COordinating Earth observation data validation for RE-analysis for CLIMAte ServiceS

Grant agreement N°: 313085

Deliverable D5.53

Procedure for comparing reanalyses, and comparing reanalyses to assimilated observations and CDRs

Version date: 17 July 2014

Theme: (SPA.2012.1.3-02) GMES Climate Change – Coordination of Earth observation data validation for re-analysis

Project start date (duration): 2013-01-17 (30 months)
Project web site: http://www.coreclimax.eu/
WP5: Intercomparing reanalysis results
Lead partner for deliverable: Finnish Meteorological Institute (FMI)
WP leader: Hilppa Gregow (FMI)
Task 5.3 Collecting and synthesizing information from global and regional reanalyses *(ECMWF, DWD)*

Revision history:
10 April 2014 Initial draft (AKW)
17 April 2014 Revision (AO, AKW)
23 April 2014 Add discussion of motivations for reanalysis inter-comparison and who would conduct it (DT)
13 May 2014 Outline modification (descriptive product comparison, direct product comparison internal metrics, and representation of natural processes), added text and figures in sections 2 and 5.1 (PP)
14 May 2014 Comments in regard to in situ pointwise data, statistically downscaled gridded data, dynamically downscaled gridded data (bias corrections using ERA-Interim) and effect of varying spatial and temporal scale in CC service and reseach (HG)
10 June 2014 Added text and figures in sections 3.2, 3.3, 5.1, 5.2, and 5.3. (PP)
10 June 2014 Commenting more about the doc, regional versus global and documents possible outreach component in reanalysis.org and CCCS (HG)
11 June 2014 Introduce notion of thematic comparison (PP)
17 June 2014 Summary tables of the main needs, further elaborations (DT).
18th June evening (AKW): Split 3rd party comparison and reanalysis inter-comparison into 2 different sections. Additions to the summary tables
19 June 2014 (DT) Harmonization and additional remarks.
20 June 2014 (HG) Comments and brief edit
7 July 2014 (DT, AKW, AO) Final edits
8 July 2014 (AR) Typo corrections and layout editing
17 July 2014 (AKW, DT) Correction and augmentation of Figure 10

This document describes procedures for comparing a reanalysis with other reanalyses or with observational Climate Data Records, as an added value for a future CCCS.

Table of contents

1. Background

The EU FP7 project CORE-CLIMAX undertakes preparatory work for shaping the envisaged Copernicus Climate Change Service (CCCS). One focus of the project is the intercomparison of reanalyses. Reanalysis intercomparison activities are a key component of characterizing reanalysis uncertainties, which in turn is of paramount importance regarding the use of reanalysis data within CCCS. The aim is to yield information that assists users in deciding which reanalysis product might be most suitable for their particular application. With increased resolution of the reanalyses, and with application aiming at high resolution processes, there is a growing need for evaluation with observations.

Uncertainty of a measurement is a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used (BIPM, 2008). Uncertainty can be due to lack of knowledge (epistemic uncertainty) or due to inherent variability (aleatoric uncertainty). In principle, applicability of the BIPM Guide to the Expression of Uncertainty in Measurement (GUM) extends beyond measurement uncertainty to information derived from measurements using application of complex models and filtering techniques which recycle information. In practice, the ability to extend formal error propagation to the more complex systems is hampered by difficulties in establishing traceable uncertainties for the auxiliary inputs and parameters of these more complex systems. Consequently, the concepts defined by the BIPM stop short of providing a complete way of defining uncertainty for gridded fields and other spatio-temporally complete estimates of our environment as produced by reanalysis. Although the situation is improving, through increasing attention to traceability in more aspects of Earth-system observational datasets, it remains prudent and pragmatic to complement quantitative uncertainty estimates with documentation on qualitative aspects of uncertainty, for example that sources of uncertainty in reanalysis systems include the specification of geophysical forcings (e.g. sea surface temperature fields) and the representation of geophysical processes (e.g. the sub-grid parametrizations).

The most popular reanalysis products are four-dimensional (space-time) gridded data of environmental variables. These are hereafter referred to as gridded fields. Examples of gridded fields include for example temperatures, which can be described at various altitudes for the atmosphere, at various soil depths for the land-surface, and at various ocean depths for the ocean.

The present document presents a set of procedures for comparing reanalyses, and comparing reanalyses to assimilated observations and CDRs. To do so, five categories of comparisons are identified. These are accompanied by two complexity ratings. The first rates the complexity of conducting the procedure (simple, moderate, difficult), and the second rates the complexity of interpreting the results (simple, moderate, difficult):

1. **descriptive product comparison (simple to conduct, simple to interpret)**

2. **comparison with third party observation-based CDRs (moderate to conduct, moderate to interpret)**

3. **inter-comparison between different reanalyses (moderate to conduct, moderate/difficult to interpret)**

4. **thematic comparison (difficult to conduct, difficult to interpret)**

5. **internal metrics comparison (difficult to conduct, moderate to interpret)**

Figure 1 illustrates the different categories schematically. From the user questionnaire and literature studies (D5.52) it is clear that users would benefit from all of these categories of comparisons. They all help drawing conclusions on the value and on the proper use of the reanalysis products for specific applications. Diversity is large though, and there is a need to build capacity for conducting and interpreting such intercomparisons. Generic intercomparisons are valuable but cannot cover all the specifics of particular applications. There is thus a need to empower users to conduct and interpret intercomparisons tailored to their own applications. Education and training will be critical to raise the level of user expertise above what is needed for purely descriptive comparisons (category 1). For instance, many traditional users still rely on gridded fields based on observations only (mainly for the surface parameters), and a lot of valuable information obtained with method 5 (internal metrics comparison) may simply be not known to the users.

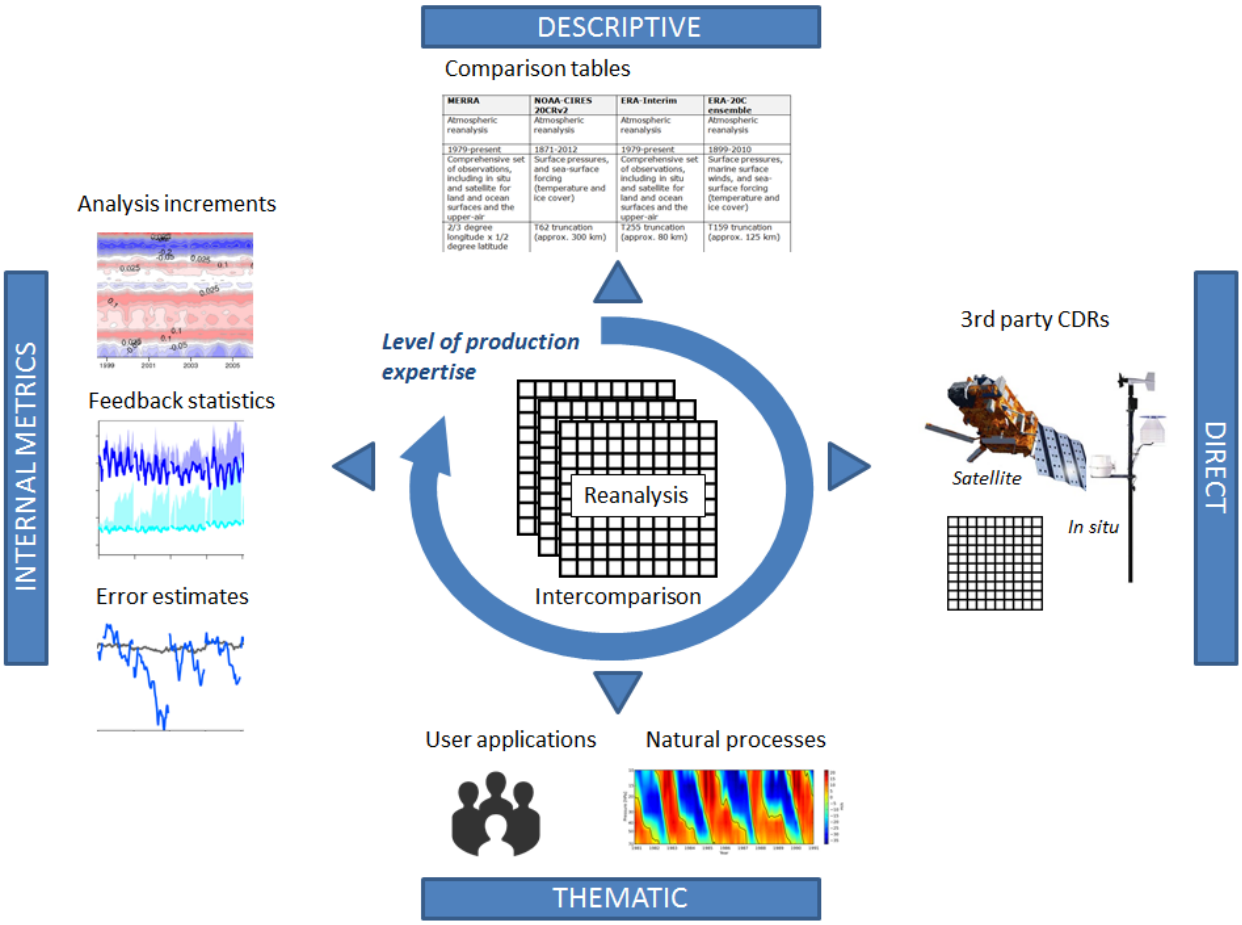

Figure 1: Procedures for comparing reanalyses.

The current document concentrates on technical descriptions of these procedures, drawing on current best-practice. Best-practice continues to evolve and there are good prospects for improvement during the Climate Service timeframe, so we also highlight some implied/associated needs that that will effect the transition from ad-hoc/research activities to operational Climate Services. The service-related issues raised here will be consolidated with findings from other Core-Climax workpackages/tasks, in a subsequent Core-Climax document (Deliverable 5.54).

2. Descriptive product comparison

Making a descriptive comparison of products and their generation from reanalysis or observation-based CDRs, is often overlooked as being a non-necessary, 'obvious' task. Yet, such comparison provides the first level of information needed before proceeding further with any other comparison. It is hence the first point of entry into a comparison exercise.

There is a tendency for descriptive comparison to be done without much explicit thought or documentation, because the investigators know (or consider that they know) the generation process and characteristics of the various products they are comparing. However, it is common to find that mistakes in the interpretation of the results come from an incomplete prior knowledge of how each product was derived. Consequently, a well-documented descriptive comparison is a basic foundation to any comparison, even for experts.

The advent of Climate Services will bring greater demands on the breadth and depth of product intercomparison activities. This will require more extensive participation from both product producers and product users (often in collaboration). The expertise of participants will be more varied than at present, and continuing with the current "ad hoc" approach to descriptive product comparison would increase the likelihood of erroneous interpretations of intercomparison results. There are thus compelling reasons to make the descriptive product comparison more systematic.

