ERA report series

18 The observation feedback archive for the ICOADS and ISPD data sets

Hans Hersbach, Paul Poli and Dick Dee

Series: ERA Report Series

A full list of ECMWF Publications can be found on our web site under:

http://www.ecmwf.int/en/research/publications

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

Abstract

This document describes the import of the ISPDv2.2, ISPDv3.2.6, and the ICOADSv2.5 historical datasets from their original formats into the observation feedback archive (OFA) at the European Centre for Medium-Range Weather Forecasts (ECMWF) which includes information on their usage into the NOAA/CIRES 20th century (20CR) and ECMWF ERA-20C reanalyses. Both ISPD data sets contain observations for surface pressure and mean sea level pressure, world-wide, from 1768 to 2010. ICOADSv2.5, which ranges back from 1662, contains marine observations for a number of geophysical quantities. Although it is being extended in near real time, the OFA archive at ECMWF covers until December 2011.

The OFA is the unified data model at ECMWF for observations that have been passed through the data-assimilation system. Key element is the addition of feedback, embracing collocated model values and information on data usage. Three feedback archives are presented. The first OFA regards ISPDv2.2 with feedback from 20CR that had used this data set, and which is available in the data set itself. The second OFA regards ICOADSv2.5, which does not contain feedback, but at ECMWF model information from 20CR was added by an offline collocation with available model fields, instead. The third data set embraces feedback from the ERA-20C analysis for the data sets it was based on, i.e., the later version ISPDv3.2.6 and ICOADSv2.5. Although the ISPDv2.2, ICOADSv2.5 and ERA-20C OFA's contain many overlaps, their attached feedback differs.

The format of OFA is the simplified representation of the ECMWF observation data base (ODB), which was introduced at ECMWF recently to enable archiving into the ECMWF Meteorological Archival and Retrieval System (MARS). The datasets presented here are the first historical datasets that were ingested into this format and made publicly available on the ECMWF data server. A detailed description is provided on which elements in the native datasets are mapped where into ODB.

1 Introduction

The reconstruction of the past climate requires an adequate network of historical observations. During the 20th century the observing system has evolved dramatically. Although synoptic observations of surface wind, pressure and sea-surface temperature have been available for more than 150 years, their distribution has been sparse during most of this period. Upper-air data have a shorter history. The conventional observing system had a boost in the International Geographical Year (1957), when the observing system for the Southern Hemisphere was improved significantly. Substantial amounts of satellite data has been available since the 1980s.

Besides an incomplete and geographically highly uneven distribution, a large amount of early data has never been digitized, and as a result never been re-used. Worldwide, several data recovery initiatives have emerged to improve on this situation, often in coordination with the Atmospheric Circulation Reconstructions over the Earth (ACRE) project (http://www.met-acre.org).

One project that was involved in the recovery, digitization and usage of observational datasets suitable for global climate studies, with focus on the past 100 years is ERA-CLIM. This three-year initiative (2011-2013) was funded under the Seventh Framework Programme of the European Union, embraced eight partners and was coordinated by the European Centre for Medium-Range Weather Forecasts (ECMWF). A concise overview of the project is presented in Dee *et al.* (2012).

A specific aim of ERA-CLIM was to improve the quality and consistency of climate observations through global reanalyses. Following the foot steps of the NOAA-CIRES 20th Century Reanalysis Version II

(20CR) by Compo *et al.* (2011) a global reanalysis (ERA-20C) was integrated at ECMWF that is based on surface observations, only (Poli *et al.*, 2013). These observations originated from the International Surface Pressure Databank (ISPD) (Compo *et al.* (2010), Cram *et al.* (2015)) and the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Woodruff *et al.*, 2010).

Another deliverable within ERA-CLIM was the development of an observation feedback archive (OFA). This new facility allowed for the development of a permanent archive for observations as used in the past, present and future reanalyses on the ECMWF public data server (http://apps.ecmwf.int/datasets/). A key feature of the OFA is the inclusion of feedback information about data quality generated from the reanalyses. Such quantities allows users to gauge the observational information contents of reanalysis products, and to make meaningful uncertainty assessments for the observations.

