CMUG Deliverable
Number: D3.3
Due date: March 2013
Submission date: 16 April 2013
Version: 0.7

Climate Modelling User Group

Deliverable 3.3

Technical note on Use of Uncertainties in Models and Reanalyses

Centres providing input: MOHC, MPI-M, ECMWF, MétéoFrance

Version nr.	Date	Status
0.1	30 Oct 2012	Initial Template RS
0.2	30 Jan 2013	ECMWF input
0.3	25 Feb 2013	MetO input
0.4	7 March 2013	Combined inputs
0.5	15 Mar 2013	Inputs from others
0.6	8 April 2013	Updated Climate section from Mark + Serge's inputs
0.7	16 April 2013	Final changes made with Mark's input

CMUG Deliverable

Number:	D3.3
Due date:	March 2013
Submission date:	16 April 2013
Version:	0.7

Deliverable 3.3

Technical note

Use of Uncertainties in Models and Reanalyses

Table of Contents

Use of Uncertainties in Models and Reanalyses

1. Purpose and scope of the Technical note

One of the main requirements for the new CCI climate data records (CDRs) is to include the associated uncertainties along with the measured variables for each measurement. This has not been available to date in most satellite climate data records. This document describes how these uncertainties will be used by the modelling and reanalysis communities in order to help guide the data providers who are providing the uncertainties in their datasets.

2. Treatment of uncertainties in climate modelling

2.1 Model evaluation and development

Model evaluation is perhaps the most obvious example of using satellite-derived observations in climate modelling. It is, however, important to remember that this is not simply a question of comparing a model with the observations but also of using the information gained from such comparisons to improve the model, for example by developing better parameterizations of physical processes. This has important implications, both for the way the data is used by modellers but also for the types of product, including uncertainty estimates,that they require from the data providers. Whilst it is probably fair to say that the use of observational uncertainties by climate modellers is still in its infancy, it also true that there is now an increasing demand for this information and that this is likely to grow very quickly in the near future. This is particularly the case when the difference – as measured by some suitable metric – between the model outputs and the observations is of the same order of magnitude as the differences between the available observational products.

(a) The basic problem

The three most basic questions we seek to answer when comparing our models to observations are: (i) How good is our model? (ii) Is our model improving or getting worse as a result of the changes we have made to it? (iii) What is the level of confidence of the reference observational product? This last one is particularly important when the model performance is considered to be reasonable.

In our model we wish to simulate a particular physical quantity, $\mathbf{X_{MOD}}$, and we would like to know how close this is to reality, as defined by the best available observations, $\mathbf{X_{OBS}}$. We want to avoid overconfidence in our simulation, i.e. inferring from $\mathbf{X_{MOD} = X_{OBS}}$ that our model is performing much better than it actually is but we also do not want to reject or penalise the model unfairly, i.e. inferring from $\mathbf{X_{MOD} \neq X_{OBS}}$ that our model is worse than it is.

We would thus like to have some estimate of the observational uncertainty, $\mathbf{X_{OBS} \pm \Delta X_{OBS}}$. In the simplest case we can then determine if our model simulation is plausible or credible, in the sense that it lies within the range of the observational uncertainty, $\mathbf{\Delta X_{OBS}}$. In addition, if we can also estimate our model uncertainty, $\mathbf{X_{MOD}} \pm \mathbf{\Delta X_{MOD}}$, for example by running an ensemble of simulations, we then have an even better basis for assessing the credibility of the model. A further level of sophistication can also be added to our evaluation if we have multiple observational data sets for a given quantity, $\mathbf{X_{OBS1}} \pm \mathbf{\Delta X_{OBS1}}$, $\mathbf{X_{OBS2}} \pm \mathbf{\Delta X_{OBS2}}$, etc

In general modellers will probably assume that $\mathbf{\Delta X_{OBS}}$ is the data provider's best estimate of the observational uncertainty, i.e. that all of the relevant contributions to $\mathbf{\Delta X_{OBS}}$ such as measurement errors, calibration errors, spatial and temporal sampling, structural uncertainty, etc, have been accounted for. It is therefore very important to know if this is *not* in fact the case, so that a meaningful comparison between the model and observation can actually be made. A good indication of the type of information and level of detail required by modellers is the documentation of requirements for the Obs4MIPS activity which is part of CMIP5 (http://obs4mips.llnl.gov:8080/wiki/FrontPage).

(b) The wider context

The above remarks can be generalised to the broader, multi-model context. We may then seek to answer questions such as:

- Are climate models improving with time, e.g. between CMIP3 and CMIP5 or as reported in successive IPCC reports?
- Is it possible to say that some models are demonstrably better than others, i.e. in a clearly-defined, objective, and quantified sense? Is the ranking of models dependent on the chosen metric?
- Are some particular physical quantities more robustly simulated in models than others? Are there consistent strengths or weaknesses across the range of available, independent models?
- Does a better comparison with currently-available observations imply more reliable projections?
- Can we weight models, in terms of their skill, based on our comparisons with the observations?

The above questions have stimulated a growing interest in the development of objective methods to assess model performance and the construction of reliable metrics to do this (e.g. Gleckler et al., 2008). This activity includes the formation of WGNE/WGCM metrics panel (http://www-metrics-panel.llnl.gov/wiki). They have also motivated an IPCC expert meeting on Assessing and Combining Multi Model Climate Projections in preparation for the publication of the Fifth Assessment Report (Knutti et al., 2010).

(c) Model evaluation in the absence of observational uncertainties

It is often the case, however, that reliable observational uncertainties are not available. What are the options for climate modellers if this is so?

The simplest approach is to treat all data sets as being equally plausible. Here we might be said to be applying a “principle of indifference” and we do this because we have no evidence to suggest that doing otherwise is any more valid. Following on from this we might then choose to define our "observational uncertainty" as, for example, the range of a particular parameter spanned by the available data sets. This was the approach chosen by Gleckler et al. (2008) when they evaluated the simulation of twentieth century climate by the CMIP3 models. To give some indication of the effects of observational uncertainty, for most fields they provided a comparison of the model simulations with two different reference data sets.

Alternatively, we might assume that technological or scientific developments necessarily lead to improved data sets, i.e. the most recent data sets are always better than their predecessors. This could be because they include enhanced information content (e.g. more channels, active vs. passive sensors to detect rainfall, etc); or improved retrieval algorithms and data processing methods; or more up-to-date technology (improved sensors/instrumentation). Note that we almost always assume that a new version of an existing data set will be an improvement on the last.

It is also the case that we sometimes make a subjective assessment of observational data sets based upon our prior experience or expertise. For example, we might consider that the observed values of a quantity that we are presented with are so far from our theoretical expectations that the most likely explanation is that they are in error.

Finally, we may decide to make an approximate estimate of the observational uncertainty. This is often based on input from the data providers themselves and is usually quite conservative, e.g. ± 100%.

Climate modellers do of course recognize that all of these solutions are far from ideal: it is thus relatively straightforward to make the case for the determination and provision of reliable observational uncertainties by data providers.

(d) Examples

Figure 1 illustrates the potential dilemma when no observational uncertainties are available. According to the first data set (ISCCP) the newer version of the model is an improvement compared to the older version, but according to the second data set (CERES) the overall quality of the simulation deteriorates. There are also considerable regional differences between the comparisons depending on which of the two data sets is used for the evaluation, e.g. over the Atlantic Ocean.

Figure 2 shows how this relatively simple approach can be made more quantitative and applied to a whole set of related parameters, in this case quantities of relevance to the hydrological cycle. In this example multiple observational data sets are used for evaluating each parameter of interest and the observational uncertainty has been estimated assuming that each of these data sets is equally valid.

Figure 3 is taken from Jiang et al. (2012), which describes the evaluation of the vertical distributions of atmospheric water vapour and clouds in the new generation of climate models submitted to CMIP5. As water vapour is strongly coupled with the cloud liquid/ice water content (LWC/IWC) it is informative to analyse the models' simulations of these quantities simultaneously. We can see that there is a much larger spread in model performance across the ensemble in the upper troposphere than in the middle and lower troposphere. Moreover, while the model simulations tend to lie within the estimated ranges of observational uncertainty in the lower troposphere they clearly do not higher up. Taking these results as a starting point we can then investigate how strongly these two model errors are coupled, and how much the models' physical parameterization schemes (e.g. cloud microphysics or convection) contribute to the biases we see.

Figure 4 illustrates the problem that arises when the differences between model outputs and observations are of the same order of magnitude as the differences between two observational products. In this case it is difficult to draw any firm conclusions on the model’s performance without complementary information on the uncertainties of observational products. Moreover, information on the strengths and weaknesses of the different products is specifically required over the spatial domains and time periods being used for the model evaluation (in this case the inter-tropical region in the lower stratosphere or at the time of ozone hole deepening).