An example of descriptive product comparison for reanalyses is given in Table 1.

Similar tables are available, for example:

- covering 11 atmospheric reanalyses, at the following URL http://www.reanalyses.org/atmosphere/overview-current-reanalyses (accessed 13 May 2014)
- covering 13 ocean reanalyses, at the following URL http://www.reanalyses.org/ocean/overview-current-reanalyses (accessed 13 May 2014)

Table 1: Example of descriptive product comparison for a selection of upper-air temperatures products

Feature	RSS v3.2	RICH	MERRA	NOAA-CIRES 20CRv2	ERA-Interim	ERA-20C ensemble
Type of product	Satellite observation-based CDR	In situ observation-based CDR	Atmospheric reanalysis	Atmospheric reanalysis	Atmospheric reanalysis	Atmospheric reanalysis
Time range	1979-present	1958-present	1979-present	1871-2012	1979-present	1899-2010
Observation input	Microwave sounder radiance (MSU, AMSU-A)	Radiosonde temperature	Comprehensive set of observations, including in situ and satellite for land and ocean surfaces and the upper-air	Surface pressures, and sea-surface forcing (temperature and ice cover)	Comprehensive set of observations, including in situ and satellite for land and ocean surfaces and the upper-air	Surface pressures, marine surface winds, and sea-surface forcing (temperature and ice cover)
Product horizontal resolution	2.5 degree longitude x 2.5 degree latitude	10 degree longitude x 5 degree latitude resolution (also available: individual, monthly adjusted, station time-series, usually twice-daily)	2/3 degree longitude x 1/2 degree latitude	T62 truncation (approx. 300 km)	T255 truncation (approx. 80 km)	T159 truncation (approx. 125 km)
Product vertical resolution	3 layers (middle troposphere, troposphere-stratosphere, lower	16 levels between 1000 hPa and 10 hPa	72 levels between surface and 0.01 hPa	28 levels between surface and 10 hPa	60 levels between surface and 0.1 hPa	91 levels between surface and 0.01 hPa

Feature	RSS v3.2	RICH	MERRA	NOAA-CIRES 20CRv2	ERA-Interim	ERA-20C ensemble
	stratosphere)					
Product temporal resolution	Monthly	Monthly	Hourly	3-hourly	6-hourly	3-hourly
Filtering technique to reduce random noise	Averaging	Averaging	6-hour 3DVAR data assimilation with Incremental Analysis Update	6-hour Ensemble Kalman Filter	12-hour 4DVAR data assimilation	24-hour ensemble of 4DVAR data assimilations
Bias correction to reduce systematic errors	Series of adjustments to correct in particular for orbital drift, viewing geometry change, local time change, and instrument changes	Homogenization	Variational bias correction for radiances, and other schemes for in situ observations (incl. RAOBCORE v1.4 for radiosondes temperature)	Removal of mean difference observation minus forecast for past 60 days	Variational bias correction for radiances, and other schemes for in situ observations (incl. RAOBCORE v1.3 for radiosondes temperature)	Variational bias correction for surface pressures
Main reference	Mears and Wentz, 2009, doi: 10.1175/2008JTECHA1176.1	Haimberger et al., 2012, doi: 10.1175/JCLI-D-11-00668.1	Rienecker et al., 2011, doi: 10.1175/JCLI-D-11-00015.1	Compo et al., 2011, doi: 10.1002/qj.776	Dee et al., 2011, doi: 10.1002/qj.828	Poli et al., 2013, ERA Report Series 14

Core-Climax Workpackage 2 has prepared some of the groundwork for facilitating descriptive product comparison. In particular, it has developed a standard Dataset Description Document for ECV products. Reanalysis producers who complete the Dataset Description Document for their reanalysis datasets will be providing the basic information needed for descriptive product comparison. A further need will be the co-ordinated collection of the basic information (into a common and readily accessible database for instance) and subsequent synthesis into the tabular comparisons such as Table 1. For maximum effectiveness, the co-ordination functions need to incorporate information about newly generated products in a timely fashion, on an on-going basis, and to make the synthesis results widely available. These co-ordination functions are well-suited to being elements of an operational Climate Service.

The main needs for Descriptive Product Comparison are summarized in Table 2 below.

Table 2: Summary of needs for Descriptive Product Comparison

Identified Need	**Purpose**	**Who should be involved?**	**At what stage?**	**Further comments**
Descriptive Product Comparison	Provide basic foundation for other in-depth comparisons	Dataset providers, co-ordinating body	See following entries	
comprising …				
Dataset Description Documents for individual datasets	Provide standard description of key characteristics of the individual datasets	Dataset provider	Upon completion/release of the individual dataset	Template generated by Core-Climax Workpackage 2.
Co-ordinated compilation/synthesis of individual Dataset Description Documents	Collate/present the information from multiple datasets in a way that facilitates intercomparison.	Co-ordinating body	On an on-going basis, updating when new Dataset Descriptions are received/updated.	Such functions are well-suited to being elements of an operational Climate Service.

3. Comparison with third-party products

We use the term "third-party products" to refer to observation-based Climate Data Records that have been produced via non-reanalysis procedures. Validation of such products is itself a challenging but necessary task (Core-Climax, Workpackage 3).

3.1. Comparison with third-party gridded observation-based CDRs

Such a comparison is relatively simple to implement and entails considering, on the one hand, observation-based CDR products which are already 'filled-in' (i.e., spatio-temporally complete, averaged and/or interpolated), and, on the other hand, reanalysis fields (spatio-temporally complete by design). The GPCP datasets of monthly precipitation derived from satellite and surface measurements (http://www.gewex.org/gpcp.html) are archtypical examples of third-party observation-based CDRs.

However, interpretation of the results requires more knowledge of how each product was derived, in particular regarding resolution, representativeness and exact domain area of validity. For example two reanalyses or CDRs may use slightly different land-sea masks, and a careful comparison requires considering only points of matching characteristics in all datasets to avoid misinterpretation. In such situations, it is helpful for users to have access to information about the land-sea masks, either in the form of figures or in the form of the data plus auxiliary tools to examine areas of interest specific to their application (e.g. coastal areas and cities.

Parameters can be sensitive to processing:

1) Diurnally averaged precipitation can be sensitive to the representativeness of the sample being averaged, for example if the observations are limited to certain times of day.
2) The equivalent parameter from a reanalysis dataset will depend on the interplay between the underlying forecast model (with its representation of the relevant physics/dynamics) and the assimilated observations, and may be affected by systematic differences (relative biases) between the two.
3) Coastal values of temperature and wind are highly dependent on spatial resolution and grid spacing. Dynamical downscaling requires very high spatial and temporal resolution to capture the variations in boundary layer-processes.

With the reanalysis grid production, these effects are carefully monitored and characterized. With traditional grid production, we have in principle the same issues: prior estimate taken from a climatology that requires its own quality to be assessed, unknown or changing spatial representativity of observations, subjective choice of smoothing length scale, homogeneity of measurement procedures and instruments (e.g. a switch to automated stations), or a change in the observing system (e.g. change of station numbers over time in a particular grid cell). Although some sophisticated methods may be applied, there is more often a less mathematically stringent characterization of these issues in "cheaper" productions. If there is sufficient observation coverage, neglection of these issues may be justified. However, users and producers might in many cases not be aware of these issues at all, and likely to attribute the differences too quickly to the reanalysis, simply because such issues are given thought there.

Interpretation should also take into account inter-dependence between the products being compared: inter-dependence can arise for example when the underlying observations used to generate gridded observation-based CDRs are also part of the reanalysis production (either directly via active assimilation, or even indirectly e.g. as part of bias correction of other observations)

Figure 2 shows average precipitation over global oceans from various sources. The first apparent feature is that the more recent products on this plot (JRA-25, ERA-Interim) feature fewer spurious jumps than earlier ones (ERA-40, NCEP-DOE Reanalysis 2), when compared to the observational CDR provided by the Global Precipitation Climatology Project (GPCP). Another feature is that all reanalyses shown here over-estimate the precipitation by as much as 50% of the observational estimate.

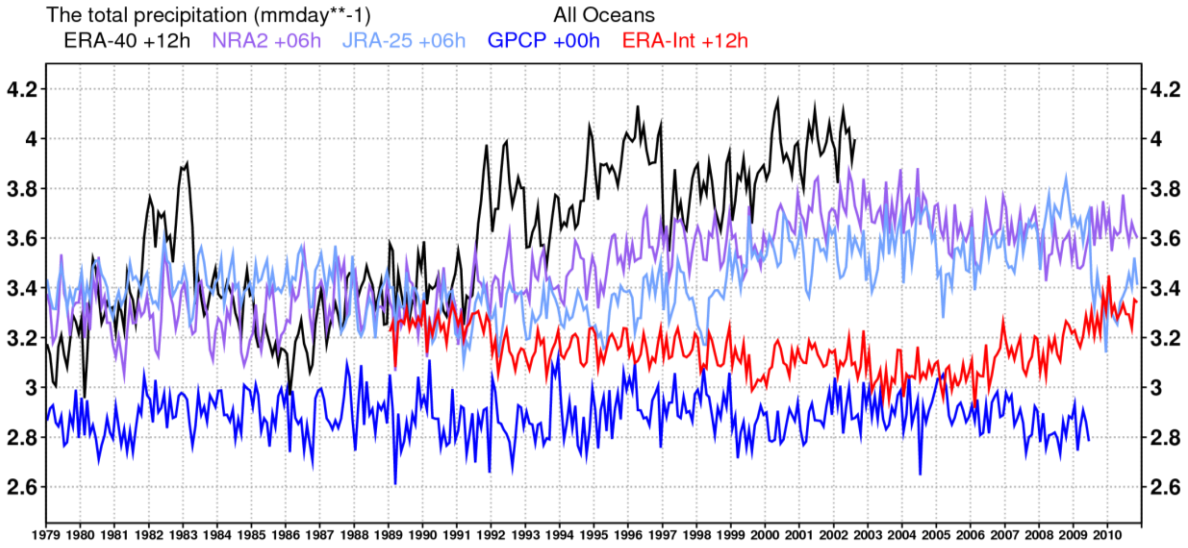

Figure 2: Time-series of average total daily precipitation (mm/day) over global oceans according to several reanalyses (ERA-40, NCEP-DOE Reanalysis 2, JRA-25, ERA-Interim) and an observational CDR (GPCP)

Other comparisons abound, for example for global budgets. One can point to the work of Kevin Trenberth and his group, who have probably authored the most highly cited publications in this area. The Google Scholar page of this NCAR Distinguished Senior Scientist includes citations and links to his articles (http://scholar.google.com/citations?hl=en&user=ovnjqjMAAAAJ&view_op=list_works).

