in EN4.
2. Profiles with the same (or very close to the same) location and time are identified and only one is retained. This procedure is used to find duplicates of the same profile from different data sources, to remove low-vertical resolution versions of CTD profiles in WOD09 that are also available at high vertical resolution, and to thin data from types of instruments that can yield extremely dense observations.
3. Profiles with matching temperature, salinity, and depth values are found. This catches duplicates where one or both profiles have incorrect location or time of observation. This test also tries to find copies of profiles where one has pressure as its vertical coordinate and the other depth.
4. A more detailed version of the third step is applied that tries to find duplicate profiles where one or both have been altered before they were archived. For example, a profile may have had some of its observations

Figure 1. Flow of processing performed on the data.
removed or its data values truncated. This step is only applied to profiles that are close to each other in space and time because of the amount of processing this requires.

[15] Application of the duplicate check results in the deletion of large numbers of profiles from the 1990s onward due to the high overlap between Argo, GTSPPP, and WOD09. Prior to 1990, when WOD09 supplies almost all the input data, the duplicate check typically removes 0–10% of the input profiles. In comparison, Gronell and Wijffels [2008] state that they find 6–10% of profiles to be duplicates within a data set.

[16] Various other processing steps are applied to the data, which are largely as described for the previous versions of the data set. If profiles from moored buoys are available more frequently than 1 day⁻¹, a daily averaging is performed. XBT profiles can be affected by errors near the surface [see e.g., Kizu and Hanawa, 2002] and Ingleby and Huddleston [2007] found that XBT data are of lower quality at depth. Therefore, bathythermograph measurements recorded at (i) 4 m or shallower and (ii) 950 m or deeper are rejected.

[17] XBT depths are, where necessary, adjusted according to Hanawa et al. [1995] or Kizu et al. [2005]—see Ingleby and Huddleston [2007] for more details. However, a change in processing compared to the previous versions of the data set is that the adjustment to XBT depths is not water temperature dependent. Time varying biases have been found to occur in XBT (and also mechanical bathythermograph (MBT)) data [Gouretski and Koltermann, 2007], affecting both the depths and temperatures. As mentioned in the introduction, there have recently been many proposals for adjusting for these but at present there is no clear evidence to choose any single method over the others. The approach taken here is to apply these adjustments post-processing to create an ensemble of adjusted data sets that represents the uncertainty in how to adjust the data. The most appropriate starting point for this is data without the water temperature-dependent adjustments.

[18] The profile data are then subjected to a series of quality control checks. These are described briefly in Table 1. Further details for many of the checks can be found in Ingleby and Huddleston [2007]. Below, we briefly describe new checks that have been added since that paper was written.

[19] Argo profiles contained in externally sourced lists of suspect data are rejected in order to reduce the possibility of bad Argo data being passed as good in the EN4 data set. The lists used are the Argo gray lists of suspect floats and a list based on the results of quality control using altimetry data [Guinehut et al., 2009].
The “bathymetry check” aims to detect profiles with latitude and longitude that are over land rather than water, which obviously indicates an error in the location information. Land topography and ocean bathymetry information is obtained from the ETOPO1 1 arc-minute data set [Amante and Eakins, 2009]. For each profile, the four ETOPO1 grid points surrounding the profile location are found. If all four are land points, the profile is rejected. Interpolation of the bathymetry data is not performed owing to the risk of steep land topography making offshore locations close to the coast appear as though they are on the land.

The “measurement depth check” aims to identify suspect depth values within a profile. Impossible depths (depth values <0 m or >11,000 m) are rejected. Each observation level in each profile is then compared to the others to check if depth is increasing throughout the profile. If this is not the case, then the level that has a depth that is most inconsistent with the others is rejected. For example, if the first observation in the profile, which should be the shallowest, has a depth that is deeper than all the others, it would be rejected. If the level to reject is ambiguous, the algorithm checks whether one has already been rejected by the profile check (which rejects observations where there are spikes, steps etc. in the profile) and, if so, leaves the other as not rejected. Otherwise, both levels are rejected. This process is repeated until all the remaining nonrejected levels have depths that increase monotonically throughout the profile.