The design and technical development of the OFA facility was modelled after the Meteorological Archival and Retrieval System (MARS), the main repository of meteorological data at ECMWF, and its recent ability to store observations from the ECMWF Observation Data Base (ODB) format. ODB is the data model that is used in the data assimilation part of the ECMWF Integrated Forecasting System (IFS). It includes quality control flags, data departures relative to both model first guess (i.e. prior to assimilation) and analysis (after assimilation), and bias estimates.

This document describes the first three data sets that were ingested into the observation feedback archive. These are the conversion of 1) the ISPDv2.2 and 2) ICOADSv2.5 data sets from their native format into ODB. Feedback information was imported from the 20CR reanalysis that had used the ISPD pressure observations. The third data set embodies the usage of the more recent ISPDv3.2.6 and ICOADSv2.5 into the ERA-20C reanalysis, and where the feedback information now reflects how these observations were used in ERA-20C.

An overview of the ODB format is provided in section 2, which includes some elements that were specifically introduced for historical data sets. Sections 3 and 4 describe the import of the ISPDv2.2 and ICOADSv2.5 datasets, respectively, while Section 5 presents the ERA-20C feedback archive. Concluding remarks are formulated in Section 6.

2 The ODB data model

2.1 Original representation

The observation database (ODB) was introduced at ECMWF about 10 years ago to improve the efficiency of observation processing inside the 4D-Var data assimilation system (Saarinen, 2004). A detailed manual can be found at the ECMWF ODB portal (https://software.ecmwf.int/wiki/display/ODB/).

ODB has elements of a hierarchical database and one of its strong points is the support of a SQL engine. Its data structure is represented by a number of linked tables. Each table consists of a number of columns (named `column@table`) that describe data characteristics (such as date, time, geo-location, observed value), and a number or rows that link to individual observations. Top of the hierarchy is formed by a set of tables that specify the assimilation window. An example is the `desc` table, with, e.g., columns for analysis date and time.

A distinction needs to be made between reports and observations. Meteorological reports often contain more than one observation. An example is a radiosonde ascent, which is seen as a single report containing observations of temperature, wind and humidity for a range of levels. One step downwards in the ODB

hierarchy are tables that identify such reports, and their columns provide information that is common to all observations within a report. An example is the `hdr` table with columns for latitude, longitude and observation time. The lowest level of tables provide information about the individual observations within reports. An example is the `body` table for observed value, observed height and geophysical observable. This latter quantity (represented by the `varno@body` column) is required to distinguish between, e.g., temperature and wind observations, since each row contains information on one specific observation of one specific quantity, only. The (evolving) list of varno's can be found at the ODB-Governance portal (http://data-portal.ecmwf.int/odbgov/Varno/).

This data model allows for an efficient and compact database. The `desc` table only contains one row, the number of `hdr` rows equals the number of reports and only the `body` table has one row per actual single observation. Special columns provide the link between tables, e.g., each row (report) in the `hdr` table has two columns that specify which rows in the `body` table belong to that report.

ODB is very flexible. The actual number of tables depends on the type of observations. For instance, satellite data have dedicated tables not relevant for conventional data and vice versa. Another flexible property is that the lowest hierarchy tables contain only one observation for one geophysical quantity per row. This allows for the addition of newly available geophysical quantities by simply providing a new definition for `varno@body`. This will not affect the data structure for any existing reports. This would be more problematic in a model where one row contains all information on a complete report, since new columns would have to be added, and this would affect all existing reports.

2.2 Simplified representation

One practical drawback of ODB was that it has a typical database structure containing a number of files in an encapsulated directory. This prohibited archiving of ODB into MARS, since that dictates single files. To accommodate for that, a new representation of ODB was designed. This format, called ODB-2, or also ODB-API, is a flat file rather than a directory. This enabled the archiving in MARS, which has recently become the standard manner in which observations are archived after they have been passed through the assimilation system. To reduce data volumes not all columns in the original ODB schema are selected for archiving. While the original representation insists on the completeness of the full schema, the new format is more flexible. In addition, rows, e.g., observations for which no observed value is available, can be filtered out, as well.