The main needs for Comparison between reanalysis products and third-party gridded observation-based CDRs are summarized in Table **3** below.

Table 3: Main needs and considerations for Comparison between reanalysis products and third-party gridded observation-based CDRs.

Identified Need	**Purpose**	**Who should be involved?**	**At what stage?**	**Further comments**
Comparison between gridded reanalysis products and third-party gridded observation-based CDRs	Identify/explain differences in gridded products at the level of specific Essential Climate Variables. Provide guidance on the interpretation/use of different products	Dataset providers, dataset users	Prior to dataset release (production quality control), and after release (dataset evaluation and feedback)	Pro:Quick sanity check, Con:Gridding introduces changes in frequency distributions and some correlations, interpretation requires

	(both reanalysis and observation-based). Provide feedback to providers of reanalysis products and to providers of observational datasets.			knowledge of both production methods and re-gridding effects.
typically addressing topics such as ...				
representativeness				Gridding smoothes over this issue
systematic differences, relative biases			Needed when users consider to switch from using traditional grid fields to using reanalysis grids	When comparing, there may be no dataset that can be considered the "truth".
inter-dependence of the datasets being compared				statistical implications of transforming both to a common grid.

Co-ordinating functions	Collate/disseminate the information from bi-lateral and multi-lateral comparisons. Review/update the topics to be addressed.	Co-ordinating body	On an on-going basis, updating when new comparisons are received/updated.	

3.2. Comparison with third-party in situ or swath (satellite) observation-based CDRs, at the observation times and locations

Such a comparison entails comparing data that are unevenly distributed with reanalysis. For such a comparison, one maps the latter to the observation location date and time. Two examples are given here: comparison with in situ data, and comparison with satellite data.

Comparison with *in situ* observations

The first example uses observations of downwelling longwave radiation measured by a radiation sensor on a buoy. Such measurements over ocean are quite rare, and usual estimates for radiation at the ocean surface are typically formed by applying bulk formulae to meteorological measurements. The data used here come from a NOAA National Data Buoy Center (NDBC) buoy augmented by dedicated in situ radiation sensors. The project "New England Shelf Fluxes", sponsored by JAII: Massachusetts Technology Collaborative's John Adams Innovation Institute. The observation time-series was retrieved from Woods Hole Oceanographic Institution website (http://www.whoi.edu/). The sensor is an ASIMet longwave radiation (LWR) module employing an Eppley Precision Infrared Radiometer (PIR). We compare here in Figure 3 such observations from NDBC buoy#44008 for May 2010 with estimates produced by two reanalyses. To perform such comparison, the reanalysis gridded data are interpolated spatially, bi-linearly from their native resolution, to the observation location, and use the nearest reanalysis neighbour in time (both available at 3-hourly resolution for the radiation parameter considered) to the observation (hourly). One notices first a remarkable agreement in the timing of the weather events, this estimate of radiation being dominated by meteorological events and cloud coverage.

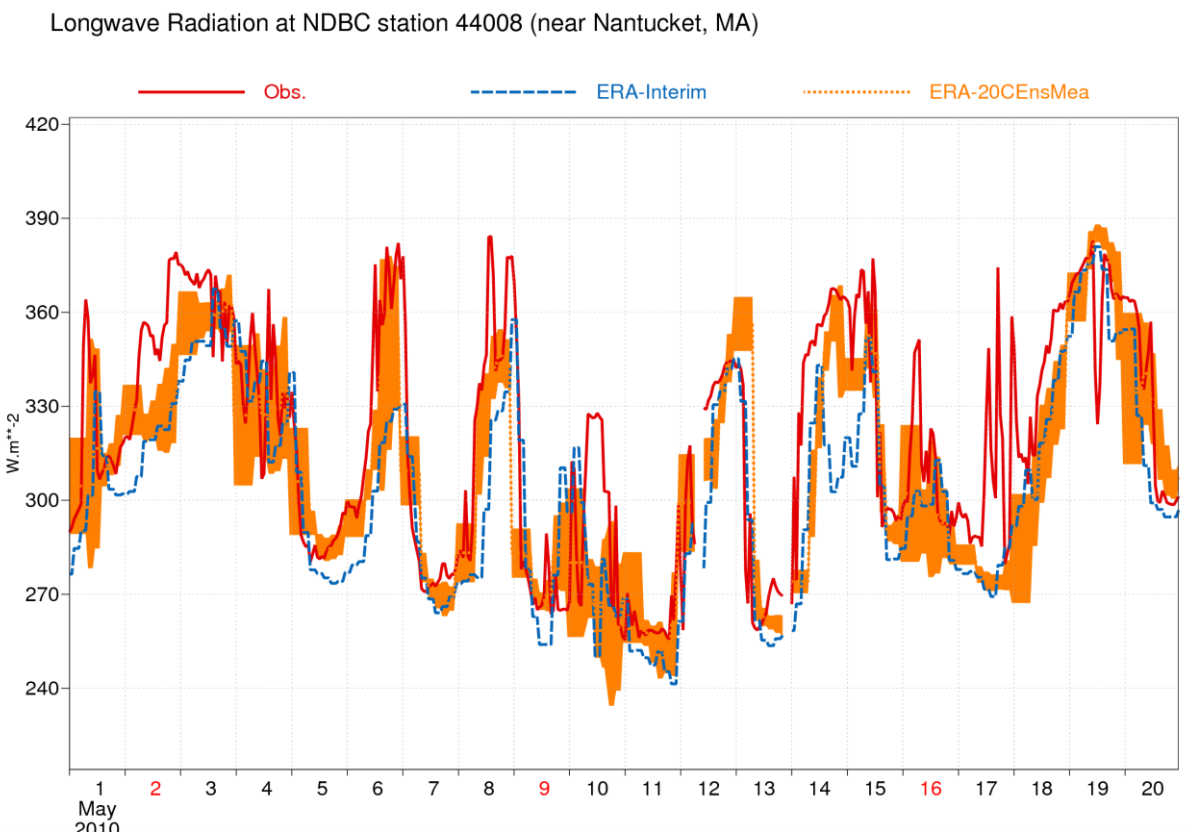

Figure 3: Time-series of in situ hourly observations of downwelling longwave radiation at a buoy with estimates from two reanalyses, ERA-Interim and ERA-20C ensemble (ensemble mean +/- ensemble spread shown in shading). All estimates valid at the observation date and location. See text for details.

One then also notices systematic offsets between the reanalysis estimates and the observations. If averaged on a larger domain and a wide variety of sites, such differences may be more difficult to interpret, but here one can further look at the correspondence by showing differences on the diurnal cycle. Figure 4 shows such averages, and there the offsets between the various estimates appear more readily. One notes also larger amplitude in the diurnal cycle in the observations than in the reanalysis estimates.

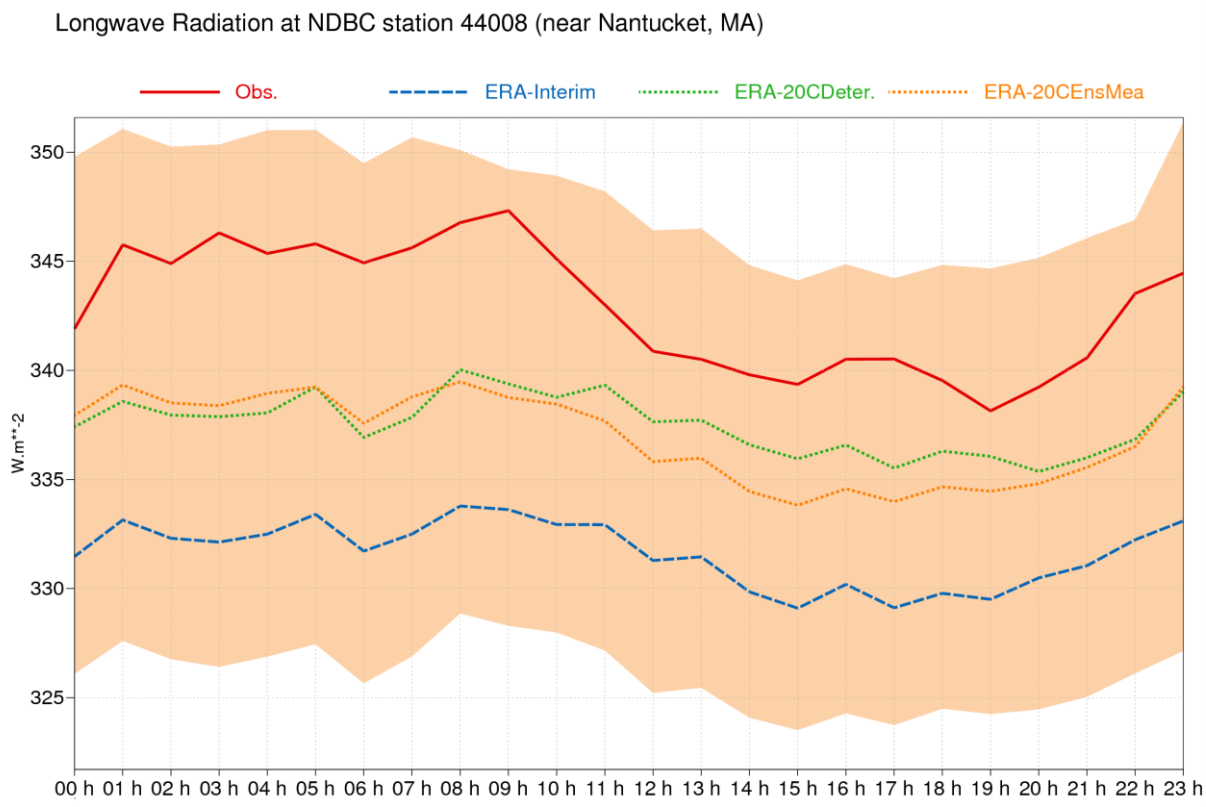

Figure 4: Average observations of downwelling longwave radiation for a buoy in May 2010 over the course a day (x-axis shows hour of the day), and various reanalysis estimates (for ERA-Interim, single-member ERA-20C deterministic, and ERA-20C 10-member ensemble with shading showing +/- the ensemble spread)

Comparison with satellite observations

The Climate Monitoring Satellite Application Facility (CM-SAF) has reprocessed the brightness temperature record collected by the Special Sensor Microwave Imager (SSM/I) onboard the Defense Meteorological Satellite Program (DMSP) satellites. The resulting Fundamental CDR (FCDR) is available from http://dx.doi.org/10.5676/EUM_SAF_CM/FCDR_SSMI/V001. This CDR is compared here with estimates from two reanalyses. The comparison process maps the reanalysis four-dimensional fields of temperature and humidity to the observations locations and times (using the same procedure as described earlier for in situ observations), and then applies a satellite simulator (here, the NWP-SAF radiative transfer model RTTOV v11, available from http://nwpsaf.eu/deliverables/rtm/index.html). Figure 5 shows the result from such comparison. ERA-20C serving as an independent comparison here, since it did not assimilate any SSM/I observations, the top panel row suggests that the inter-sensor calibration computed by the CM-SAF algorithm removes several time breaks and shifts in the observation record. The second row shows much reduced

differences between SSM/I observations and ERA-Interim than between SSM/I observations and ERA-20C. This comes from the assimilation of SSM/I observations in ERA-Interim. However, the differences with respect to ERA-20C are still reasonable, below 4K standard deviation, and much smaller that the variations within the observation record itself (see last row which shows standard deviations within the observations on the order of 12K). The small reduction of differences with respect to ERA-20C over time between 1997 and 2009, valid for all satellites, probably comes from the improvement in the quality of ERA-20C over (southern) oceans (improved wind speeds would yield improved microwave emissivities and thus better agreement with satellite observations).