The “waterfall check” is used to identify suspect salinity values in Argo data that have not already been detected in other tests. The algorithm attempts to mimic the process a human operator might use to manually quality control a sequence of profiles by plotting each next to the others, giving the appearance of a waterfall. Three months of profiles from each float are used. The sequence of profiles is used to establish a baseline salinity for each depth. Errors in observations are calculated using the Analysis Correction (AC) scheme [Ingleby and Huddleston, 2007]. The background error covariances are parameterized as a set of background error variances at each analysis grid point along with a function that defines how the errors covary spatially. The background error variances have been revised for this version of the data set (see section 3.2 for details). The covariance is specified using the combination of two second-order autoregressive (SOAR) functions with length scales of 300 and 400 km (between ~±4° N and the equator the former is increased exponentially to 1500 km). Vertical spreading of information is also applied and is governed by two correlation length scales, which are set to 200 and 100 m.

The persistence-based forecast of the ocean state for month $i$ is calculated as:

$$x_i^f = x_i^c + x(x_{i-1}^c - x_{i-1}^c)$$

This and following equations are written following the notation of Ide et al. [1997]. It indicates that the forecast ($x_i^f$) is generated from the climatology for the calendar month corresponding to $i$ ($x_i^c$) plus the anomalies from the previous month (analysis $x_{i-1}^c$ minus climatology $x_{i-1}^c$ for that calendar month), which are first reduced by a factor $x$. As in previous versions, $x$ is set to 0.9. This value corresponds to an e-folding time scale of 9.5 months.

The climatology is unchanged from previous versions. It represents the average over 1971–2000 and was formed from smoothed objective analyses from a previous version of the data set (EN2), which in turn used as its climatology the World Ocean Atlas 1998 [Antonov et al., 1998a, 1998b, 1998c; Boyer et al., 1998a, 1998b, 1998c]. In the first month of the data set (January 1900), the forecast is the January climatology.

It is important to note that the analyses will relax to climatology in the absence of any observations. Care must therefore be taken if using the analyses for applications such as identifying trends in temperature or salinity, because a trend may be unrealistic if analyzing periods when there were no observations. In this case, consideration should be given to alternatives such as reanalysis products that use a numerical model forecast as the background (e.g., see Xue et al. [2012], for an intercomparison of various examples of these) or objective methods that do not feature this relaxation to climatology (as done, for example, by Rayner et al. [2003], for sea surface temperature data).

### 3. Objective Analysis Uncertainty Estimates

In version 4 of the EN data set, uncertainty estimates are provided with the objective analyses for the first...
time. As already described, the EN4 objective analyses are produced using the AC scheme, which solves the OI equation iteratively. If using OI to produce the analyses directly, there would be an equation that could be used to calculate the analysis uncertainty. This is used, for example, by Ishii et al. [2003]. Unfortunately, the way that the AC scheme operates prevents the direct evaluation of this equation. In any case, the OI equation will only give accurate uncertainty estimates if provided with accurate estimates of background and observation error covariances, which are often not available.

[31] An alternative method to generate uncertainty estimates for the EN4 objective analyses was therefore developed. Since the AC scheme gives the same result as OI, OI can be used as a more practical means to test that alternative method. In the following sections, we describe the method, test it, and discuss its strengths and weaknesses. Finally, we present the result of applying the method to the EN4 data.

3.1. Method

[32] The method that has been developed to calculate uncertainty estimates for each EN4 objective analysis is an improved version of a scheme used for the Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) [Donlon et al., 2012]. It involves first calculating a special “observation influence” analysis. This has values of close to one where there is strong influence of observations on the analysis and values of zero where there were no observations. In general, the more that observations impact on the analysis the smaller the analysis error variance will be. Hence, it is possible to use the observation influence to infer analysis error variance.

[33] If we consider the situation where climatology is used as the background and assume a linear relationship between the analysis error variances and the observation influence values (see section 3.1.1 for justification of this), we can write:

\[ b_i^o = b_i^o + g \circ x_i^o \]  

(2)

[34] Where \( b_i^o \) is the analysis error variance in month \( i \) for each grid point of the analysis, \( b_i^o \) is the background error variance if climatology is the background, and \( g \) is the change in analysis error variance per unit change in the observation influence value \( (x_i^o) \). \( g \) should normally be negative since increasing the influence of observations should decrease the analysis error variance. In the limit where there are no observations, the observation influence values will be zero, hence \( b_i^o = b_i^o \). Conversely, where the observation influence values are close to unity the analysis will be almost independent of the background and the analysis error variance will be \( b_i^o = b_i^o + g \).