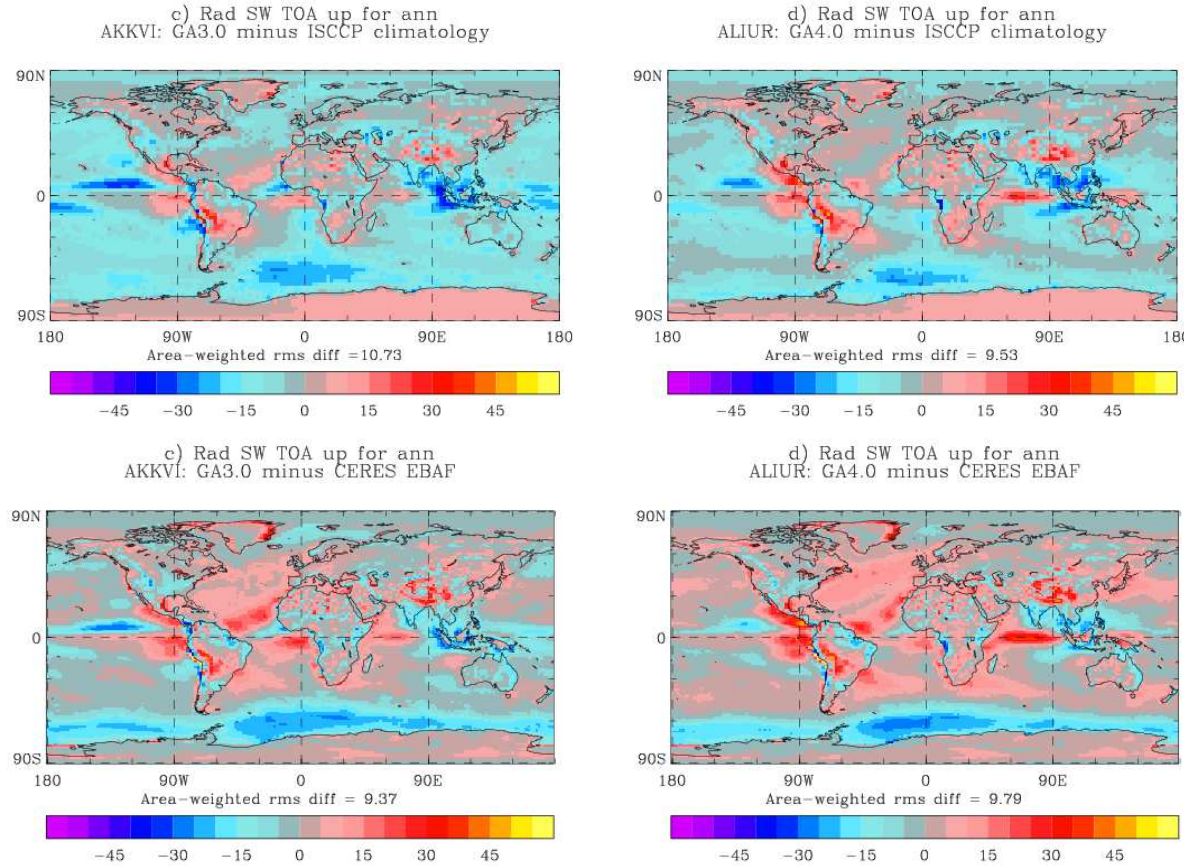

Figure 1: The annual mean reflected shortwave radiation at the top of the atmosphere simulated by two versions of the Met Office climate model compared to the ISCCP-FD (upper) and CERES-EBAF (lower) observational products. Values shown are model minus observations in Wm-2. The two versions of the model chosen (earlier version on the left, later version on the right panels) are arbitrary and are taken from current developmental work in progress.

Although the above examples focus primarily on atmospheric quantities it is important to recognize that similar efforts are being pursued across the climate modelling community. For example, Luo et al. (2012) propose a framework for assessing ("benchmarking") land models.

This framework consists of four components:

1) Identification of the aspects of models to be evaluated.
2) A set of benchmarks as standardized references to test models.
3) A scoring system (metrics) to measure and compare model performance skills.
4) Evaluation of model strengths and deficiencies for model improvement.

Clearly this framework could be applied more widely than land models to the full range of model evaluation activity, encompassing atmospheric, oceanic and cryospheric processes and parameters.

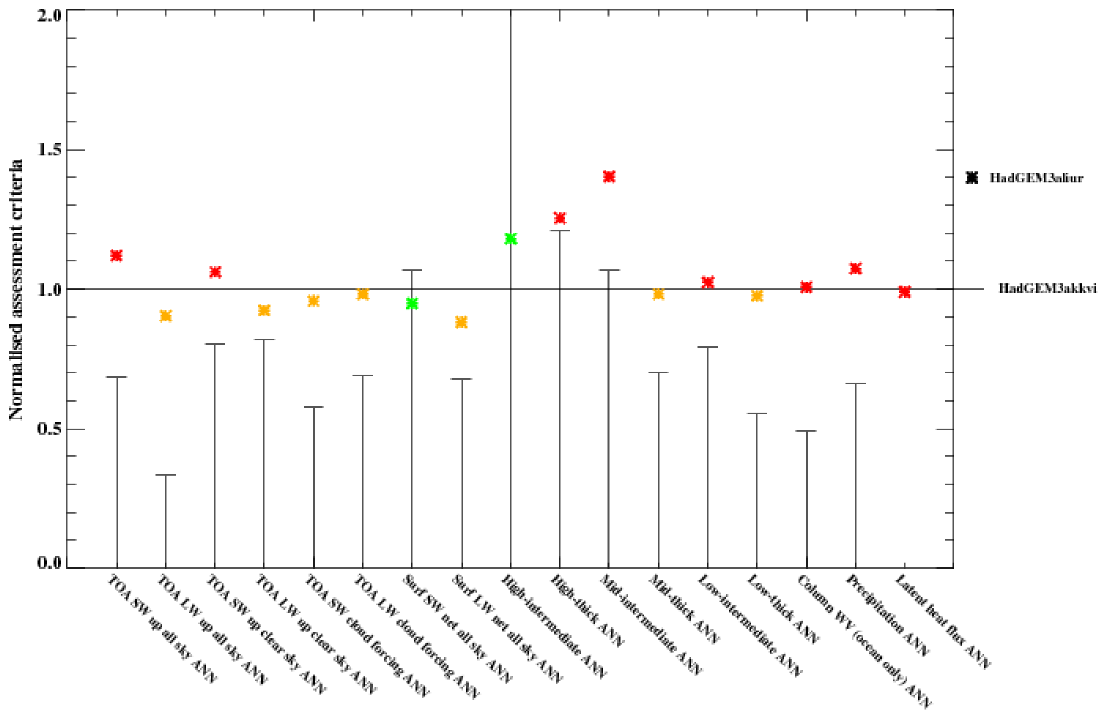

Figure 2: Normalized assessment criteria (ratios of mean field root mean square errors) for a range of radiation, cloud and hydrological cycle variables for the two simulations shown in Fig. 1. These errors are calculated relative to a reference set of observations for each of the variables shown. The whisker bars are observational uncertainty, which is calculated by comparing these with alternative data sets. The colour coding indicates whether the performance of the new version of the model has improved, deteriorated, or remains unchanged compared to its predecessor. For a complete description of this methodology see Walters et al. (2011).

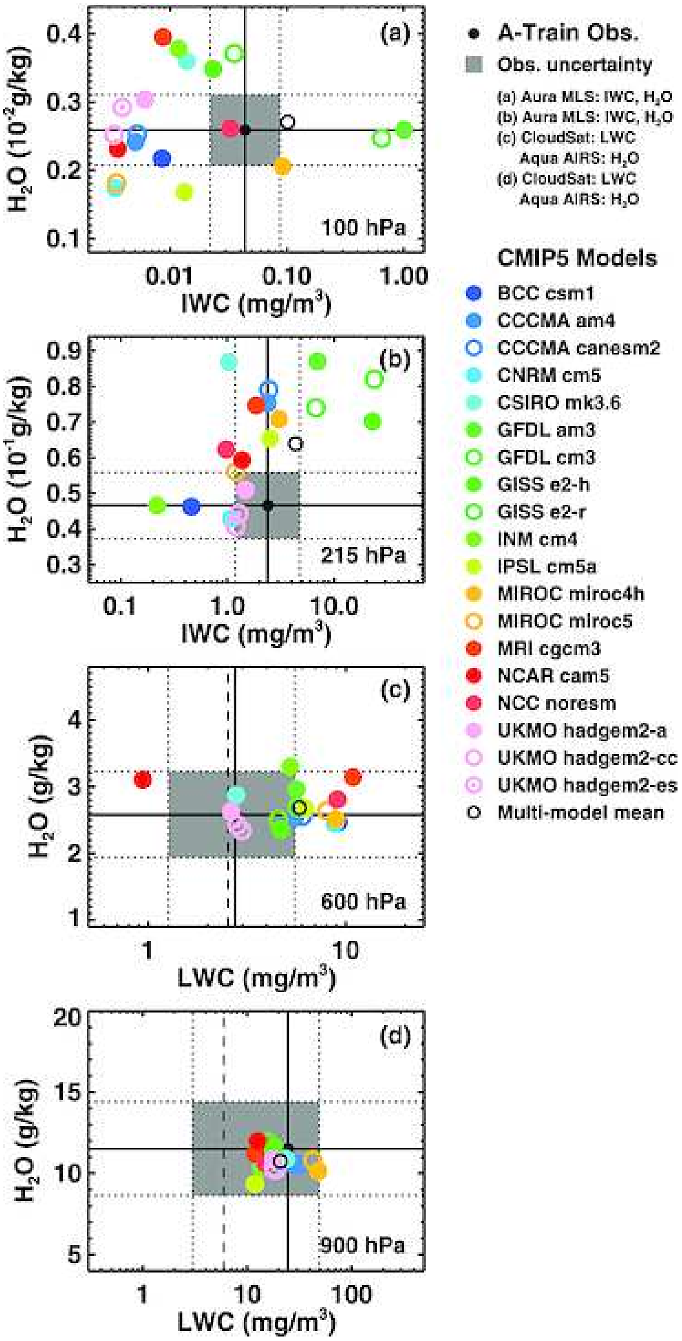

Figure 3: Scatter plots of tropical mean (30°N–30°S) oceanic multi-year mean atmospheric water vapour (H2O) versus cloud ice water content IWC at (a) 100 and (b) 215 hPa, and H2O versus cloud liquid water content (LWC) at (c) 600 and (d) 900 hPa. Black dots show the A-Train observed values and the grey area indicates the observational uncertainties. Coloured dots/circle are the values from the CMIP5 models and the black open-circles represent the multi-model means. [From Jiang et al., 2012] quality of the weighted ensemble mean. They propose and apply some relatively simple statistics to demonstrate this. However, they also suggest that uncertainties arising from the use of different datasets could be naturally included in probabilistic projections of regional climate change within a Bayesian statistical framework.