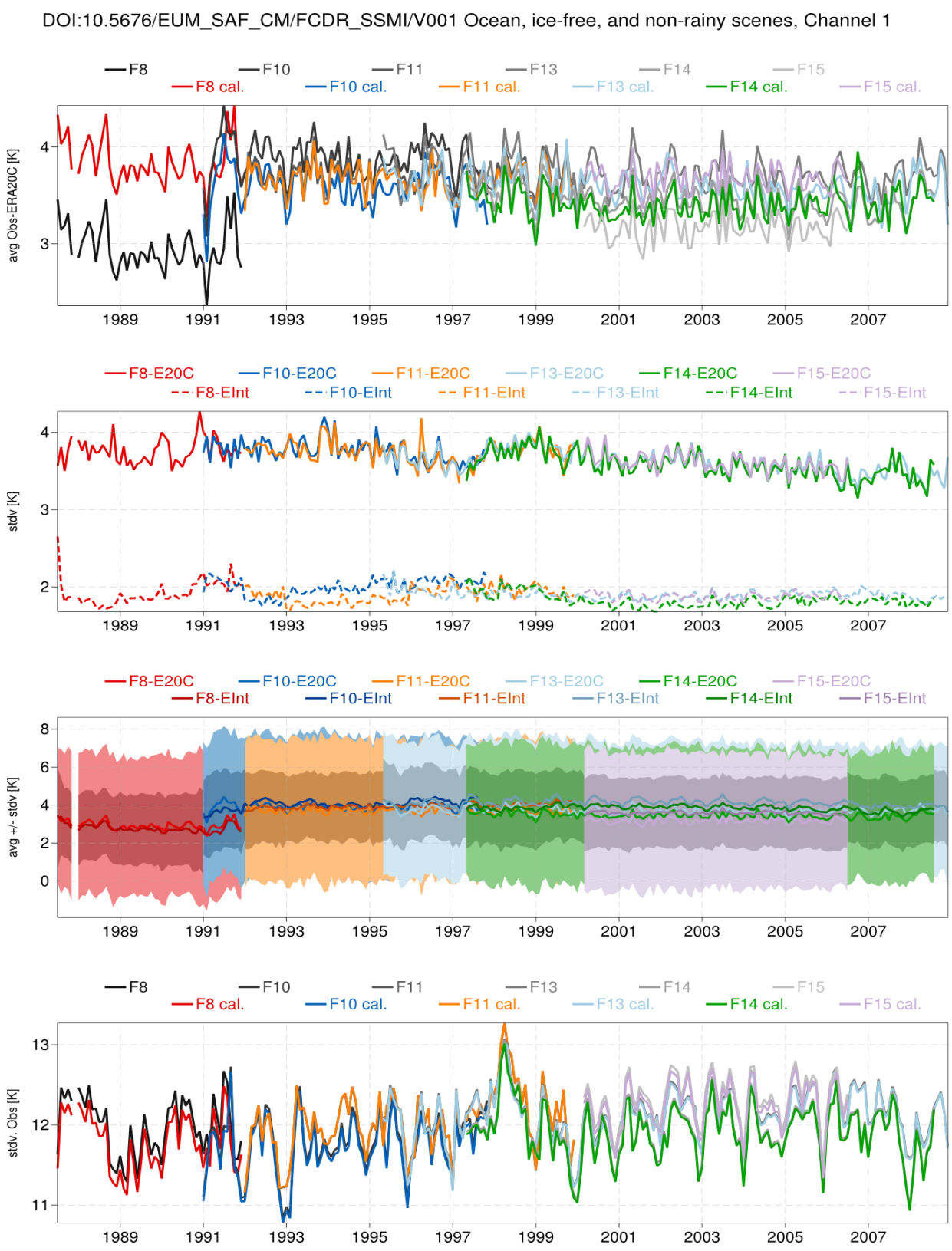

Figure 5: Top row: Monthly series for CM-SAF SSMI FCDR channel 1 (19 GHz, horizontal polarization), compared with ERA-Interim (EInt) and ERA-20C deterministic E20C). Scenes retained here only

include ocean, without sea-ice, and believed to be rain-free. Top row shows mean of the differences with respect to ERA-20C (for observations corrected by application of the inter-sensor calibration in color, and without such calibration, in grey scales). Second row shows standard deviation of differences for uncalibrated observations with respect to ERA-20C and ERA-Interim (dashed curves). Third row shows average +/- standard deviation of differences with respect to ERA-20C (brighter tones) and ERA-Interim (darker tones). Bottom, fourth row shows the standard deviation of the observations (within the month and domain).

The main needs for Comparisons between gridded reanalysis products and third-party in-situ/satellite-swath observation-based CDRs are summarized in Table 4 below.

Table 4: Main needs and considerations for Comparisons between gridded reanalysis products and third-party in-situ/satellite-swath observation-based CDRs

Identified Need	**Purpose**	**Who should be involved?**	**At what stage?**	**Further comments**
Comparison between gridded reanalysis products and in-situ/satellite-swath observation-based CDRs	Identify/explain differences in datasets at the level of specific Essential Climate Variables or at the level of upstream observational parameters (radiances, brightness temperatures etc) Provide guidance on the interpretation/use of different products (both reanalysis and observation-based). Provide feedback to providers of reanalysis products and to providers of observational datasets.	Dataset providers, dataset users	Prior to dataset release (production quality control), and after release (dataset evaluation and feedback)	Is of high interest to users
typically addressing topics such as …				
Representativeness Temporal stability, spatio-temporal scales	To assist meaningful interpretation by the users	CDR Dataset providers, reanalysis providers	Information required at application	Can be affected by variations in satellite simulators (i.e. the observation operators employed

which can be interpreted with confidence				to map reanalysis products to the parameters measured by satellite instruments)
systematic differences, relative biases	To assist interpretation of climatologies (frequency distributions, histograms)	Crowd sourced	Information required at application	Bias correction
inter-dependence of the datasets being compared	To assist scientifically sound conclusions	Observation providers	Information required at application	A priori content in both reanalysis and in satellite retrievals have to be considered.

Co-ordinating functions	Collate/disseminate the information from bi-lateral and multi-lateral comparisons. Review/update the topics to be addressed. Co-ordinate the development/use of standard tools, in particular the satellite simulators (observation operators) needed to map reanalysis products to the parameters measured by satellite instruments.	Co-ordinating body	On an on-going basis, updating when new comparisons are received/updated.	

4. Inter-comparison between reanalysis products

Users are attracted by the highest given nominal resolution of the reanalysis gridded fields, possibly guided by the descriptive comparison. Inter-comparison between gridded reanalysis fields however can help pointing to the inherent feature resolution of the different reanalyses.

4.1 Global versus global reanalysis

Considering the feature resolution, parameter fields from global reanalyses can be directly compared with each other. Good agreement might not necessarily indicate a reduction in uncertainty, but could be related to the sameness of methodology or technical parameters (e.g., choice of error covariances). Any dis-agreement is easier to interpret as a sign to raise alertness that the observations might be imperfect (biased), or the observations do not constrain the model sufficiently, or model errors might play a role.

4.2 Regional versus global reanalysis

In ideal circumstances, a regional reanalysis will inherit the bulk of large-scale variability from the global reanalysis that constitutes the boundary conditions. The greater the consistency between the global reanalysis and the regional system, the more potential there is for benefit from regional reanalysis (with its meso-scale modelling plus regional scale data assimilation, higher resolved topography and surface boundary). Real-world circumstances are not ideal. Thus, the uncertainty of the regional reanalysis is a combination of the uncertainty of the global reanalysis and subsequent modifications/additions by the regional assimilation process (which introduces structure on smaller scales). The regional assimilation process can similarly modify or add to the temporal variability present in the global reanalysis.

One aspect worth examining concerns the extent to which long-term variability in the regional reanalysis is dictated by features resolved in the global reanalysis, or whether the regional assimilation process modifies this significantly. With the higher resolution of the regional system, local-scale extremes should be better resolved (heat waves, daily precipitation extremes) as regional effects on temperature, precipitation, snow cover, clouds, surface winds are modelled and constrained by more regional observations. Regional reanalysis might be expected to have higher applicability than global reanalysis where complex topography is important (sea breezes, alpine processes) or smaller meteorological scales (e.g. storms) and applications with special interest in extremes (e.g., hydrology). It is not trivial to show this, as it involves a subjective sorting which observations are considered to be representative for

regional (but not local) features. For those, regional reanalysis should show a better fit. Of course this is not a fair comparison of the gridded fields if such regional features (e.g., daily cycle) are not resolved in the global reanalysis (as can be seen in the descriptive comparison). For this and other reasons, it is important to complement regional-global intercomparison of gridded fields with comparison of internal metrics (Section 6.2).

4.3 Regional versus regional reanalysis

How regional renanalysis compare will depend on whether they use different global renanalyses as boundary condition. This aside, the categories discussed in this document can be applied likewise.

4.4 Remarks

The reader may be surprised that the complexity of this category (intercomparison of reanalysis gridded products) is rated as moderate to conduct and moderate/difficult to interpret. Although the computation of differences and subsequent statistical processing is relatively simple, we raised the conduct complexity rating to moderate because of other considerations: e.g. some attention must be given to differences between the reanalysis grids and possible effects of interpolations from one to another. The interpretation complexity is deceptively difficult, because each reanalysis dataset consists of a large number of ECV products that are inter-related through the reanalysis process. Interpretation thus relies on sound knowledge of how the different reanalyses have been produced and of the role played by observations in the data assimilation component of the assimilation process (which is available from comparison category 5, internal metrics).

The main needs for Comparison between gridded fields from different reanalysis products are summarized in Table 5 below.