[35] For the EN4 objective analyses, the background is a persistence-based forecast rather than climatology and therefore its error variance will lie somewhere inbetween the two limits described above. We approximate it as:

\[ b_i^o = b_i^o + g \circ x_i^o \]  

(3)

[36] This equation states that the error variance of the persistence-based forecast for month \( i \) \( (b_i^o) \) is related to the previous month’s observation influence analysis \( (x_i^{o,1}) \) by the same coefficients \( b_i^o \) and \( g \) used in equation (3). However, the observation influence values are reduced by a factor \( \beta \) to reflect the temporal decorrelation of anomalies between months and the fact that the persistence-based forecasts are relaxed back to climatological values. Based on examination of results, \( \beta \) is set to 0.65.

[37] When using the persistence-based forecast as the background, the error variance of the resulting analysis must lie between \( b_i^o \) (if there are no observations) and \( b_i^o + g \) (the situation where the analysis is almost completely determined by observations and so it is largely irrelevant what the background is). To achieve this, the observation influence values \( x_i^o \) are replaced by the quantity:

\[ \beta x_{i-1}^o + x_i^o (1 - \beta x_{i-1}^o) \]  

(4)

[38] This represents a linear transition between \( \beta x_{i-1}^o \) when no observations are available to 1 when the analysis is completely determined by the observations. This new definition of \( x_i^o \) is used when calculating analysis error variance using equation (2). By doing this, both previous and current observations impact on the analysis error variance estimates that are calculated.

[39] To obtain the estimates of analysis error variance, we therefore only need our observation influence analyses (section 3.1.1) and estimates of \( b_i^o \) and \( g \) (see section 3.1.2 for details of how these are obtained).

3.1.1. The Observation Influence Analysis

[40] The observation influence analysis is generated by repeating the objective analysis of a month of data but with all observation values set to one, the background values to zero, and the length scales that determine the covariance in the background errors shortened (see below). No other modification of the analysis scheme is required. The aim of the observation influence analysis is to produce an analysis that has values that are linearly related to the analysis error variance at each grid point. Using the OI equations, it can be demonstrated that this is achievable for a simple case. From OI, the equation for the analysis error variance \( (B_i^o) \) is:

\[ B_i^o = (I - KH)B_i^o \]  

(5)

\[ K = B_i^o H^T [HB_i^o H^T + R_i^o]^{-1} \]  

(6)

[41] The equation depends on the background error covariance matrix, \( B_i^o \), and the matrix containing the weights applied to the observation increments when forming the analysis, \( K \). It also uses the identity matrix, \( I \), and the linearized observation operator, \( H \) (which transforms data on the analysis grid points to observation space). As well as depending on the background error covariance matrix, \( B_i^o \), this depends on the observation error covariance matrix, \( R_i^o \).

[42] Take the simple situation where we have a single observation at location \( o \) and wish to calculate the error variance at a single analysis point \( a \). We first make the assumption that the error in the background covaries between locations as \( B_{oa} = B_{oo} \exp(-\frac{d^2}{2}) \), where \( B_{oa} \) is the background error covariance between the analysis and
observation locations, $B'_{oo}$ is the background error variance at the analysis grid point, $r$ is the distance between analysis and observation locations and $l$ is a length scale. The weight applied to the observation increment, $K_i$, for a particular length scale then becomes:

$$K_i = \frac{B'_{oo}}{B_{oo} + R} = \frac{B'_{oo} \exp\left(-\left|\frac{r}{l}\right|\right)}{B_{oo} + R}$$  \hspace{1cm} (7)

\[41\] Note that $B'_{oo}$ is the background error variance at the observation location. Equation (5) then simplifies to:

$$B' = B'_{oo} - K_i B'_{oo}$$

$$= B'_{oo} \left[ \frac{B'_{oo} \exp\left(-\left|\frac{r}{l}\right|\right)}{B_{oo} + R} \right]$$

$$= B'_{oo} \left( 1 - \frac{B'_{oo} \exp\left(-\left|\frac{r}{l}\right|\right)}{B_{oo} + R} \right)$$

\[8\]

\[44\] We have found that analysis error variance is related linearly to $K_i$. This is the same quantity that is calculated by an observation influence analysis if the background error covariance length scale is halved. Therefore, in this test case we have shown that the observation influence is linearly related to analysis error variance.