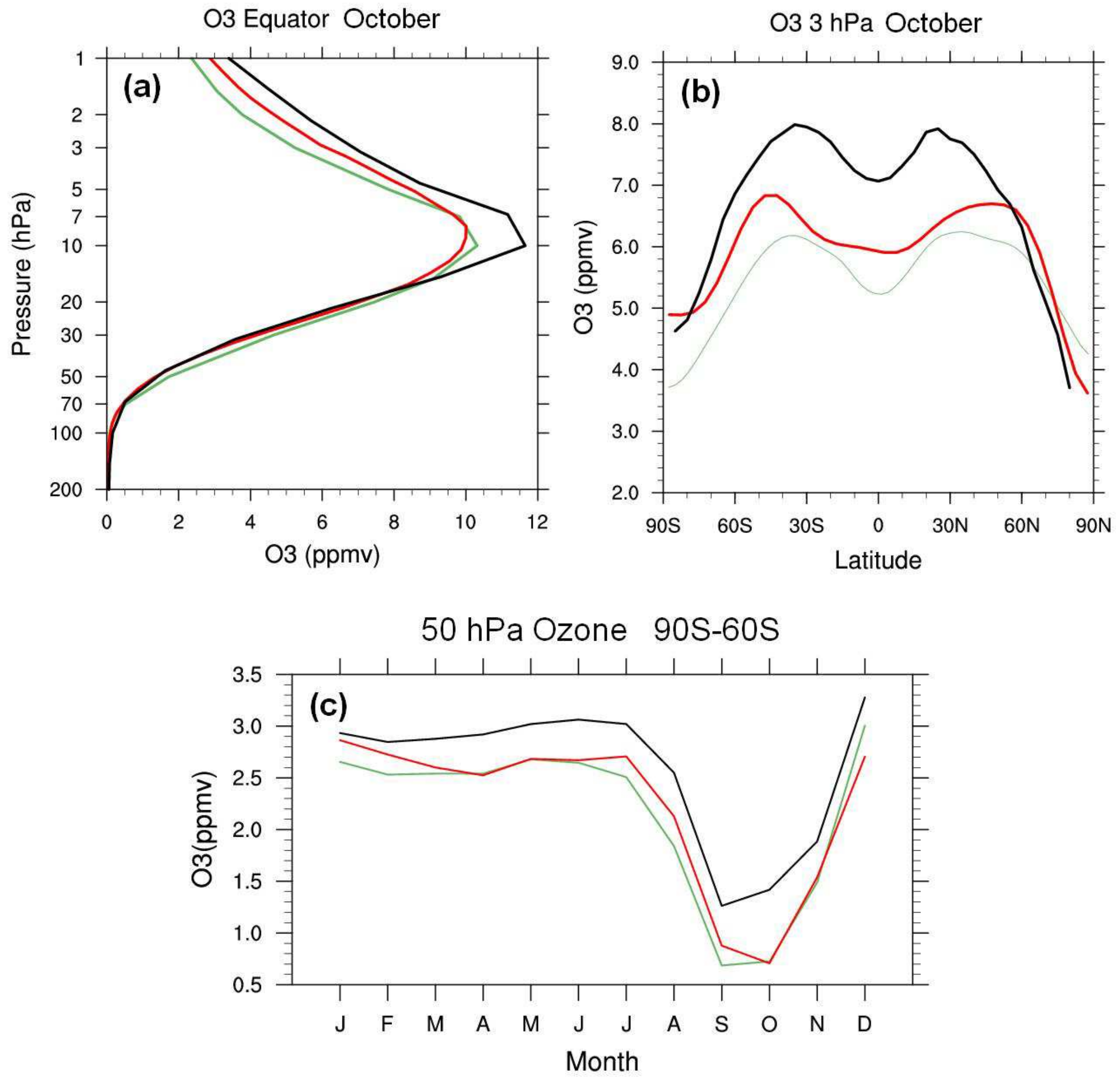

Figure 4: Diagnoses of comparisons between the CCI O3 preliminary L3 product from Sciamachy (black lines), the CCMVal2 O3 observational product (red lines) and the CNRM-CCM O3 simulated in a nudging experiment towards temperatures and winds of ERA-Interim reanalysis (green lines). Shown are (a) the mean October 2006 vertical profiles at the equator, (b) zonal averages of the mean October 2006 at 3 hPa, and (c) the monthly means of O3 at 50hPa for 2006. The CCMVal2 O3 observational product was developed by Greg Bodeker (Bodeker Scientific) and Birgit Hassler (NOAA) and combines measurements made by satellite-based instruments, ozonesondes, aircraft-based instruments and lidars (version 1.1.0.6 of the data base). The CNRM-CCM is an atmospheric climate model including an on-line comprehensive representation of chemical processes in the stratosphere (Michou et al., 2011).

The global observing system – including the ESA CCI – continues to offer an increasing number of data sets which potentially could become candidates for benchmarking climate models. It must be recognized, however, that many of these data sets have limited information content and are not always suitable for our purposes. This means that we need to thoroughly assess the range of available data sets in order to develop reliable benchmarks against which model performance can then be both effectively and objectively evaluated.

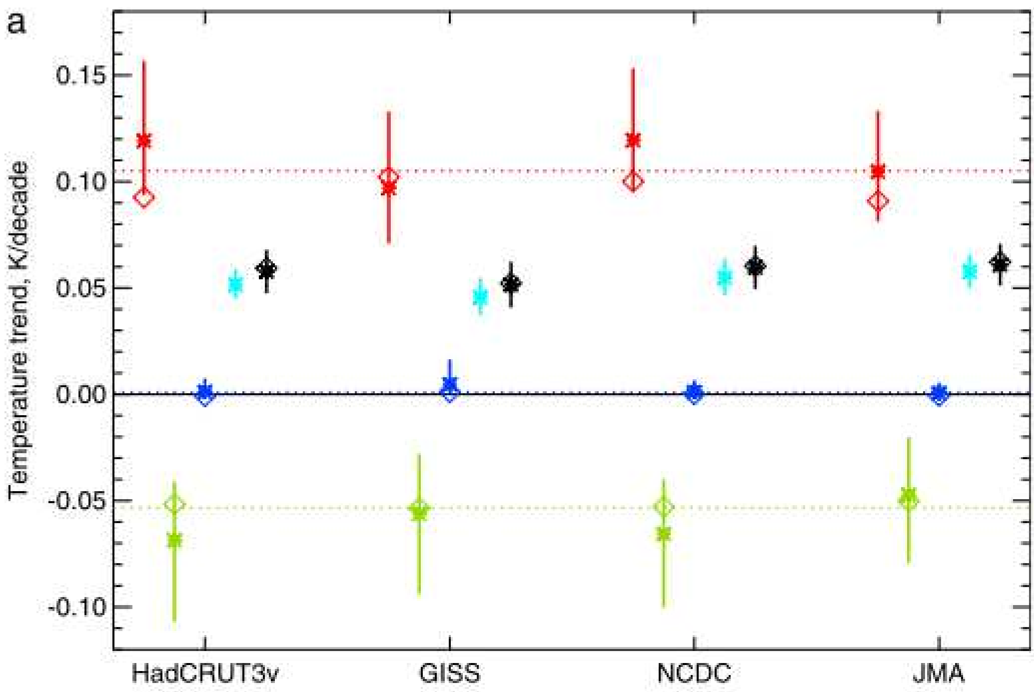

Figure 5: Attributed global mean temperature trends for 1900–1999 contributing to each of the four observational datasets indicated on the x-axis. Estimates of the attributed trends (represented with the asterisks) of greenhouse gases, other anthropogenic, and natural factors, together with their sum. The 5–95% limits of the attributed trends are indicated by the vertical lines. Trends in the observations are also shown (black asterisk symbol) with the 5–95% uncertainty range representing an estimate of internal climate variability deduced from the climate model control simulation. The diamonds show the trends calculated after masking by the observational coverage. [From Jones and Stott, 2011]

2.2 Detection and attribution of climate change

Although numerous studies have clearly demonstrated that much of the recent warming in global near-surface temperatures can be attributed to increases in anthropogenic greenhouse gases (IDAG, 2005; Hegerl et al., 2007), very little has been done to assess the sensitivity of these findings to the choice of observational data sets, and thus to the observational uncertainty. Errors in these measurements arise due to a number of factors, e.g. grid-box sampling, instrumental biases, and changes to the global coverage. Jones and Stott (2011) is the first exploration of the full impact of observational uncertainty on attribution. They performed a standard detection and attribution analysis using four independently-processed near-surface temperature data sets. The main results are summarised in Figure 5.

Their principal finding is that the ‘headline’ IPCC conclusion on attribution is indeed robust to observational uncertainty. Clearly this type of study needs to be both developed and expanded to parameters other than the global mean surface temperature (Vautard and Yiou, 2012) and the precise details of such attribution results may well depend on the observational data set, e.g. for smaller sub-regions. A further recommendation of Jones and Stott (2011) –

itself following Thorne et al. (2010) – is that structural uncertainty in data sets, as determined by differences between different reconstructions, should be accounted for along with other types of uncertainty when making comparisons with models. In particular, as the spatial pattern of the observed trends is often the focus for detection and attribution studies, the homogenization of the time series is a key component of the data reconstruction process that needs to be considered with specific attention to the uncertainty analysis.

2.3 Constraining climate projections

In a recent study Gómez-Navarro et al. (2012) examine the degree to which the evaluation and ranking of an ensemble of regional climate models – based on their ability to reproduce observed climatologies of surface temperature and rainfall – is sensitive to the choice of reference observational dataset. They demonstrate that, even in areas covered by a dense observational network (Spain in their case); uncertainties in the observations are comparable to those in state-of-the-art regional climate models. The clear implication of this is that model evaluation needs to account for the observational uncertainties. Furthermore, they point out that weighting models according to how well they perform against a single observational dataset, without acknowledging the observational uncertainties, might actually reduce the quality of the weighted ensemble mean. They propose and apply some relatively simple statistics to demonstrate this. However, they also suggest that uncertainties arising from the use of different datasets could be naturally included in probabilistic projections of regional climate change within a Bayesian statistical framework.

In fact, work in this direction has already commenced (Sexton et al. 2012; Sexton and Murphy 2012). Sexton et al. (2012) outline a method for producing probabilistic projections of climate change at both global and regional scales. In particular they consider the response to increasing atmospheric CO_2 on both global, annual-mean surface temperature and regional climate change in summer and winter temperature and precipitation over Northern Europe.

Their approach combines information from a perturbed physics ensemble (of a single climate model), a multi-model ensemble (CMIP3/IPCC AR4), and observations and is based on a multivariate Bayesian framework: this enables the prediction of a joint probability distribution for several variables constrained by multiple observational metrics. The use of multiple metrics is important because, unlike using a single metric, it reduces the risk of rewarding a model which scores well fortuitously (for example due compensation between large errors of opposite sign) rather than because it is providing a realistic simulation of the observed quantity. Here then, model skill can be defined as the likelihood of a model given the observations. Clearly, this approach also requires quantitative estimates of uncertainties in the observations used to constrain the climate projections. These are obtained using two or three alternative observational data sets for each quantity and then generating 100 “pseudo-observations” by adding random linear combinations of the different data sources.

An example of the method – applied to the global-mean surface temperature response to doubling CO_2 – is shown in Figure 6. This shows that neglecting the observational uncertainty has some important negative consequences: it narrows the distribution of the surface temperature response; it reduces the effective sample size; and it unfairly excludes some models.