Table 5: Main needs and considerations for Comparison between gridded fields from different reanalysis products

Identified Need	**Purpose**	**Who should be involved?**	**At what stage?**	**Further comments**
Comparison between different reanalysis products	Identify/explain differences in reanalysis products at the level of specific Essential Climate Variables. Provide guidance on the interpretation of different resolutions.	Dataset providers, dataset users	Prior to dataset release (production quality control), and after release (dataset evaluation and feedback)	
typically addressing topics such as ...				
regional variations		Dataset providers, and crowd sourced		users could add to the information data providers give
representation of extreme events/values		Dataset providers, and crowd sourced		Is of high interest to the users

Co-ordinating functions	Collate/disseminate the information from bi-lateral and multi-lateral comparisons. Review/update the topics to be addressed.	Co-ordinating body	On an on-going basis, updating when new comparisons are received/updated.	

5. Thematic comparison

Such type of comparison follows from the earlier three as requiring broader knowledge of reanalysis. One of the challenges for the future is to ensure that participants do not bring breadth at the expense of depth. Thematic comparison consists in evaluating how well a basket of reanalysis products fares when applied to understand a particular problem. One may separate between at least two categories there, climate service user applications and scientific applications of reanalysis (to understand natural processes). As there are thousands of reanalysis users in the first category, one may refer to this first type of comparisons as 'crowd comparisons'.

5.1 Climate service user application comparisons (or crowd comparisons)

Whenever a user of reanalysis products downloads data for his/her application, one of the first actions is to compare these products with other products already in his/her possession, to see how well they agree.

A nicely organized example of this is given by Gil Lizcano, Research and Development Director of a wind energy company, in his presentation "Some guidelines to infer and assess wind climate variability uncertainty from modelled time series" at the European Wind Energy Association Resource Assessment Workshop 2013 (http://www.ewea.org/events/workshops/wp-content/uploads/2013/06/EWEA-RA2013-Dublin-1-3-Gil-Lizcano-Vortex.pdf). One of the first steps in his approach is to compare the reanalysis product with wind mast observations, to assess whether the reanalysis data are fit for purpose.

Scaled over the thousands of users of reanalysis data, there is thus a great amount of distributed knowledge in terms of how well reanalyses compare for each application. Yet, there is no integrated platform to collect all this knowledge in a systematic fashion.

The website reanalyses.org provides an extremely valuable forum, but being developed and updated by the reanalysis producers, it may lack the engagement of non-expert users. This is especially true if they do not feel comfortable enough to ask simple questions and receive simple answers.

In the comparison procedure to be proposed to the community, one may thus include as part of the climate services user desk a platform to collect all these 'crowd comparisons'.

The main needs for Thematic Comparison of reanalysis products (sub-category: climate service user applications) are summarized in Table 6 below.

Table 6: Main needs and considerations for Thematic Comparison of reanalysis products (sub-category: climate service user applications)

Identified Need	**Purpose**	**Who should be involved?**	**At what stage?**	**Further comments**
Thematic Comparison - climate service user applications	Document the effectiveness of reanalyses when used as input to sectorial climate-services. Share experience of using reanalysis products in climate-service applications. Provide feedback to providers of sectorial climate services and to reanalysis providers.	Developers and users of sectorial climate services	After release of reanalysis dataset (dataset evaluation and feedback)	Developers and users of sectorial climate services will need a basic level of understanding of non-thematic comparisons.
typically addressing topics such as …				
suitability of reanalyses as input datasets for sectorial applications		Successful users		Is of high user interest
placing reanalysis uncertainty within the overall uncertainty of the sectorial application		Successful users, scientific users		

Record the benefit of reanalysis over traditional approaches	Identify potential for improvement (both traditional approaches and reanalysis)	Successful users	After demonstration of successful use cases	Most convincing way to attract new users

Co-ordinating functions	Mobilize the sectorial applications community on a sustained basis Collate/disseminate the information from a diverse range of applications.	Co-ordinating body		

5.2 Natural processes representation comparison

This is an area where further development would be extremely valuable to an operational Climate Service.

To date, a sizeable uptake of reanalysis products has been for the purposes of understanding of natural processes, largely by the scientific research community. The breadth of this community encompasses atmospheric process studies on weather timescales (hours to days) as well as climate timescales (years to decades). They communicate their findings via the traditional scientific routes, namely peer-reviewed literature and conference presentations, often supplemented by internal institutional reports.

Reanalysis producers also contribute to the comparison of natural process representation, often in collaboration with the scientific users, and using the same communication channels. The reanalysis reference papers for MERRA, 20CR, and ERA-Interim (see Table 1) featured discussions of the representation of the following natural processes:

- stratosphere-troposphere exchange (by showing that the stratospheric tape recorder is represented in the products)
- Brewer-Dobson circulation (by comparing age-of-air in the stratosphere from transport simulations with high-altitude aircraft observations)
- Quasi Biennial Oscillation (QBO) and Semi-Annual Oscillation (SAO) (by showing that these are found in the products)
- global conservation properties (by computing various water, mass, and energy budgets and assessing the overall imbalances)
- surface fluxes over the ocean (by using these to drive an ocean model and check the quality of its forecast)
- Madden–Julian Oscillation (by showing that this oscillation is found in the products)
- extreme weather events (for example for the 1987 European storm or tropical storms)
- frequency of weather events of a given type (for example frequency of blocking events over Europe)
- regional climate indices (for example the Pacific Walker Circulation (PWC), the North Atlantic Oscillation (NAO), and the Pacific North America (PNA), comparing them with observation-based estimates).

Many of the existing comparisons of natural process representation are essentially bi-lateral in nature, i.e. they compare two datasets. Some are multi-

lateral but in a limited sense, in that the number of datasets compared is more than two but does not cover the full range of available datasets.

Further enhancement of current activities is arguably critical to the success and usefulness of operational Climate Services. There is a clear and pressing need to capitalize on the considerable expertise that already exists for bi-lateral process comparison, and to enhance the capability to the level of comprehensive intercomparisons that would ensure the robustness of the datasets underlying the Service.

Having recognized the importance of building such a capability, the world-wide scientific and reanalysis-producer communities have already taken steps to collaborate on its development. It was a prime motivation for the inception of the SPARC Reanalysis Inter-Comparison project (S-RIP, http://s-rip.ees.hokudai.ac.jp/index.html). The clear focus on comparison of climate-relevant natural process representation is reflected in S-RIP's 4-year workplan and the outline structure of its Final Report. Chapters 3 to 11 include, respectively: 3: Climatology and Interannual Variability of Dynamical Variables, 4: Climatology and Interannual Variability of Ozone and Water Vapour, 5: Brewer–Dobson Circulation, 6: Stratosphere–Troposphere Coupling, 7: Extratropical Upper Troposphere and Lower Stratosphere, 8: Tropical Tropopause Layer, 9: Quasi-Biennial Oscillation and Tropical Variability, 10: Polar Processes, 11: Upper Stratosphere and Lower Mesosphere.

One member of the Core-Climax team (David Tan, ECMWF) is on the S-RIP preparation team that oversees the activities and direction of S-RIP. This provides opportunities to co-ordinate/align S-RIP with the future needs of operational Climate Services. In this way, Core-Climax has been able to encourage the development and sharing of common data analysis tools, and to transfer knowledge to the scientific community about reanalysis internal metrics (this document, Chapter 6). Such knowledge transfer contributes to climate-service capacity building by increasing the pool of people who have both breadth and depth of reanalysis understanding.

One of the major risks in developing and sustaining the climate-service capacity is that the participation of the international scientific community is currently highly dependent on research funding. This could be mitigated by consolidating these activities as an element of Evaluation and Quality Control within an operational Climate Service.

The main needs for Thematic Comparison of reanalysis products (sub-category: natural process representation) are summarized in the table below.

Table 7: main needs for Thematic Comparison of reanalysis products (sub-category: natural process representation)

Identified Need	**Purpose**	**Who should be involved?**	**At what stage?**	**Further comments**
Thematic Comparison - natural process representation	Establish the fidelity of reanalyses to represent the Earth-system. Provide guidance on the interpretation/use of different products (both reanalysis and observation-based). Provide feedback to providers of reanalysis products and to providers of observational datasets.	Dataset users with scientific expertise, dataset providers	Prior to dataset release (production quality control), and after release (dataset evaluation and feedback)	
typically addressing topics such as ...				
suitability of reanalyses as reference datasets for climate model validation				
confidence in quantitative reanalysis-based estimates of climate variability	Adding information on which spatio-temporal scales can be interpreted			High user interest
Co-ordinating functions	Mobilize the scientific community on a sustained basis Collate/disseminate the information from bi-lateral and multi-lateral comparisons.	Co-ordinating body		

6 Internal metrics comparison

Each system that produces reanalyses or CDRs generates internal analysis metrics for its own use. Although these are usually not formally published as products, they are extremely valuable to compare as they are intimately tied to essential product features.

Across reanalysis and observation-based CDR products, one can distinguish between 3 classes of internal metrics.

6.1 Internal metrics based on differences between a prior estimate and new estimate

Such metrics are a first measure of temporal discontinuity. In reanalysis, systematic analysis increments point to recurrent corrections that aim at correcting either a model bias or an observation bias; if these vary over time, they can introduce spurious low-frequency signals. One could thus claim that a 'good' product should feature, *on average*, zero increments. (However, the difficulty lies in defining over what space that average should apply, as there could sometimes be very good reasons why a particular average over a given sub-space should feature a non-zero increment so it would cancel out an average of opposite sign in another sub-space. Taking this argument to the limit, the sizes of sub-spaces where average increments would have to be zero would become infinitesimally small, and thus all increments would have to perfectly zero, in which case there would be no step-wise changes in time and the products would feature stationary time-series, or time-series generated by a perfect model.)

Figure 6 shows a comparison between time-series of mean analysis increments for 3 reanalyses: (a) ERA-Interim, (b) ERA-20C ensemble, and (c) a deterministic re-run of ERA-20C, for temperatures across the vertical. Descriptive comparison indicates that (a) features a different vertical resolution different than (b) and (c), and that (a) uses upper-air and satellite data, whereas (b) and (c) only rely on observations at the surface. Also, the figure titles indicate that the increments are computed at different time-steps (owing to the differences in the assimilation schemes). These three points already explain a large part of the differences.