\[45\] Although the length scale for the analysis error is half that of the background error in this simple case, this changes if a different background error covariance structure is assumed. Also, in reality, there will be many observations influencing each analysis grid point and the observation influence analysis will contain the sum of the weights applied to the observation increments for each grid point. We examined this more complicated situation in an idealized framework. We simulated sets of observations and used OI to calculate the analysis error variance and the observation influence analysis with various scaling factors applied to the background error covariance length scales (Figure 2) to determine a suitable scaling factor to use for the EN4 analyses. In Figure 2, plots a–c are shown plots of analysis error variance against observation influence. Each dot in the plots represents a set of simulated observations, which each consist of between 1 and 20 observations randomly positioned on a 21 $\times$ 21 grid. The total background error variance was chosen to be twice the observation error variance. The length scales that determine the background error covariance were set to be representative of those used in the EN4 objective analysis system (3 and 4 times the grid spacing) when calculating the analysis error covariance and to 1 (plot a), $\frac{1}{2}$ (plot b), and $\frac{1}{4}$ (plot c) times those values when calculating observation influence. The solid red lines are the result of fitting a line to the data. In plot a (scaling factor 1), the relationship between analysis error variance and observation influence is clearly not linear and there is a clustering of points at high values of observation influence. However, in plot b—with the scaling factor set to $\frac{1}{2}$ — an approximately linear relationship is achieved. Changing the scaling factor further results again in a non-linear relationship. This is illustrated quantitatively in plot d which shows the root-mean-square difference between the data points and the fitted line for various scaling factors, showing that the relationship is closest to linear if the scaling factor is $\frac{1}{4}$. The optimum scaling factor varies to $\frac{1}{4}$ if the stretched length scale used near the equator is adopted. We therefore adopt $\frac{1}{4}$ as the scaling factor.

\[46\] In summary, our observation influence analyses are generated by repeating an objective analysis but with all observation values set to one, all background values set to zero, and background error covariance length scales shortened by a factor $\frac{1}{4}$. These values are expected to be linearly related to the analysis error variance and can therefore be used to generate uncertainty estimates for the EN4 objective analyses.

### 3.1.2. Relating Observation Influence to Analysis Error Variance

\[47\] We have determined that we can produce an observation influence analysis that has a linear relationship with analysis error variance. We now need to obtain the coefficients of that relationship, i.e., we need to know $B'$ and $g$ from equation (3). That relationship is not of the form derived in equation (8) for the simplest case, which is shown as the dashed red line in Figure 2b. Also, this equation would only hold if we have good knowledge of the background error variance, which can often not be the case. We therefore needed an alternative method to obtain the coefficients.

\[48\] The method developed to determine these is to divide each month of observations into two groups with 2% of the observations in one group and the remainder in the other and make objective analyses from each, using climatology as the background. Locations in the analyses where the observation influence analysis values are high (between 0.95 and 1.00) are then found; these are locations where there are sufficient observations or observations close enough to the analysis point to give a high observation influence value, i.e., where the result of the analysis is determined mostly by the observations with little influence from the background. The analysis values here are used as an estimate of the true value of the field in those locations. The difference between the other analysis and this “truth” gives an estimate of the error in the former. These analysis error estimates and the corresponding values of observation influence are then used to obtain the coefficients of the linear relationship in equation (2). The proportion of observations in each group was chosen to give sufficient locations with high values of observation influence in one set of analyses with a range of observation influence values in the other.

\[49\] This method was applied to the real EN4 data. The analysis using 2% of the data is referred to as the “sparsely sampled analysis” and the analysis of the remainder, the “reference analysis.” All the analyses from the 1930s, 1940s, 1960s, 1980s, 1990s, and 2000s were used to define the relationship between observation influence and analysis error variance in each location. The remaining decades of data were reserved to provide semi-independent data for testing the resulting uncertainty estimates. Analysis values in a 15° diameter circle centered on each analysis grid point over all months were selected and searched for observation influence values between 0.95 and 1.00 to use as an
estimate of the true value of the field. These generally were found in the reference analyses but the sparsely sampled analyses were also searched for locations with high observation influence. Where these occurred, the difference between that and the other analysis and the corresponding observation influence value were saved. These difference data were binned by observation influence and the variance of the data in each bin found (as long as there were at least 10 points in the bin). If at least eight of the bins had variance estimates, a line was fitted to the data using a least absolute difference criterion. This fitting method was used to reduce the influence of outliers on the fit. If it was not possible to fit a line, the window diameter was increased by \( \frac{2}{C_{14}} \) and the process repeated up to a maximum of \( \frac{45}{C_{14}} \) diameter. If, at observation influence values of zero or one, the fitted line had negative values, it was rejected. The coefficients of these line fits defined the contents of \( b \) and \( g \) described in equation (3).