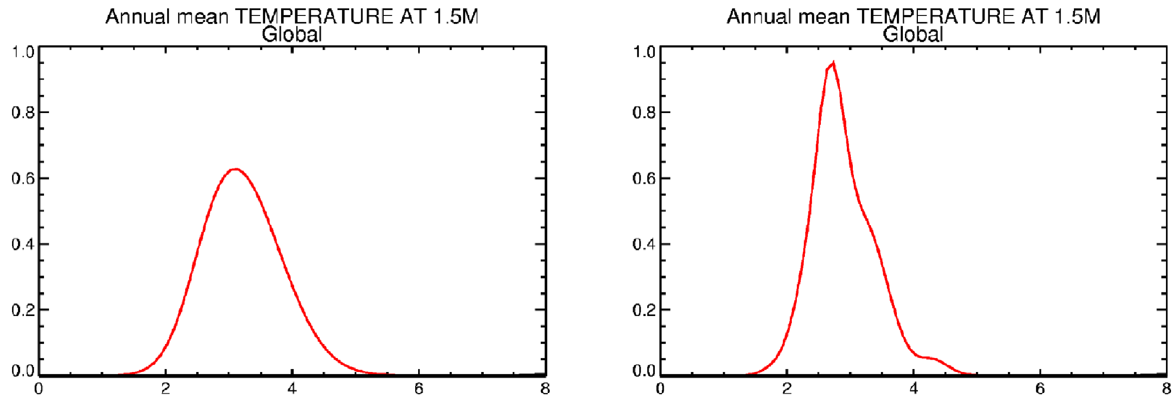

Figure 6: Probability distribution function (pdf) of the global mean temperature response (K) to doubling atmospheric CO2 derived from a large ensemble of climate model simulations and applying the Bayesian approach described in the text. The two pdfs show the impact of including (left) and excluding (right) the observational uncertainty [From Sexton et al., 2012]

Another type of approach is suggested by studies such as Hall and Qu (2007), who try to constrain the snow albedo feedback in climate models using observations of the present-day seasonal cycle. Following on from this Fernandes et al. (2009) attempted to quantify the snow albedo feedback using satellite observations. Hall and Qu (2007) suggest that this approach could possibly be extended to other processes, such as the sea-ice feedback in the Arctic, although this has yet to be demonstrated. Note however that making a link between the performance of models on present or past climate conditions and uncertainties in model projections remains challenging (Knutti et al., 2010).

2.4 Reconciling observations and models

It is now generally recognised that comparisons between climate models and observational data require as much consistency as possible between the simulated and observed quantities – e.g. effective use of the information content of the measurements, temporal and spatial sampling – in order to draw meaningful conclusions regarding model performance. This contrasts with the early work in this field which often consisted of placing satellite-derived quantities and their model ‘equivalents’ side-by-side. In a sense climate model evaluation is moving more towards the more rigorous match-ups between models and observations used in data assimilation as part of numerical weather prediction systems.

We can attempt to achieve this consistency in two ways: (i) producing observational datasets which match model-simulated parameters and diagnostics (the “satellite-to-model” approach); and (ii) simulating in the model what is actually observed by the satellite sensor (the “model-to-satellite” approach). Here we give examples of both.

(a) Top-of atmosphere radiation budget: CERES EBAF

The standard NASA CERES top-of-atmosphere (TOA), global-mean radiative fluxes have a positive net imbalance of around 6.5 Wm^{-2}. This is much larger than the best estimate of 0.85

Wm^{-2} based on observations of ocean heat content data and model simulations, with the major sources of uncertainty being related to the CERES instrument absolute calibration.

This makes using the data for coupled climate model evaluation problematic, as these models need to be close to balance for present-day conditions to be used reliably. It also has implications for attempts to estimate the Earth's global energy budget and for inferring meridional heat transports.

To alleviate these difficulties the CERES Energy balance and filled (EBAF) TOA product (Loeb et al., 2009) was designed specifically for climate modellers needing a net imbalance constrained to the ocean heat storage term. The CERES team undertook a detailed uncertainty analysis and determined that the CERES instrument calibration was the largest uncertainty, with other aspects of the processing chain making smaller, but not negligible, contributions. They then derived and applied a set of adjustments to the various terms to produce a radiation budget data set that was consistent with the requirements of the climate modelling community. The CERES EBAF data set now extends from 2000 to the present and is considered the standard reference for climate model evaluation and other studies.

(b) The GCM Oriented CALIPSO Cloud Product (GOCCP)

CALIPSO combines an active lidar instrument with passive infrared and visible imagers to examine the vertical structure and properties of clouds and aerosols. The need for a specific model-oriented CALIPSO data set arises because the interpretation of the lidar backscatter ratio in terms of cloud products (e.g. cloud fraction) requires a set of criteria that depends on the vertical resolution at which the lidar scattering ratio is measured or computed.

In order to allow consistent comparisons between models and the CALIPSO data, the GCM Oriented CALIPSO Cloud Product (GOCCP) data set has been produced (Chepfer et al., 2010). This is an entirely new product that has been derived from the original CALIPSO Level-1 data. An example of its application to climate model evaluation is shown in Figure 7. In addition, the GOCCP data set is consistent with the CALIPSO simulator outputs derived from models using the satellite simulator COSP (see below).

(c) Forward modelling and satellite simulators

As noted above, the 'traditional' approach to model evaluation assumes that model-simulated and satellite-retrieved versions of physical quantities are essentially equivalent. This is of course rarely the case in practise. This lack of consistency has stimulated the development of satellite simulators which aim to avoid the inherent ambiguities between model and satellite-derived parameters and allow us to make full use of the information content of measurements. The greatest amount of progress in this direction has been made by the cloud modelling community and has led to the development of the CFMIP observational simulator package (COSP), which has already been described in previous deliverables (CMUG, 2011; CMUG, 2012). An example of using COSP, which compares the Met Office model to five different sensors, is shown in Figure 8.

Using satellite simulators obviously means that comparisons between models and observations – including the use of uncertainties – then take place in the space of the simulated radiances, radar reflectivities, lidar backscatter, etc, rather than derived quantities such as cloud fraction or cloud top altitude. This clearly has potentially important implications for the types of data sets required by the climate modelling community.

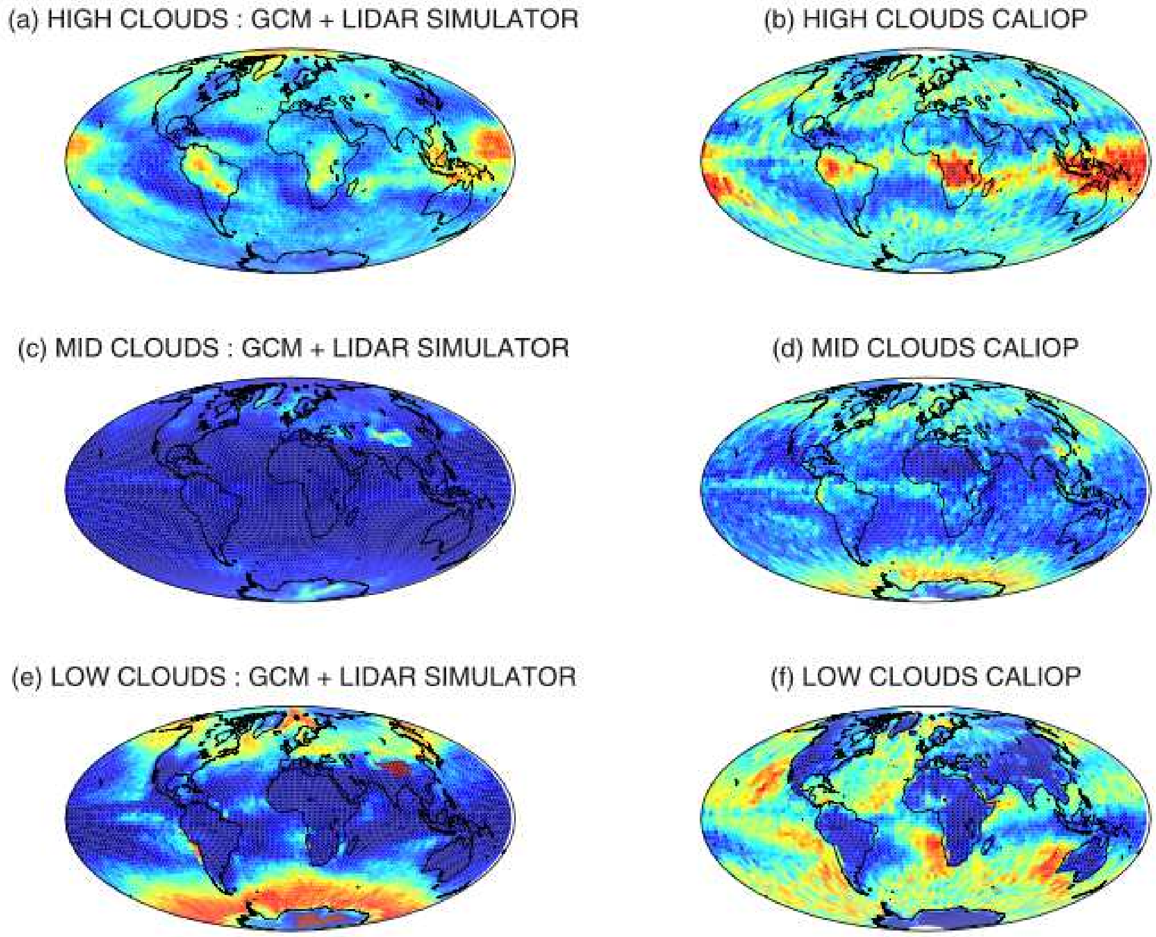

Figure 7: Comparison of the IPSL climate model (left) and the GCM Oriented CALIPSO Cloud Product (right) for high-level, middle-level and low-level clouds.[From Chepfer et al., 2010]

2.5 High resolution modelling

An important current development in climate modelling is the move to much higher grid resolutions, both horizontal and vertical. Some centres have already submitted versions of their models at horizontal resolutions as high as 20km to the CMIP5 archive and this tendency will undoubtedly increase over the coming years and for CMIP6. The development of the CORDEX (COordinated Regional climate Downscaling Experiment) international project is also fostering the development of regional climate modelling at horizontal resolutions that are commonly of order 50km but including several intercomparison exercises at higher resolution