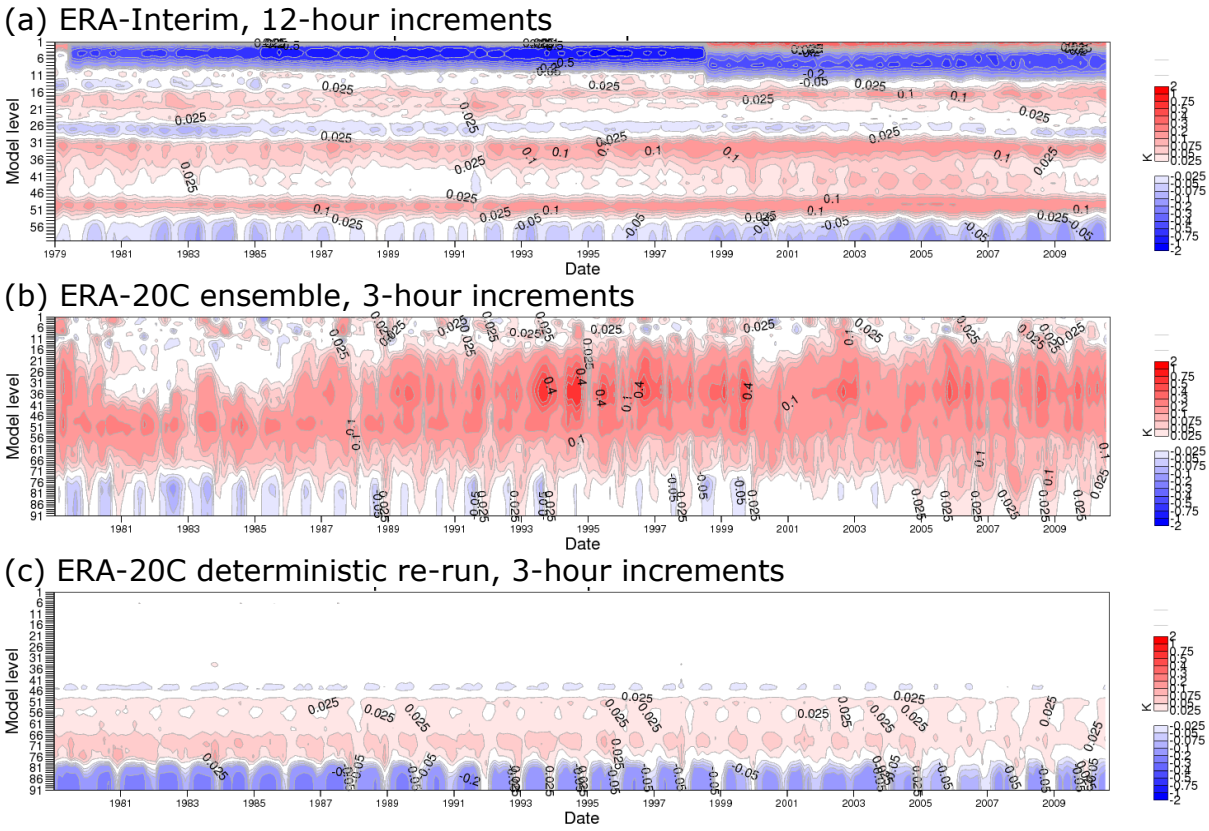

Figure 6: Monthly average of temperature analysis increments in 3 atmospheric reanalyses (2 of which are described in Table 1), from January 1979 until September 2010. Color scales are identical in all figures above and show absolute values mainly in the range between -0.5 and 0.4 .

These metrics are also a measure of observation-added value in the product generation system. A system which sees no observation for some time would retain prior properties but would more likely lack realism.

Figure 7 shows the RMS of temperature analysis increments in 3 reanalyses. The two reanalyses that only use surface observations feature, as expected, the smallest increments at the top levels (below 0.2 K), whereas the reanalysis which uses satellite data features large increments there. There again a prior knowledge of the descriptive comparison is important to understand what the comparison shows.

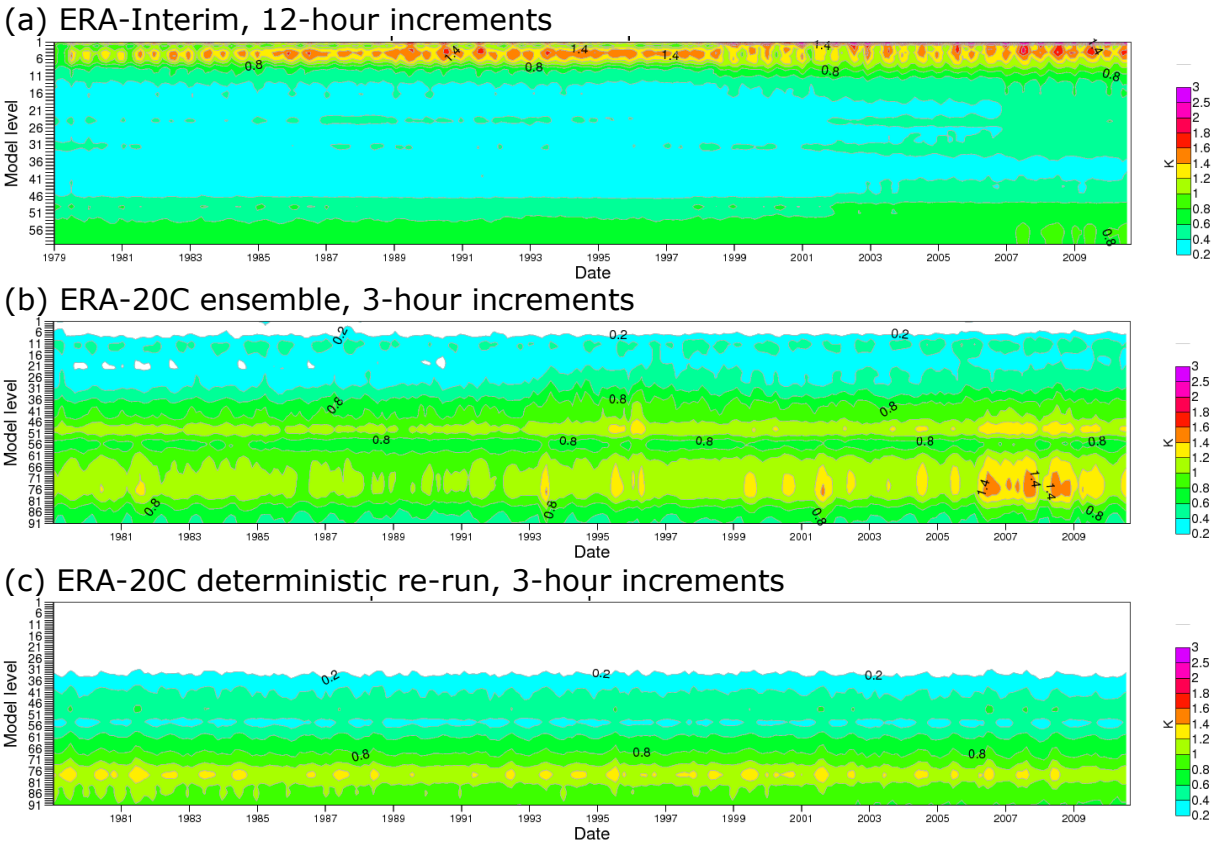

Figure 7: Same as Figure 6, but showing monthly average RMS of temperature analysis increments. Color scales are identical in all figures, and show only values larger than 0.2 K.

Finally, metrics based on differences between a prior estimate and new estimate, when taken over a longer time-range, such as a medium-forecast, can provide a validation to the extent that the 'new estimate' is superior in quality and can be taken as a reference. This is commonly referred to as forecast scores by the Numerical Weather Prediction (NWP) community. Figure 8 shows an example of such metrics, for two reanalyses, for predictions of geopotential height at 500 hPa for days 1, 3, 5, and 7 (all verified against ERA-Interim analyses). This figure shows that the 1-day forecasts from ERA-20C are of similar quality to the 3-day forecasts from ERA-Interim (similarly: 3-day from ERA-20C similar to 5-day from ERA-Interim, and 5-day from ERA-20C similar to 7-day from ERA-20C). Of interest, one also notices that the forecasts from ERA-20C improve drastically over the Southern hemisphere extra-tropics in the last decade, probably due to the large increase in the number of observations from drifting buoys in the southern oceans.

Forecast scores *Diagnostics (short-term or medium-term forecast based on last reanalysis field verified against next reanalysis)* → forecast scores *(Tmin, Tmax, Precip skills like ETS etc.). WMO publishes standards for verification:*

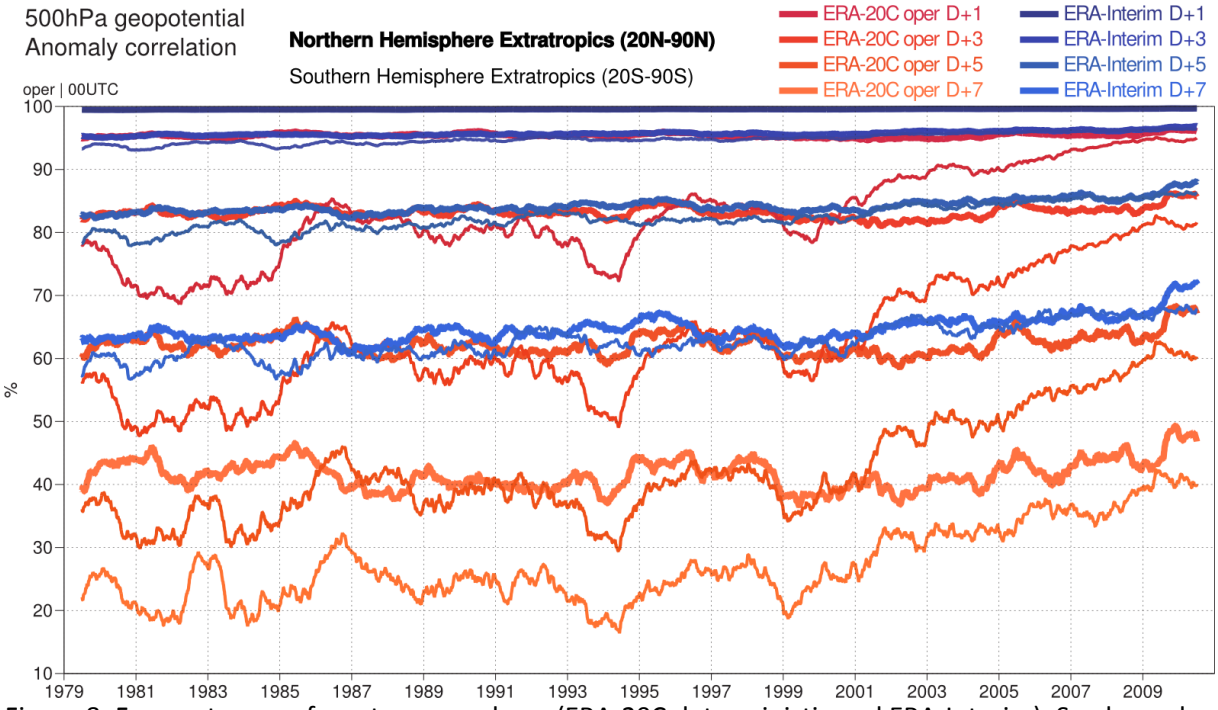

Figure 8: Forecast scores from two reanalyses (ERA-20C deterministic and ERA-Interim). See legend at the top for description of the various curves, and text for details.