For the potential temperature analyses, we were able to associate line fits using this method for greater than 90% of the ocean area shallower than 1000 m. Below that depth the proportion of the ocean filled with line fits decreases rapidly. Therefore, levels below 1000 m depth were not investigated further and do not have uncertainty estimates calculated. Generally, we were able to associate line fits to a lower proportion of the ocean area for the salinity data than for potential temperature. At least 70% of the ocean area was filled for the salinity analyses at each depth above 1000 m. However, for consistency, it was decided to derive uncertainty estimates for the same depth

![Figure 2](image1.png)

**Figure 2.** Relationship between analysis error variance and observation influence if the scaling factor applied to the background error covariance length scales is (a) 1, (b) 1/1.7, and (c) 1/2. Solid red lines are lines fitted to the data; the dashed red line in Figure 2b is as would be obtained if a relationship as derived in equation (8) is assumed. Plot (d) shows how the linearity of the data points varies with scale factor.

![Figure 3](image2.png)

**Figure 3.** Gradients of lines relating observation influence to analysis error variance (cf. Figure 2) for the top analysis level for potential temperature on the first iteration of the method.
range as for potential temperature. Any unfilled locations in \( b' \) and \( g \) were infilled using a nearest neighbor algorithm to give spatially complete maps of the intercept and gradients of the lines fitted to the data.

[51] We note that a drawback to this scheme is that the “truth” analyses contain errors and share the same background as the analyses they are compared to. We used simulated sets of observations to investigate these issues. The method proposed to find the equation linking observation influence to analysis error variance was found to sometimes underestimate the analysis error variance if the observation influence is low, including in the situation where the background and observation error covariances are known perfectly. However, it was found that the method is more robust to inaccurate knowledge of the error covariances than the OI analysis error variance equation. In particular, it provides clear benefits in the situation where the background error variance is misspecified. As we show below, this robustness also means that the method is useful for identifying regions where there are misspecifications. While the underestimation of the analysis error variance when observation influence is low is clearly an issue, the

\[ \text{Figure 4.} \text{ Intercepts of the lines relating observation influence to analysis error variance for selected levels calculated during the first iteration of the method. These were subsequently used as the background error variances input to the analyses, subject to them having a minimum value of a quarter of the observation error variance.} \]
misspecification of background error variance is potentially a much larger problem and justifies the use of the method in the current study. If we were confident that the background and observation error covariances were well specified, it would be possible to compensate for this underestimation. However, as shown in section 3.2, this is not the case. We therefore leave the underestimation of the analysis error variance in some situations as a problem to be considered in the future.

3.2. Results

[52] An example of the gradients of the lines (corresponding to g in equation (2)) calculated from the potential temperature data in the top level of the analysis is shown in Figure 3. These indicate how the analysis error variance alters as the influence of observations on the analysis changes. Unexpectedly, we found that the gradients are positive in some locations. These tend to be in regions characterized by high variability, for example in the Gulf Stream region and the Agulhas region south of Africa. This undesirable feature implies that the analyses have larger error variances than climatology in those locations.

[53] Further investigation using simulated observations and background fields as described previously confirmed that this can occur in limited circumstances when the inputs to the analysis system are in error. Eyre and Hilton [2013] investigated the effect of misspecifying the background error variance. They found that analysis error variance can become larger than the background error variance if the latter is smaller in reality than that specified to the analysis system, resulting in the observations being fitted too closely. Therefore, in addition to providing analysis error variance estimates, our method is also useful in highlighting regions with misspecified error variances [similar, e.g., to the methods of Desroziers et al., 2005].

[54] The intercepts of the fitted lines for selected levels are shown in Figure 4. These represent the error variance when no observations are available and therefore are estimates of the background error variance. Given the above, and the fact that we find that the method for deriving the relationship between observation influence and analysis error variance is robust to misspecification of background error variance, we attempted to improve the background error variance input to the analysis system using our results. The intercepts of the line fits were used as the estimates of the background error covariance. The variance was split equally between the two length-scale components of the background error covariance. As errors in line fits could result in unrealistically low background error variance estimates, the minimum allowed total error variance was set to be a quarter of the observation error variance.