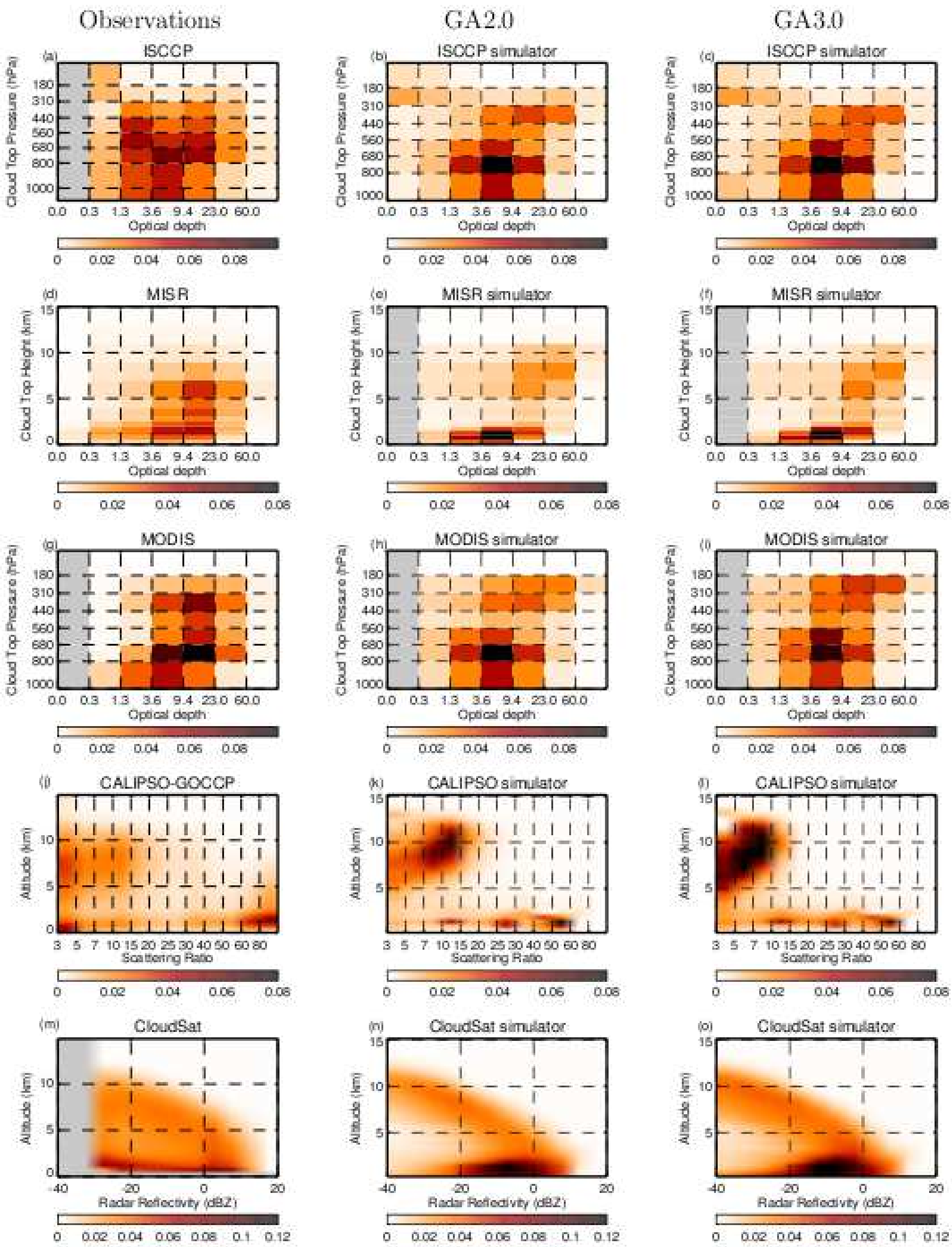

Figure 8. Observational and COSP diagnostics averaged over the Southern Ocean (40º-70ºS) for the DJF season. Shown are (left) the observational results for five sensors and the equivalent COSP diagnostics from two versions of the Met Office model referred to as (middle) GA2.0 and (right) GA3.0. The histograms show frequency of occurrence in each bin (0–1). The ISCCP, MODIS, and MISR histograms are normalized by the total population of the histogram (i.e., the sum of all the bins gives the total cloud fraction), whereas the CloudSat and CALIPSO histograms are normalized level-by-level. [From Bodas-Salcedo et al., 2012]

up to 12km as in MedCORDEX. Figure 9 shows an illustrative example of the new Met Office global climate model run at a range of resolutions from 135km down to 12km. Clearly there is increased regional detail as the resolution increases but is this detail plausible and how can we verify this?

If we are to both examine the potential benefits of increased resolution and use such models for projections we clearly need to evaluate their performance using reliable observational data sets with appropriate uncertainties. A key point is that simply averaging or interpolating to different grids is unlikely to be adequate and we will probably require versions of data sets at multiple resolutions each with their own uncertainty estimates, taking into account sub-grid variability, for example.

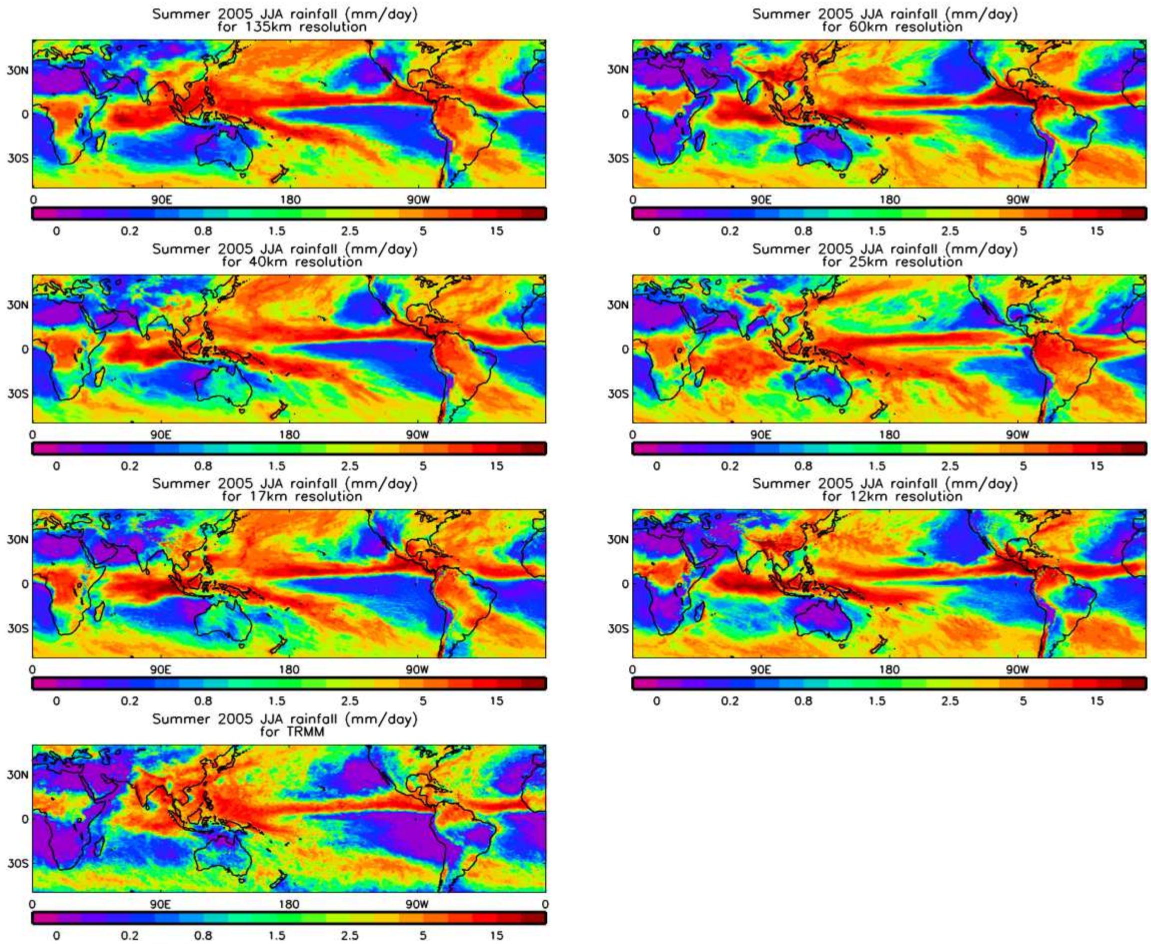

Figure 9: Simulation of seasonal mean rainfall for JJA 2005 using the Met Office climate model run at a range of horizontal resolutions: 135 km, 60 km, 40 km, 25 km, 17 km, and 12 km. Also shown are TRMM observations of rainfall (bottom left) for the same season. Note that this is a version of the model which is currently under development and the comparison is shown for illustration only. [Courtesy Malcolm Roberts, Met Office Hadley Centre].

2.6 Summary

A central aim of the CCI is to address the requirements of the climate modeling community and to increase the use of ESA-derived data sets by climate modelers generally. In addition, a key deliverable for all of the projects is to produce reliable uncertainty estimates for each of the variables they derive. The above discussion has aimed to highlight the following points:

- Climate modelling requirements are multiple and evolving, often quite rapidly, and will certainly continue to evolve over the lifetime of the CCI.
- The previous "one size fits all" approach, i.e. simply matching monthly mean Level-3 products to presumed model equivalents, is now only part of what model evaluation entails.
- Multiple data sets, e.g. at different temporal/spatial resolutions, with appropriate uncertainties, may be required for certain applications.
- More systematic intercomparisons of different data sets within the context of specific applications are required to measure the strengths/weaknesses of the different reconstructions.
- In some circumstances data sets produced specifically to address modelling applications may be needed.
- Use of the forward modelling or simulator approach is increasing: observation groups should be encouraged to consider the development of appropriate simulator modules in collaboration with modelling community.
- It is important to consider structural uncertainty when deriving observational data sets (e.g. multiple realisations of the retrieval algorithms with different settings/choices), and this is considered essential for applications such as detection and attribution of climate trends, in particular the contribution of the uncertainties in the homogenization processes.

The principal uses of observational data in climate modelling discussed here are:

- Model development and evaluation, including improving physical parameterisations.
- Development of reliable metrics for multi-model inter-comparisons.
- Detection and attribution of climate change and trends.
- Testing the benefit and utility of increasing model resolution at both global and regional scales.

In addition we can also include seasonal-to-decadal prediction and the generation of model ensembles. In the former, this generally involves using the observations as part of the process of generating the initial analyses, so that the inclusion of uncertainties is handled in a similar manner to numerical weather prediction. The latter is basically self-explanatory and includes, for example, producing multiple model simulations using boundary conditions (SST, land cover, etc) derived either from different data sets or from several realisations of the same data set.

The key message is that we potentially need to consider all of these climate modelling applications when constructing observational data sets and determining the associated uncertainties. We also stress the requirement to thoroughly assess both available and future observational data sets in order to determine their suitability for assessing the performance of climate models.

3. Treatment of uncertainties in reanalyses

This chapter describes some key issues and challenges in the production of high-quality climate reanalyses. The methodologies and tools that can be used for data quality assessment are discussed with a focus on homogeneity (e.g. related to changes in the observing system, including calibration issues) and uncertainty (e.g. accuracy and precision related to measurement error, processing methods, etc.).

Reanalysis makes use of advanced statistical methods to assimilate observations from multiple sources into a state-of-the-art atmospheric forecast model. This generates a physically and dynamically coherent global dataset, typically extending over several decades and containing estimates of many essential climate variables (ECVs). The use of a model-based data assimilation technique ensures that the ECV estimates are consistent with observations, but also with the laws of physics, and therefore with each other. Since it is produced with a single version of a data assimilation system, a reanalysis is more suitable for climate monitoring and climate research generally than archived weather analyses from operational forecasting systems.