6.2 *Internal metrics based on differences between new information (observations) and past information (e.g. from persistence or from a forecast model)*

These metrics, called innovations in data assimilation, are a measure of quality with respect to the observations. Two important caveats apply when interpreting these metrics: some information from these observations may already somehow be included in the product (through error correlation or biases), and this comparison is only of use to the extent that these observations are not characterized by gross errors.

Feedback statistics terminology	
o	observations
x	model values
b	a priori or background estimate of the state vector
a	analysis
a - b	analysis increments
o - b	background departure *or* first-guess departure *or* innovation
o - a	analysis departure *or* residual

Since the most valuable data are often ingested into the assimilation system, stable time series of independent observations covering a long time-period are scarce. A way around this obstacle is to use the observations, but not to compare against the reanalysis fields, but against the free forecasts (or background fields) which were started from the re-analysis a few hours earlier. These so-called *feedback statistics* can be routinely produced by the data assimilation system, and relate assimilated observations, so-called free forecasts (i.e., background fields), analysis results or analysis increments to each other. They yield valuable additional information, e.g on upper error bounds of the analysis error or on systematic changes in increments due to biases in observations or model (or both), indicating the deficiencies in the system. Favourable statistics may show that the frequency distribution and time series of observed and reanalysed parameters are matching. Thus, it is potentially of high practical value for the user, to take into account the results of feedback statistics.

User can interpret these feedback statistics (see inbox above), as comparison of the reanalysis fields against chosen observations yielding an estimate on reanalysis uncertainties.

Figure 9 shows an example for observations of surface pressure assimilated in one reanalysis. Over time, the background departures become smaller, which here is explained by the reanalysis estimates becoming more accurate (as they are constrained by an increasing amount of observations). The impact of the bias correction (shown by the shading) also reduces over time, suggesting that the today's observations are better and more often calibrated than earlier ones. The increases in differences during the two world wars coincide by degraded observation network and practices. Between 1899 and 2009, the reduction in RMS of background departures exceeds a factor 2, and the reduction in mean differences is even more significant.

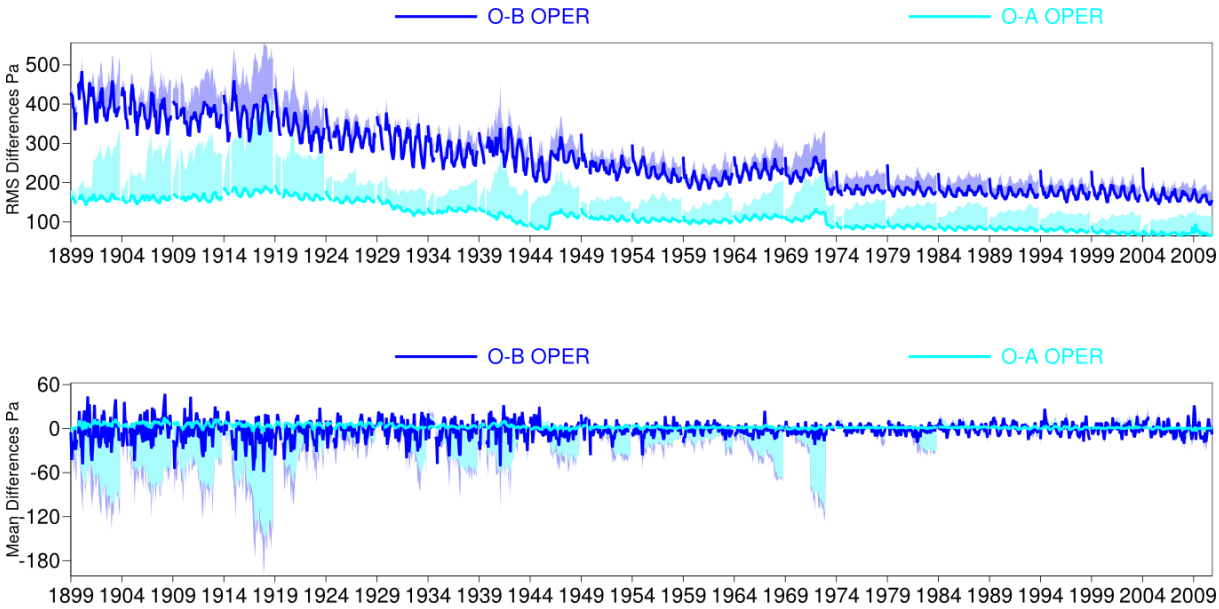

Figure 9: Time-series of Root Mean Square (RMS) differences (top row) between assimilated observations of surface pressure and ERA-20C deterministic background (o-b) and analysis (o-a). Second row shows mean of differences. Shading is bounded by two curves: curve closer to zero is from differences after application of bias correction, and curve farther from zero is from differences before application of the bias correction.

The analysis of feedback statistics can also be a powerful means when it comes to the comparison of global and regional reanalysis (see also Section 4.2). Figure 10 compares ERA-Interim to the COSMO-based high-resolution HErZ regional reanalysis (University of Bonn, Germany) for one radiosonde station. The figure shows background departures (o-b) and analysis departures (o-a) of wind speed from both reanalyses at the Lindenberg radiosonde station (monthly mean). The other German stations show a similar behaviour (not shown).

The choice of comparable metrics depends on how the assimilations are set up. This is illustrated below with a quick plot of reanalyses output in Figure 10 (left, not recommended). Compared at 12 UTC with ERA-Interim (black lines), the HErZ regional reanalysis (red lines) seems to have smaller RMSE of o-b as well as o-a at the 1000, 925, 850 and 700 hPa levels, which would be a desirable behaviour. However, this can be explained by the different forecast lead times of both systems and the different data assimilation windows of the initial conditions. The forecast lead time for the 12 UTC COSMO background is 6 hours, whereas for ERA-Interim it is more like 12 hours (3 hours initial condition error + 9 hours forecast error). In addition, we have an effect from the difference in assimilation windows of both analyses which provide the initial conditions for the free forecasts. The COSMO 12 UTC background is a 6h free forecast, starting from the 6 UTC COSMO-reanalysis, the latter had assimilated the 6 UTC soundings. The ERA-Interim 12 UTC background does "not know about" the 6 UTC soundings, as

its forecast was started from the 0 UTC ERA-Interim reanalysis which had not assimilated the 6 UTC soundings. More comparable metrics are shown in Figure 10 (right). Here, COSMO has a forecast length of 6 hours, the ERA-interim background is comparable to a 6 hours forecast length (with 3h initial error plus 3h analysis error), and the sounding data that have influenced the initial conditions for the background forecast are more comparable. With the more comparable metrics, the perceived advantage of the regional reanalysis is not apparent (at least for the statistical sample shown here).

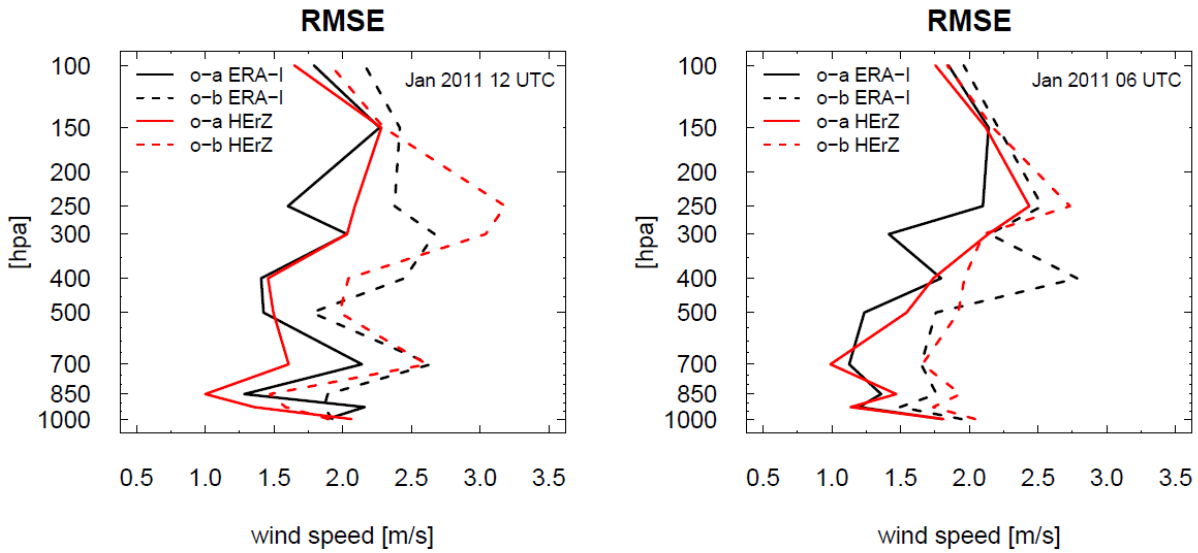

Figure 10: Root mean square errors of background and analysis departures for radiosonde observations of wind speed at standard pressure levels in January 2011 at station Lindenberg, shown for ERA-Interim and HErZ regional reanalysis (produced by the University of Bonn). Left: For 12 UTC, different forecast lead times have been compared (not recommended) Right: for 6 UTC, the metrics are comparable (recommended).

As expected from the fact that HErZ is driven by ERA-Interim as boundary condition, the figure reveals a similar performance of both reanalysis systems. Generally, these kind of statistics might help the user to interpret the benefits that can be expected by regional reanalyses, and investigate the performance for particular times or places of interest. Further metrics like histograms (representation of extreme values) and daily cycles might also be regarded in this context.

Generally, a comparison of o-b is more meaningful than o-a, because the analysis depends on the observations in a degree which varies over the different assimilation systems.

6.3 *Internal metrics characterizing the error estimates produced by the system*

Such error estimates include for example bias corrections or adjustments, ensemble spread, random error estimates.

These should be taken with precaution as these estimates may of course be incorrect (in sign or by up to an order of magnitude), but still represent a condensed summary of the best knowledge about uncertainties in the product generation system. If these are found to be incorrect on some time-scales, then one should not expect to have such uncertainties being necessarily properly corrected in the products.

Figure 11 shows an example of observation bias correction estimates for two reanalyses, using about the same observations as input. One notices in the time-series jumps or breaks every 20th year for ERA-20C ensemble, and every 5th year for ERA-20C deterministic. These correspond to the length of each production stream and indicate that in the first year of each product one finds remnants of the spin-up the bias correction scheme. Owing the large disparity between the two estimates, one may postulate that either one or both of them is/are quite incorrect.