[55] The estimation of the relationships between observation influence and analysis error variance was then repeated. It was found that the occurrence of locations with high observation influence values was decreased in this new version of the analyses. Therefore, two modifications to the method were made. First, the reference analyses from the first iteration of the method were compared to the new sparsely sampled analyses. This is justified since the main requirement for the reference analyses is to generate analysis points that are strongly determined by the observations. Simulated data were generated that demonstrated that this does not impact on the effectiveness of the method. Second, the number of bins that were required to be filled in order to perform a line fit was halved.

[56] The result of this iteration of the method is shown in Figure 5. The areas with positive gradients in the northwest Pacific and Atlantic are reduced in magnitude in this iteration. However, new areas of large positive gradients have appeared, most noticeably in the Indian Ocean. These are not due to overestimation of the background error variances since this was decreased between iterations. They appear to be caused by profiles with no near-surface observations (because they did not pass the quality control). The analysis method instead spreads the observation increments from depth, where the variability is larger, affecting the analysis error variance at the surface. Again, this demonstrates that the method is useful in highlighting issues with the analysis method. However, it is beyond the scope of this study to try to resolve the problem.

[57] Despite the new areas of positive gradients, since the new background error variances should, in general, be an improvement over those used previously, we decided to continue to use these when generating the EN4 analyses. Further work is required to understand more completely the cause of the positive gradients. In particular, the other inputs to the analysis system (the background error covariance length scales and the observation error variances) need to be examined and perhaps refined.

[58] Finally, in Figure 6, we show some examples of the observation influence analyses (first column), the same observation influence analyses adjusted for the impact of using a persistence forecast as background (equation (4)) (second column), and the corresponding analysis error variance fields (third column). The first row shows the results for potential temperature for January 1965 for level 21 (373 m). The restricted spatial sampling of the oceans to that depth at that time is reflected in the observation influence plots. In the second row (showing January 1975), the situation after the introduction of XBTs is illustrated. Observational sampling has clearly improved and this is reflected in the reduction in the analysis error variance between the first two rows. For example, decreases in analysis error
variances are observed to the north/north-east of New Zealand. The decrease to the north-east results from the persistence of information from previous month of data, which demonstrates the value of the persistence-based forecasts.

In the third and fourth rows of Figure 6 is shown, the change in observation influence, persisted observation influence, and analysis error variance between July 2000 and July 2012 for salinity in the top level of the analyses. This illustrates very clearly the change in observational sampling that has been brought about by the Argo project. A very obvious difference in analysis error variance can be seen between the Argo period and pre-Argo.

3.3. Validation of Uncertainty Estimates

The uncertainty estimates have been validated using observation minus analysis statistics. This was done using the same EN4 data split into two groups as before. However, the data were analyzed in a different way to that used for deriving the uncertainty estimates. If the two different methods agree, this gives confidence in the uncertainty estimates. To give further independence between the derivation of the uncertainty estimates and the validation, the validation is applied to the data that were withheld from the derivation of the uncertainty estimates.
The validation method is to apply the EN4 processing system to the group of data containing 2% of the observations in each month over the full length of the data set. The resulting analyses are compared to the observations from the other group (the “reference observations”). To compare the observations and analyses in this way, additional uncertainty components must be accounted for—the representivity uncertainty induced by the point observations sampling processes that occur on smaller scales than the resolution of the analyses and the error variance in the reference observations. These components were estimated together in each analysis grid cell by interpolating the reference observations to the analysis levels and then calculating the variance of the differences between nearby profiles. Only profiles that were not used in deriving the uncertainty estimates were used for this.