3.1 Uncertainties in reanalysis

Users often view reanalysis data as a “true representation of the atmosphere according to observations" or simply “observations”. In fact, reanalysis combines inaccurate and incomplete observations with imperfect models, using methods and procedures that are technically and scientifically complex. A realistic analysis (as in true to nature) is possible only if the degrees of freedom in the modelling system can be adequately constrained by available observations. The actual impact of any given set of observations on the reanalysis depends on many factors, including limitations of the forecast model used for data assimilation, the choice of analysis method, and the description of error characteristics of the data. It also depends on various choices and assumptions made in the technical implementation of the reanalysis system (e.g. production in parallel streams).

The use of atmospheric reanalysis data for climate change assessment has been, and still is, somewhat controversial (e.g. see Thorne and Vose 2010, and comments by Dee et al 2011a). This is due to well-known difficulties with the representation of low-frequency variability in reanalysis. Early generations of reanalyses, as well as some recent ones, show spurious shifts and other artefacts that can be identified with changes in the observing system, improper use of observations, transitions between multiple production streams, or various mistakes that can occur in a complex reanalysis production. Considerable progress has been achieved in this area in recent years, mainly due to advances in data assimilation related to the treatment of biases in satellite observations (Dee and Uppala 2009). It has been demonstrated that near-surface temperature and humidity anomalies estimated from reanalysis data closely match those obtained independently from station observations (Simmons et al. 2004, 2010), and reanalysis data have begun to be routinely used to assess global climate change, e.g. in the annual State of the Climate special issues of the Bulletin of the American Meteorological Society (SOC 2010; 2011; 2012).

It is always preferable to verify trend estimates derived from reanalysis by comparing with independent data sets, but in many cases it is not possible to do so. Users must be aware that the accuracy of trend estimates (and the conclusions they lead to) can differ greatly from one reanalysis to another, depending on the data assimilation methodology used and choice of observing system. Reanalysis is a relatively young field that has seen rapid progress in recent years. For example, Paltridge et al. (2009) showed that specific humidity in the NCEP/NCAR reanalysis had a negative trend with time. Based on this, they cast doubt on the general consensus that the global water vapour feedback was strongly positive (e.g. Dessler and Sherwood, 2009). Dessler and Davis (2010) analysed several reanalysis datasets and found that the NCEP/NCAR was the only one affected by such a negative trend. They suggested as a possible explanation the fact that the specific humidity field in the NCEP/NCAR reanalysis was only constrained by radiosonde humidity observations, whereas humidity fields in more recent reanalyses are additionally constrained by the assimilation of satellite radiances.

In general, uncertainty assessment for specific variables estimated by reanalysis involves the following questions:

- How strongly is the variable constrained by observations? Is it directly or indirectly observed? How accurate are the constraining observations?
- What is the spatial and temporal distribution of the assimilated observations? How does this change in time?
- How accurately can the assimilating model represent the variable? Does the model have skill in extrapolating and/or predicting it?

Climate users interested in the quality of low-frequency variability and/or trend estimates need to consider these aspects throughout the time period in question. In particular, the complexity of the observing system and its evolution over time, associated with changing biases in observations, can introduce spurious low-frequency signals in the reanalysis. Unfortunately many users do not have access to sufficient information in order to fully address the difficult questions listed above. On the other hand, producers of reanalysis data do not have the resources (nor the application-specific knowledge) to answer them either. Part of the solution is to provide better tools and information systems to support users in making their own uncertainty assessments. In particular, it should be made much easier for a user to get detailed information about the observations used in reanalysis, including the quality assessment and any adjustments produced by the reanalysis process itself.

In summary, limitations and caveats of reanalysis products mainly result from

- Lack of observations. The atmosphere is not now, nor ever has been, fully observed.
- Errors in the observations, and lack of information about those errors.
- Shortcomings in the assimilating model, and lack of information about model errors.
- Shortcomings in data assimilation methodology.

- Technical errors and mistakes in reanalysis production.
- Computational limitations (e.g. limitations in spatial and temporal resolution)

The most important of these items are due to lack of information; there are fundamental limitations as to what can be achieved with incomplete observations and imperfect models.

3.2 Quality of input observations

Ultimately, and apart from all technicalities, the attainable realism of a reanalysis depends on the quality of the input observations used, and on the available information about their uncertainties. Especially in the modern observing period (i.e. in recent decades), the accuracy of current reanalysis products such as ERA-Interim (Dee et al. 2011b) is sufficiently high that observations must meet strict quality requirements before they can be usefully assimilated. The key requirements are:

- The relationship between the observations and the model state variables must be accurately represented in the observation operator;
- The errors in the observations must be sufficiently well understood to allow their statistical characterization, e.g. in terms of biases and error covariances;
- Adequate quality control and bias correction procedures for the observations must be available;
- The remaining signal in the observations (i.e. after quality control and bias correction) must add useful information to the reanalysis.

Reanalysis provides a useful framework for assessing the quality of ECV products derived from different components of the global observing system. This is illustrated in figure 10 which compares the relative departures of ozone concentrations retrieved from several instruments from their collocated ERA-Interim ozone analyses, at 10 hPa in the tropics. This type of comparison gives confidence on the quality of the reanalysis and of the data, but also helps spotting issues. An example is the comparison with SCIAMACHY limb data in figure 11, which differs from ERA-Interim estimates by 40%. While the reasons for such large differences are not yet clear, residuals of this magnitude were only seen around 10 hPa in the tropics. The level of agreement between SCIAMACHY limb ozone profiles and the ERA-Interim ozone analyses was better than 10% in the rest of the tropical stratosphere and at all levels in the extra-tropics.

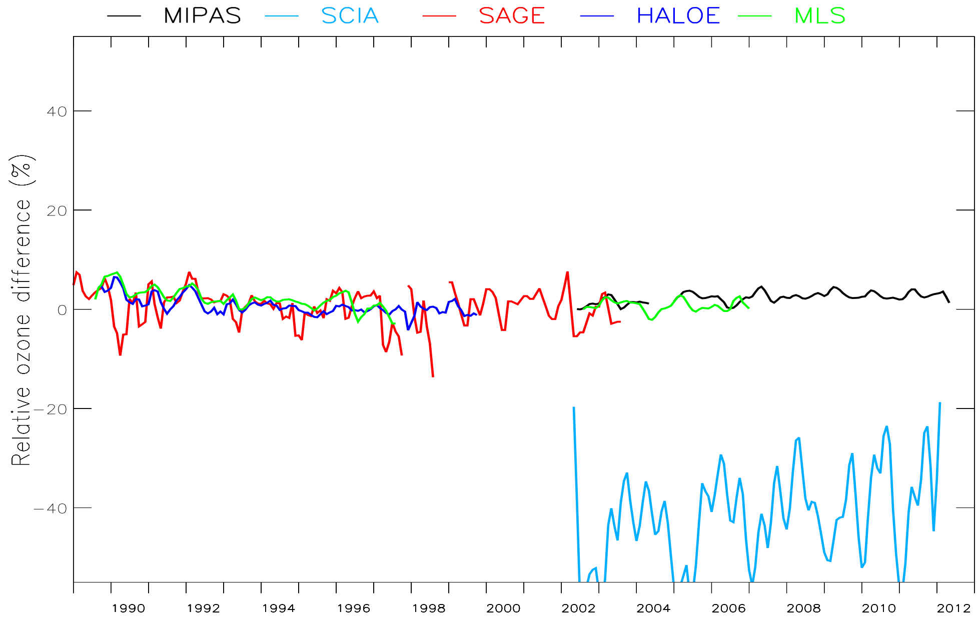

Figure 10: Relative mean difference between the ERA-Interim ozone analyses and reprocessed ozone data from several instruments (as given in the legend, where SCIA is a short name for SCIAMACHY limb ozone profiles) at 10 hPa in the tropics over the period 1989-2012. With the exception of MIPAS during the period 2003-2004, none of the ozone data was assimilated in ERA-Interim. The calculation has been done as described in Dragani (2011). The plot is an adaptation of figure 5 (panel d) of Dragani (2011).

The absolute observation minus reanalysis residuals can also be used to assess observation errors, e.g. using approaches such as the triple collocation method (e.g. Janssen *et al*., 2007) – but this requires that at least two sets of data are available in addition to the reanalysis.

For observations that are assimilated in a reanalysis, the reanalysis provides a continuous on-line observation-model confrontation. In that case the reanalysis output is clearly not independent of the observations. Nevertheless, the data assimilation process itself generates a wealth of information about the uncertainties in the input observations. For example, inconsistencies among the different sources of ozone information used in the reanalysis would be clearly visible in figure 10. More interestingly, the sequential time-stepping data assimilation procedure used to generate the reanalysis involves production of a short (typically 12-24 hours) forecast to provide a first prediction of all observations used in the next analysis step. This so-called background estimate depends only on past observations, and is therefore independent of the observations used in the next analysis. The “background departures” or observed-minus-forecast residuals are stored and can be used for posterior statistical error analysis.

The background departures generated during data assimilation are part of the “reanalysis feedback”, which may also include estimates of observation bias generated during the reanalysis (see next section), and output of the automated quality control embedded in the reanalysis. The reanalysis feedback is an important resource for data quality assessment, which can be exploited to improve the description of input data uncertainties for subsequent reanalyses. Various methods are available to estimate error covariances from background departures, e.g. see Dee and da Silva (1998), and Desroziers (2006). Reanalysis feedback has also been used to detect breakpoints in upper-air temperature data from radiosonde stations, resulting in demonstrable improvements in the historic radiosonde record (Haimberger 2007).

3.3 Treatment of biases in reanalysis

The requirement for a realistic low-frequency variability in climate reanalysis means that special efforts must be made to remove biases from the input observations as well as from the assimilating model. Systematic errors in any of the input sources inevitably introduce biases in the reanalysis. Most observations in fact require substantial adjustments for bias before they can be usefully assimilated. Standard data assimilation methods (such as the Four Dimensional Variational data assimilation scheme, 4D-Var, used in ERA-Interim) were originally designed under the assumption that all assimilated observations have well-characterised uncertainties resulting from random errors only. It has been only recently that data assimilation has become “bias-aware” (Dee, 2005).

Treatment of biases in satellite observations is especially critical. Systematic errors in radiance measurements reflect the complexity of the instruments and the indirect nature of the measurement, and can include large-scale flow-dependent components. In addition to the effects of instrument and calibration errors, biases in satellite data assimilation can result from systematic errors in the radiative transfer models that are either embedded in the assimilation system as in the case of level-1b data (radiance) assimilation or in the retrieval scheme in the case of derived data (retrieval) assimilation. It is important to recognize that it is generally preferable to assimilate radiances rather than retrievals. The reason is that error characterization of derived products, which include additional information to the raw measurements, is much more difficult. The importance of this principle was clearly demonstrated in the 1990s with the development of variational data assimilation methods at ECMWF and at NCEP, which resulted in major improvements in weather forecasting (e.g. Rabier et al., 2000).