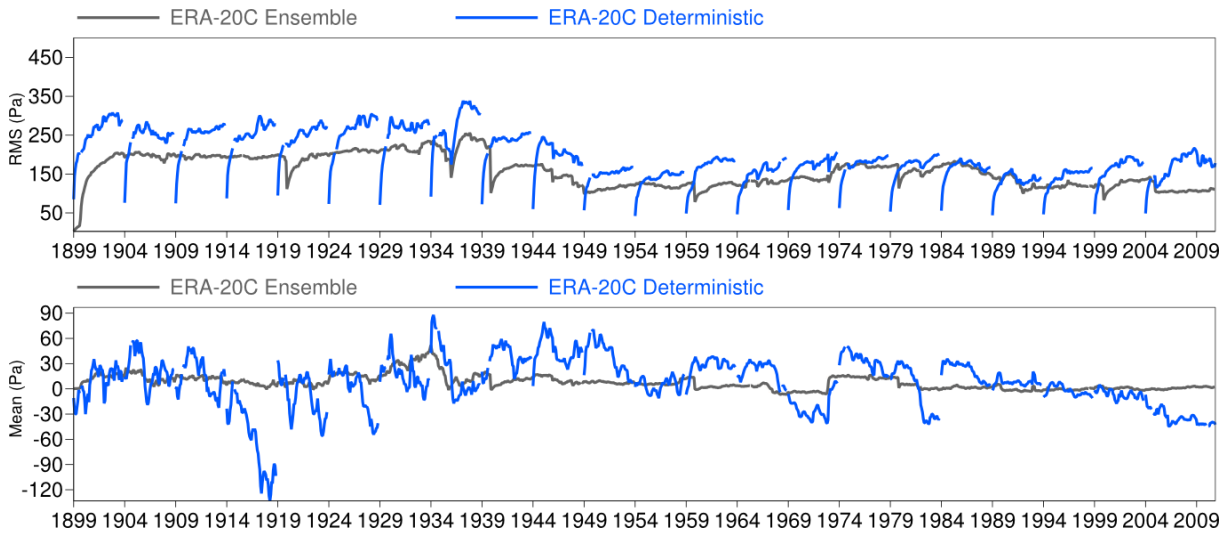

Figure 11: Time-series of bias corrections for observations of surface pressure assimilated by two different reanalyses, ERA-20C ensemble and ERA-20C deterministic. First row shows average, second row show RMS.

Regarding the estimates of random error in the reanalysis products, the ensemble technique can provide a tool to estimate this uncertainty. The ERA-20C reanalysis uses such a technique and the spread of the 10 members is shown in

Figure 12 added to the observation error. The figure also shows, for verification, the expected RMS difference in terms of background departures. If the background errors (hence ensemble spread) and observation errors were correct, then the two curves would agree perfectly. Obviously the match is not perfect, but still allows one to draw qualitative conclusions regarding which of the estimates is probably incorrect. More about this can be found in ERA Report Series 14 (available from http://www.ecmwf.int).

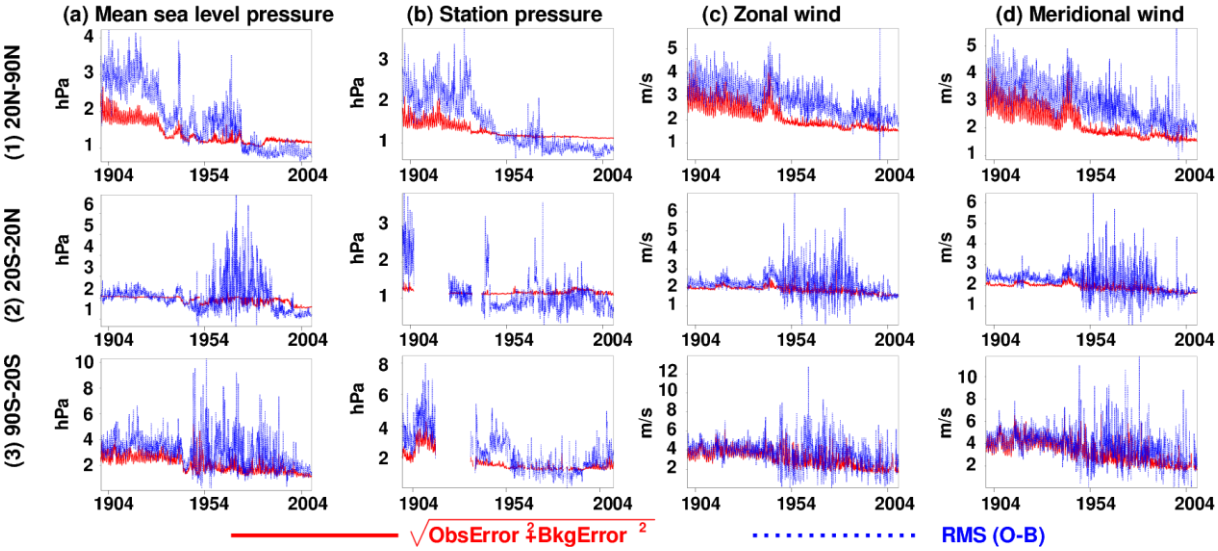

Figure 12: Time-series of expected total error (in red) and observed background departure RMS (in blue) for observations assimilated in ERA-20C ensemble, by geophysical variable (each column) and by geographical domain (each row).

6.4 *Limitations and difficulties implementing such comparisons of internal metrics*

As all results from such comparisons draw from internal metrics, they do not represent independent information. Consequently, all interpretations require stepping back and considering to what extent these metrics can be trusted.

Second, access to these internal metrics is not straightforward and even though great effort may be spent in trying to encourage everyone to use the same standards (e.g. for forecast scores: WMO standards for forecast verification), there remains some room for individual participants to remove some 'outliers' from the set of input they consider.

Finally, owing to the fact that these metrics are tied to each particular system that generates them, it is sometimes impossible to enforce the same baseline of

comparison: for example, the forecast lead time of the background considered by two reanalyses with different data assimilation cycling may differ by design.

Taking into account these considerations, the category of internal metrics comparison has been given a provisional rating for complexity of difficult to conduct and moderate to interpret - with recognition of the scope to re-rate in future as simple/moderate to conduct and simple/moderate to interpret. The scope for lowering the complexity to conduct stems from the potential to improve exchange and access to the internal metrics, and to support these with tools for statistical analysis. The scope for lowering the complexity to interpret hinges on building the capacity of the community to understand the practical implications of internal reanalysis metrics. Such understanding underpins proper interpretation of the preceding categories of reanalysis comparison, and so capacity-building targeting the understanding of internal reanalysis metrics is arguably a priority for a Climate Service.

The main needs for Internal-metrics Comparison of reanalysis products are summarized in Table 8 below.

Table 8: Main needs and considerations for Internal-metrics Comparison of reanalysis products

Identified Need	Purpose	Who should be involved?	At what stage?	Further comments
Internal Metrics Comparison	Establish the underlying consistency of reanalysis products for representing the Earth-system. Provide guidance on whether reanalysis products can be regarded as "climate-quality" datasets. Provide feedback to providers of reanalysis products and to providers of observational	Reanalysis producers, providers of observational datasets.	Prior to dataset release (production quality control), and after release (dataset evaluation and feedback)	Many users are not aware that these metrics exist or which information can be drawn from it.

	datasets.			
typically addressing topics such as ...				
temporal discontinuities, systematic errors, reliability of long-term trends, reliability of error-estimation				Is of high user interest.

Skilled personnel to conduct such comparisons				Requires in-depth understanding of different reanalysis systems and their internal metrics, and how to account for such differences when conducting intercomparisons
Capacity-building	Transition comparison of internal reanalysis metrics from research activity to operational activity.			The requirements for capacity-building include mechanisms to exchange internal metrics between reanalysis producers, and the development software tools to perform comparisons.

Co-ordinating functions	Facilitate exchange of internal metrics between reanalysis producers Collate/disseminate the information from bi-lateral and multi-lateral comparisons.	Co-ordinating body		

7 Concluding remarks

The present document presents a set of procedures for comparing reanalyses, and comparing reanalyses to assimilated observations and CDRs. To do so, five categories of comparisons are identified, accompanied by two complexity ratings. The first rates the complexity of conducting the procedure (simple, moderate, difficult), and the second rates the complexity of interpreting:

1. descriptive product comparison (simple, simple)
2. comparison with third-party observation-based CDRs (moderate, moderate)
3. inter-comparison between different reanalyses (moderate, moderate/difficult)
4. thematic comparison (difficult, difficult)
5. internal metrics comparison (difficult, moderate)

The current document concentrated on technical descriptions of these procedures, drawing on current best-practice. It has also identified and documented some areas in which best-practice would need to evolve to transition reanalysis comparisons from the level of research activities to operational Climate Services. The service-related issues raised here will be consolidated with findings from other Core-Climax workpackages/tasks, in a subsequent Core-Climax document (Deliverable 5.54).

All five categories of comparison are fundamental in evaluating and quality-controlling the use of reanalyses. Each category benefits from the findings of the other categories. The breadth and depth of expertise and amount of effort required for each category is considerable. For these reasons, it is thus vitally important to co-ordinate the efforts of many individuals, and to use resources efficiently.

Internal metrics comparison (category 5) arguably underpins categories 2-4, so there are compelling reasons to place high priority on capacity-building measures to reduce the complexity levels for internal metrics comparison.

Within each category, the sections above consistently identified a need for co-ordination, not least to mobilize the participants, to collate/disseminate findings, and to promote the sharing/use of common software tools. Efficiency is also enhanced when reanalysis intercomparisons are preceded by standalone evaluation of individual reanalyses under the responsibility of each producer. It could be argued that a minimum level of self-evaluated quality should be reached and documented prior to undertaking extensive intercomparison - this

level could be characterized by achieving threshold levels (to be defined) in the Core-Climax System Maturity Matrix developed in another Workpackage. More generally, the timeline of the intercomparison workflow will need a set of triggers (decision points) in order to establish when the various activities can be usefully undertaken.

The notion that reanalysis characterization (evaluation/quality/uncertainty) can be condensed into a single, unique, comprehensive and independently verifiable measure remains elusive. Such a measure is often sought in the belief that it would allow an inexperienced user to pick a suitable dataset and use it with limited understanding of its characteristics. Experience suggests that reanalysis applications are too diverse to be treated in this way. Arguably, a better model would be to undertake capacity-building so that users make better-informed decisions - either through raising their own experience at conducting and interpreting the comparisons detailed above, and/or developing the Service infrastructure that enables them to collaborate with others who can pass on their greater experience.

We shall return to these issues in Deliverable 5.54.

References

BIPM, 2008: International vocabulary of metrology — Basic and general concepts and associated terms (VIM), BIPM, JCGM 200:2008 (http://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2008.pdf).

Simmons, A. J., Poli, P., Dee, D. P., Berrisford, P., Hersbach, H., Kobayashi, S. and Peubey, C. (2014), Estimating low-frequency variability and trends in atmospheric temperature using ERA-Interim. Q.J.R. Meteorol. Soc., 140: 329–353. doi: 10.1002/qj.2317