The squared differences between the reference observations and the nearest analysis grid points were normalized by the sum of the analysis error variance at the observation location and the corresponding estimate of representivity and observation error variance. If there was more than one observation within a grid cell, one was arbitrarily chosen to calculate this difference and the rest discarded to avoid heavily observed areas dominating the statistics. The square root of the mean of the normalized differences was calculated for each month. The results are shown in Figure 7a. The normalized score is close to one throughout the record, indicating that the analysis error variance values are not unreasonable. No differences are seen between the periods where the data were used to derive uncertainty estimates (the 1930s, 1940s, 1960s, 1970s, 1990s, 2000s) and those where they were not. However, when the results for each analysis level were aggregated over the time period, it is seen that the normalized score deviations from one at the deepest levels. However, in the latter case where the analyses are more strongly influenced by observations the statistics are relatively close to one. In shallower levels, there is no dependence of normalized score on observation influence. We hypothesize that the difference at deeper levels is due to the historical lack of observations at depth, which means that ocean variability is not fully sampled. If this is the case, the problem should reduce in future years as the record of deeper observations extends in time.

4. Conclusions

We have presented a new version of the Met Office Hadley Centre EN series of data sets, EN4. The data set runs from 1900 to present and is based on subsurface ocean temperature and salinity profile data obtained from the WOD09, GTSP, Argo, and ASBO collections. All data were first compared to identify and remove duplicates. They were then subjected to a series of quality control procedures and quality flags assigned. These included three new checks introduced in this version of the data set. The first of these checked the depth at the location of each profile in the ETOP01 relief data set to confirm that it lies in the ocean. The second ensured that depths increase monotonically throughout each profile and rejected levels that appeared erroneous. The third implemented automatically a check that might be done by a human operator. It compared sequences of salinity profiles recorded by an Argo float to find errors. The profile data are output into monthly NetCDF files and are made available for research and private study from http://www.metoffice.gov.uk/hadobs.

Monthly potential temperature and salinity objective analyses were calculated from the quality controlled ocean data. These have a regular 1° horizontal grid and 42 levels in the vertical. The backgrounds for the analyses were forecasts of the ocean state generated by persisting anomalies from the previous month. These anomalies were damped so that the analyses relax to climatology in the absence of observations. A new method of calculating the uncertainty in the analyses has been described. It was found that this method is robust to misspecification of the background error variances, that it is useful for detecting issues with the
analyses, and that it can be used to improve the background error statistics used by the analysis system. However, despite updates to the background error variance, it was found that some regions of the analyses display the undesirable property of having higher error variance than climatology. More investigation is needed to determine the causes of some of these issues and to resolve the problems. Additionally, at depth there is evidence that the climatological error variance is underestimated. In these regions, caution is urged when using the analyses and uncertainty estimates. As with the profile data, the analysis files are made available from the Met Office website.

In the future, the EN4 data will be kept up to date (with approximately half a month lag) using data made available in the GTSPP and Argo collections. Bias adjustments proposed for MBT and XBT data will be applied to the data to make an ensemble of possible bias adjusted data. These can be made available to users on request and provide a mechanism to understand the sensitivity of an application to the uncertainty in how to adjust for biases in the data.

Acknowledgments. This work was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01001) and by ERA-CLIM, a collaborative project (2011-2013) funded by the European Union under the 7th Framework Programme. Thanks to Kristian Mogenes for the program to generate the combined Argo gray list. We would like to thank the three anonymous reviewers for their comments, which significantly improved the paper. Acknowledgments for data (also see references in section 2.1) now follow. Argo: data were collected and made freely available by the International Argo Program and the national programs that contribute to it. (http://www.argo.ucsd.edu, http://wo.jCOMMops.org/cgi-bin/WebObjects/Argo). The Argo Program is part of the Global Ocean Observing System. ASBO: The ASBO data were collated as part of the Arctic Synoptic Basin-wide Observations (ASBO) project by Takamasa Tsubouchi (http://www.southampton.ac.uk/oes/research/projects/arctic_synoptic_basinwide_observations_asbo.page). Beaufort Gyre experiment: The data were collected and made available by the Beaufort Gyre Exploration Program based at the Woods Hole Oceanographic Institution (http://www.whoi.edu/beaufortgyre/). North Pole Environmental Observatory (NPEO): http://psc.apl.washington.edu/northpole—NSF grants OPP-9910350, OPP-0352754, and ARC-0634226. Freshwater switchyard of the Arctic project: M. Steele, W. Smethie, and P. Cabanes, CTD data from the Arctic Switchyard Project, Office of Polar Programs, National Science Foundation: http://psc.apl.washington.edu/switchyard/. Nansen and Amundsen Basins Observational System (NABOS), Canadian Basin Observational System (CABOS): http://nabo-s.iarc.uaf.edu/.

References