A reanalysis assimilates a large volume of observations from different sources, often constraining the same model variable. It is not unusual that these different sources of observations are biased one with respect to another. This is particularly the case for satellite data that now represent the vast majority of all existing atmospheric observations. Space agencies and other data providers are now investing substantial efforts to reprocess the raw measurement data from satellites in order to remove inter-satellite biases and generally to improve the information content of the data (e.g. the Global Space-Based Inter-Calibration System project; more information about the GSICS project can be found at http://www.star.nesdis.noaa.gov/smcd/spb/calibration/icvs/GSICS).

It is now common practice in data assimilation to use bias predictor models to estimate (and then remove) systematic errors in the assimilated observations. Whether observations are assimilated in the form of radiances or retrievals, their bias varies in space and time, and may also depend on atmospheric conditions at the time they were observed. To account for this complexity, bias is typically represented by a predictor model involving properties of the observed atmospheric column (such as the integrated lapse rate) as well as the state of the instrument (such as its field of view). The bias in the input observations is then described by a relatively small (say, < 10) number of parameters, which are the unknown coefficients for the predictor model. These bias parameters can be estimated separately for each channel, for example, by regression to some reference dataset (Harris and Kelly, 2001).

In ERA-Interim, the estimation of bias parameters for satellite radiance data is handled automatically by a variational bias correction system (Dee and Uppala 2009). This system detects the appearance of a new satellite data stream, and it then initialises, updates, and keeps track of bias estimates for radiance observations from all channels for each sensor flying on the satellite. The bias parameters are updated during each analysis cycle by including them in the control vector used to minimise the 4D-Var cost function. This ensures that the bias estimates are continuously adjusted to maintain consistency of the bias-corrected radiances with all other information used in the analysis, which includes the conventional observations as well as the model background. An important practical advantage of this approach is that it removes the need for manual tuning procedures, which are prone to error and simply impractical in the modern age.

Figure 11 shows an example of temporal consistency and homogeneity of the bias corrected observation minus the ERA-Interim model background departures (top panel) computed for the Microwave Sounding Unit (MSU) channel-2 radiances flown by several NOAA satellites over a period of over 30 years. The bias corrections applied to each instrument, to achieve such homogeneity, are plotted in the bottom panel. Instruments on different satellites are biased relative to each other and differences can be as large as 1.5K in brightness temperature.

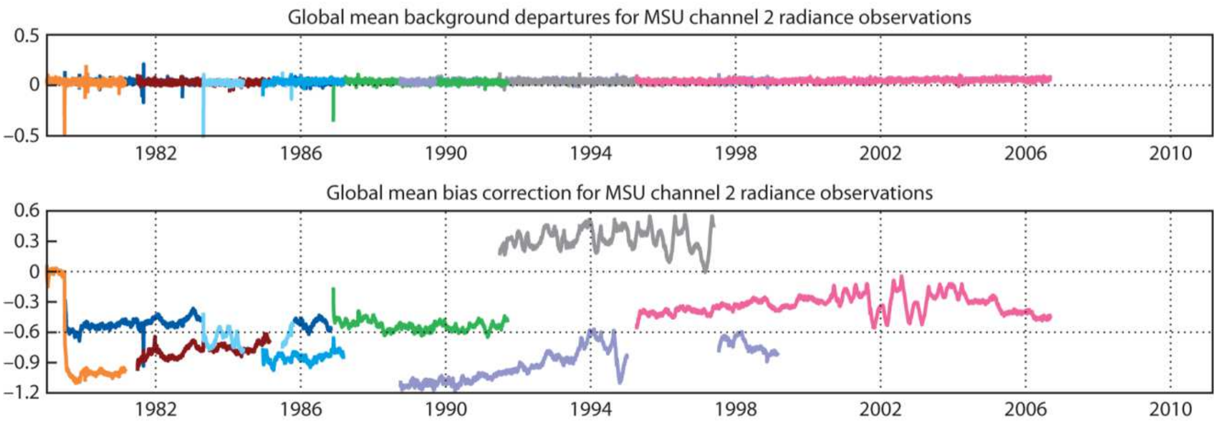

Figure 11: Bias-corrected radiance measurements from Microwave Sounding Unit flown on successive NOAA satellites (top panel; colours indicate different satellites). The global mean bias corrections for the MSU data, produced by the variational analysis in ERA-Interim, account for calibration differences, orbital drifts and various other instrument errors (bottom panel).

Successful examples, such as the one presented in figure 11, increase the confidence in the latest reanalyses to address well-documented contamination of climate signals by changes in the observing system and possibly to accurately simulate the long-term trend in those signals. However, users need to be cautious when using reanalysis data for climate change assessment. Temporal variation in the observational constraint can still produce, in some cases, artificial shifts in the reanalysis time series - especially if the assimilating model has systematic errors -

even when using bias-corrected observations. Dee and Uppala (2008) discussed a stratospheric example, reproduced in figure 12 that shows how changes in the observing system (in this case the switch between SSU and AMSU-A) affected the ERA-Interim stratospheric temperature in summer 1998. The discontinuity in the upper stratosphere (at 5 hPa and above) occurs because the assimilating model has large temperature biases there, while the two instruments with different measurement characteristics can only partly counteract those biases.

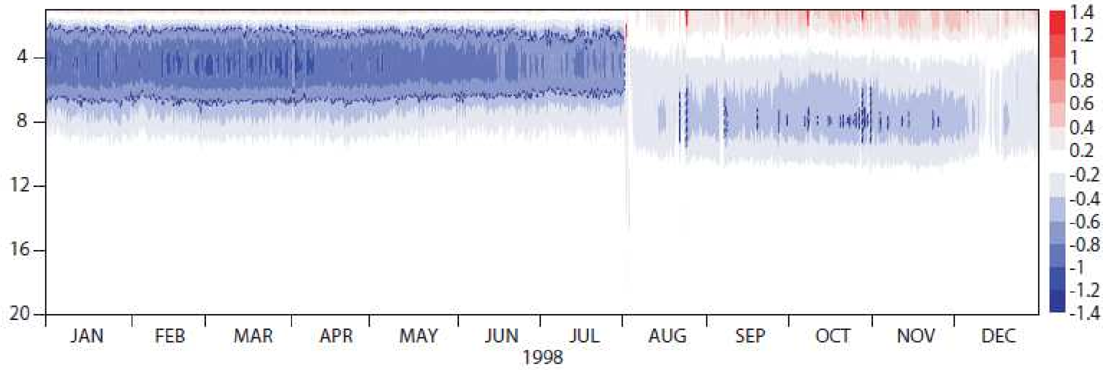

Figure 12: Globally averaged analysis increments for upper-stratospheric temperature (30 hPa and up) in ERA-Interim during 1998, when the switch from SSU to AMSU-A took place. Courtesy of Dee and Uppala (2008).

This example is typical of a situation where the model performs relatively poorly and the observing system is sparse; in this case no additional information is available to improve the estimates.

3.4 Additional benefits (and challenges) in using a complex reanalysis system

The complete description of a physically plausible atmosphere consistent with observations provided by reanalysis makes it possible to do many things that simply cannot be achieved otherwise. For example, it permits estimation of a large set of climate variables, even for variables that are not well observed, e.g. stratospheric winds, radiative fluxes, root-zone soil moisture, etc. These estimates are important because they are indirectly constrained by the observations used to initialise the model. An example is detailed diagnostics of the global energy budget and the hydrological cycle (Trenberth et al. 2011). Such diagnostics are especially useful if they involve known time-invariant properties of the climate system. These are (usually) conserved by the assimilating model in a reanalysis, but tend to be destroyed by the assimilation increments, depending on the nature of the observational constraints and on the method of assimilation. Budget diagnostics can be used to demonstrate shortcomings as well as progress in climate reanalysis (Berrisford et al. 2011). In other cases, it facilitated the assessment of inter-related fields to check their consistency.

The drawback, however, is that in the absence of direct observations it is difficult to quantify the uncertainties in estimates of model-generated variables, as they depend on errors in the model as well as on the strength of the (indirect) observational constraint. Some insight into the uncertainties can be obtained by using ensemble techniques, with the important caveat that it is not practical to sample more than a few selected sources of uncertainty in a reanalysis.

Another potential benefit of using a complex, coupled system such as that of a reanalysis is the ability of producing adjustment in one variable while constraining a different one. This is particularly the case of data assimilation systems based on a 4D-Var scheme. Coupled data assimilation potentially allows for better use of observations with information about both meteorology and e.g. aerosols or chemistry. These are important advantages over uncoupled or weakly coupled systems, in which either the model integration or the analysis of observations (or both) is performed in separate steps. Thus, it represents in general a desirable aspect of complex data assimilation systems, as it permits to generate information about not well observed fields (e.g. stratospheric winds) by assimilating observations of different parameters (e.g. ozone), and thus estimate a large set of climate variables. The increased complexity in the system requires a proportional increase in assumptions and choices to be made for its implementation resulting in additional degrees of freedom in the modelling system. In this case, a realistic analysis can be produced only if these additional degrees of freedom are adequately constrained by accurate observations. This has important implications for climate reanalysis, since the instrumental record available e.g. for a reanalysis of atmospheric composition is limited, both in quality and quantity.

When the observations cannot provide an adequate constraint, a negative impact can result in the analyses. During the production of ERA-Interim, it was noticed that the assimilation of ozone profile data retrieved from the ERS-2 GOME instrument could generate large and unrealistic changes in the upper stratospheric circulation, where the model background is not well constrained by observations (see Figure 13). These upper-level increments provided the most effective way for the 4D-Var analysis to accommodate the observed local changes in ozone concentration further below. It should be possible, in theory, to extract useful information about advection from stratospheric tracer observations in a 4D-Var analysis. In practice this can work well only if both the model background and the observations are sufficiently accurate, which is currently not the case.

As a direct result of the discovery of this problem in ERA-Interim, the 4D-Var analysis in the operational forecast system was modified in 2007 to prevent any changes in temperature and wind resulting directly from the analysis of ozone data. This change was implemented across all applications that are run at ECMWF, including the weather forecasting system. Recent improvements in the data assimilation system, including implementation of variational bias corrections for ozone observations, have ameliorated the problem, and it is now being investigated whether a fully coupled 4D-Var analysis for ozone can be safely reinstated.

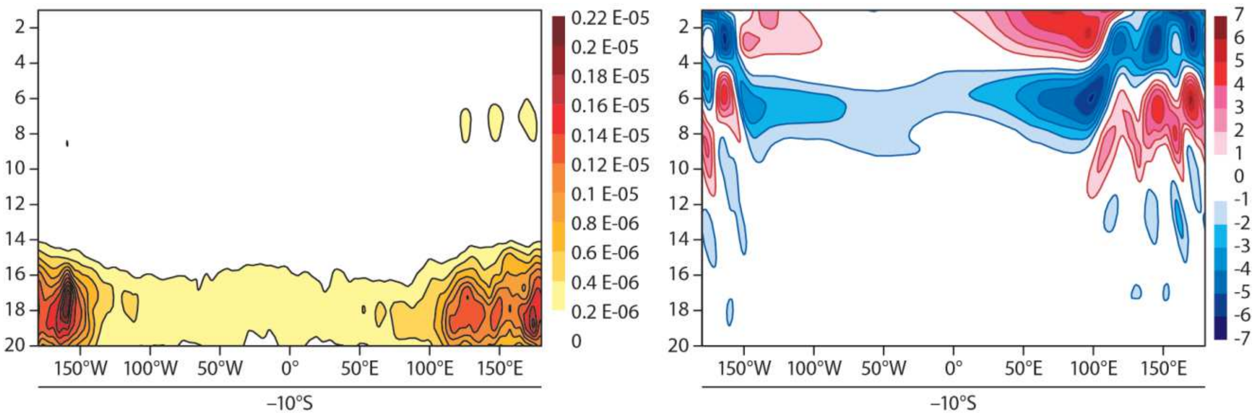

Figure 13. Impact of GOME ozone profile observations only, in a single 12h 4D-Var cycle (4 July 1995, 0 UTC), along the latitude circle 10S for the top 20 model levels (of a 60-level model, i.e., from 40hPa up to 0.1hPa). Ozone increments (left panel) with maximum values of about 2 mg/kg are concentrated in locations where the satellite track crosses 10S. They are everywhere positive in this vertical plane, because the model ozone concentrations are biased low. Unrealistic temperature increments (right panel) ranging from -6.6K to +6.3K occur at much higher levels.

4. References

Berrisford, P., P. Kållberg, S. Kobayashi, D. Dee, S. Uppala, A. J. Simmons, P. Poli, and H. Sato, 2011. Atmospheric Conservation Properties in ERA-Interim. *Q. J. R. Meteorol. Soc.*, **137** (659), 1381–1399.

Bodas-Salcedo, A. et al., 2012: The Surface Downwelling Solar Radiation Surplus over the Southern Ocean in the Met Office Model: The role of midlatitude cyclone clouds, *J. Climate*, **25**, 7467–7486, doi: http://dx.doi.org/10.1175/JCLI-D-11-00702.1.

Chepfer, H. et al., 2010: The GCM Oriented CALIPSO Cloud Product (CALIPSO-GOCCP)", *J. Geophys. Res.*, **105**, D00H16, doi: 10.1029/2009JD012251.

CMUG, 2011: Deliverable 4.1, Scientific exploitation plan (v. 2.0).

CMUG, 2012: Deliverable 1.2, User requirements document (v. 1.6).

Dee, D.P 2005: Bias and data assimilation, *Q. J. R. Meteorol. Soc*, **131**, 3323-3343.

Dee, D. P., Arlindo M. da Silva, 1999: Maximum-Likelihood Estimation of Forecast and Observation Error Covariance Parameters. Part I: Methodology. *Mon. Wea. Rev*., **127**, 1822–1834.

Dee, D. P. and S. Uppala 2008: Variational bias correction in ERA-Interim, *ECMWF Technical Memorandum*, **575**.

Dee, D. P. and S. Uppala 2009: Variational bias correction of satellite radiance data in the ERA-Interim reanalysis, *Q. J. R. Meteorol. Soc*, **135**, 1830-1841.

Dee, D. P., E Källén, A J Simmons, and L. Haimberger, 2011a: Comments on "Reanalyses Suitable for Characterizing Long-Term Trends" *Bull. Amer. Meteorol. Soc.*, **91**, 66-70.

Dee, D.P., S.M. Uppala, A.J. Simmons, P.Berrisford, P.Poli, S.Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M.

Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart 2011b:. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.*, **137**, 553-597.

Desroziers, G., L. Berre, B. Chapnik, P. Poli, 2006: Diagnosis of observation, background and analysis-error statistics in observation space. *Q. J. R. Meteorol. Soc.*, **131**, 3385-3396.

Dessler, A. E. and S. M. Davis 2010: Trends in tropospheric humidity from reanalysis systems. *J. Geophys. Res*., **115**, D19, 1984–2012.

Dessler, A. E., and S. C. Sherwood 2009: A matter of humidity, *Science*, **323**, 1020–1021.

Dragani, R. 2011: On the quality of the ERA-Interim ozone reanalyses: Comparisons with satellite ozone data. *Q. J. R. Meteorol. Soc.*, **137** (658), 1312--1326.

Fernandes R. et al., 2009: Controls on northern hemisphere snow albedo feedback quantified using satellite Earth observations, *Geophys. Res. Lett.*, **36**, L21702, doi:10.1029/2009GL040057.

Gleckler, P.J., K.E. Taylor and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, doi:10.1029/2007JD008972

Gómez-Navarro, J. J. et al., 2012: What is the role of the observational dataset in the evaluation and scoring of climate models? *Geophys. Res. Lett.*, **39**, L24701, doi:10.1029/2012GL054206.

Haimberger, L. 2007: Homogenization of radiosonde temperature time series using innovation statistics. *J. Climate*, **20**, 1377-1403

Hall, A., and X. Qu, 2007: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett*., **33**, L03502, doi:10.1029/2005GL025127.

Harris, B. A. and G. Kelly 2001: A satellite radiance-bias correction scheme for data assimilation, *Q. J. R. Meteorol. Soc*., **127**, 1453-1468.

Hegerl, G.C. et al., 2007: Understanding and Attributing Climate Change. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S. et al. (eds.)]. Cambridge University Press, Cambridge, UK and New York, NY, USA.

International Ad Hoc Detection and Attribution Group (IDAG), 2005: Detecting and attributing external influences on the climate system: A review of recent advances, *J. Clim*., **18**, 1291–1314.

Janssen, P.A.E.M., Abdalla, S., Hersbach, H. and Bidlot, J.R. 2007: Error estimation of buoy, satellite, and model wave height data, *J. Atmos. Oceanic Technol*., **24**, 1665–1677.

Jiang, J.H. et al., 2012: Evaluation of cloud and water vapor simulations in CMIP5 climate models using nasa A-Train satellite observations, *J. Geophys. Res.*, **117**, D14105, 10.1029/2011JD017237.

Jones, G.S., and P.A. Stott, 2011: Sensitivity of the attribution of near surface temperature warming to the choice of observational dataset, *Geophys. Res. Lett*., **38**, L21702, doi:10.1029/2011GL049324.

Knutti, R., G. Abramowitz, M. Collins, V., P. J. Gleckler, B. Hewitson, and Linda Mearns, 2010: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In: *Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Loeb, N.G. et al., 2009: Toward optimal closure of the Earth's top-of-atmosphere radiation budget, *J. Climate*, **22**, 748–766, doi: http://dx.doi.org/10.1175/2008JCLI2637.1.

Luo, Y.Q. et al., 2012: A framework for benchmarking land models. *Biogeosciences*, **9**, 3857–3874.

Michou, M., D. Saint-Martin, H. Teyssèdre, A. Alias, F. Karcher, D. Olivié, A. Voldoire, V.-H. Peuch, H. Clark, J.N. Lee, and F. Chéroux, 2011 : A new version of the CNRM Chemistry-Climate Model, CNRM-CCM: description and improvements from the CCMVal2 simulations, *Geoscientific Model Development*, **4**(4), 873-900, doi : 10.5194/gmd-4-873-2011.

Paltridge, G., A. Arking, and M. Pook 2009: Trends in middle and upper level tropospheric humidity from NCEP reanalysis data, *Theor. Appl. Climatol*., **98**, 351–359.

Rabier, F., H. Järvinen, E. Klinker, J.-F. Mahfouf and A. Simmons 2000: The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics, *Q. J. R. Meteorol. Soc*., **126**, 1143-1170.

Sexton D.M.H et al., 2012: Multivariate probabilistic projections using imperfect climate models. Part I: outline of methodology, *Clim Dyn*., **38**, 2513-2542, doi:10.1007/s00382-011-1208-9.

Sexton D.M.H. and J.M. Murphy, 2012: Multivariate probabilistic projections using imperfect climate models. Part II: robustness of methodological choices and consequences for climate sensitivity, *Clim Dyn*,. **38**, 2543-2558, doi: 10.1007/s00382-011-1209-8.

Simmons, A. J., P. D. Jones, V. da Costa Bechtold, A. C. M. Beljaars, P. W. Kållberg, S. Saarinen, S. M. Uppala, P. Viterbo, and N. Wedi, 2004: Comparison of trends and low-frequency variability in CRU, ERA-40 and NCEP/NCAR analyses of surface air temperature, *J. Geophys. Res*., **109**, D24115, doi:10.1029/2004JD005306.

Simmons, A. J., K. M. Willett, P. D. Jones, P. W. Thorne and D. P. Dee 2010: Low-frequency variations in surface atmospheric humidity, temperature, and precipitation: Inferences from reanalyses and monthly gridded observational data sets. *J. Geophys. Res*., **115** D1, 1984–2012.

SOC 2010: State of the Climate in 2009. *Bull. Amer. Meteor. Soc*., **91** (7), 2010.

SOC 2011: State of the Climate in 2010. *Bull. Amer. Meteor. Soc*., **92** (6), 2011.

SOC 2012: State of the Climate in 2011. *Bull. Amer. Meteor. Soc*., **93** (7), 2012.

Thorne, P. W., R. S. Vose 2010: Reanalyses suitable for characterizing long-term trends. *Bull. Amer. Meteorol. Soc*., **91**, 353–361.

Thorne, P. W., et al., 2010: Tropospheric temperature trends: History of an ongoing controversy, Wiley *Interdiscip. Rev. Clim. Change*, **2**, 66–88.

Trenberth, K. E., J. T. Fasullo, and J. Mackaro 2011: Atmospheric Moisture Transports From Ocean to Land and Global Energy Flows in Reanalyses. *Journal of Climate*, **24**, 4907-4924.

Vautard, R. and P. Yiou, 2012: Attribution: Robustness of warming attribution, *Nature Climate Change*, **2**, 26-27, doi:10.1038/nclimate1343.

Walters, D. N. et al., 2011: The Met Office Unified Model Global Atmosphere 3.0/3.1 and JULES Global Land 3.0/3.1 configurations, *Geosci. Model Dev*., **4**, 919-941, doi:10.5194/gmd-4-919-2011.