Like ASCII or BUFR (http://www.wmo.int/pages/prog/www/WMOCodes.html), concatenated ODB-2 files are also a valid ODB-2 file. Moreover, the SQL ability has been retained. The information content is identical to the original ODB structure, just the representation is different. One consequence of the flat file format is that all tables are expanded to the level of individual observations. Where previously the `desc` table contained only one line, it now equals the number of observations. Internal compression in ODB-2, though, prevents the potentially excessive increase in data volumes. The naming of ODB columns (`column@table`) has not changed. In principle the `@table` specification is redundant in a flat table structure, but it was remained since that allows for the mapping back to the original hierarchical data format. This is sometimes required, since internally the ECMWF variational assimilation system (4D-Var) remains to work with the original format.

It is this new representation, from now on simply denoted by ODB, in which the observation feedback archives as presented here were ingested, and on which the OFA is based.

Type	Missing value indicator	alias used in this document	SQL missing value
INTEGER	2147483647	IUNDEF	NULL
REAL	-2147483647.	RUNDEF	NULL
CHARACTER	”????????”	CUNDEF	NULL
BITFIELD	2147483647	IUNDEF	NULL

Table 1: Column types in ODB

2.3 Separation of reports

The original representation allows for a simple separation of reports. Each report has its own entry in the tables at report-level, such as `@hdr`, which contain pointers to the observation-level tables for their corresponding observations. In the simplified representation, this convenient link doesn't exist any more, and for this reason it is more difficult to know where one report ends and the next one starts. To facilitate this distinction, two book-keeping columns exist in ODB. The first one, `seqno@hdr` labels the report. So a next report starts where the value of `seqno@hdr` changes. Another quantity, `entryno@body` is a counter within a report. It usually starts at 1, but this is not guaranteed, since feedback archives may contain gaps due to the filtering of rows. Therefore, this latter column cannot be safely used for the distinction between reports.

2.4 Column types

Columns in ODB can either of type real, integer, 8-character string, or bitfield. When extracted (e.g. inside a Fortran or C program or dumped to ASCII) all columns are 8 bytes long; internally, the length varies subject to the acquired level of compression. Each type has its own missing number indicator. A bitfield is in principle an integer but its individual bits can be directly addresses in the SQL query language. An example is a column called `datum_status@body` which informs whether an observation was used or not. The bit `datum_status.active@body`, informs whether an observation was considered for assimilation, and the bit `datum_status.blacklisted@body` whether it was blacklisted from the outset.

2.5 ODB tools

At ECMWF several tools have been developed to work with ODB data. Central tool is **odb_sql** which allows any selection of columns (or formulae between them), subject to any conditions and ordering using an SQL-based language (select ... where ... order by ...). It works for both ODB representations. Output is by default ASCII, but it can be cast to a (new-style) ODB file as well. Similar functionality has been integrated into MARS, what is effectively used behind the screen on the public data server. Examples are available at the ECMWF ODB portal (https://software.ecmwf.int/wiki/display/ODB/Examples). Other tools (amongst others, but for new-style ODB only) are **odb_header**, printing out all column definitions, **odb_count** counting the number of rows, **odb_split** allowing to split ODB files into multiple ones according to a specified stratification and **odb_dump** which dumps the entire ODB file contents to the screen. At the reanalysis section a tool **odb_plot** was recently developed to create density plots between any two specified columns (or formulae of columns) subject to any condition for stratification.

Also, ODB libraries can be linked with Fortran and C programs that allow for a more custom-made manipulation. It is these libraries that were used for the dedicated programs that converted the ISPD and

Data set	expver (@desc)	class (@desc)	stream (@desc)	type (@desc)
ISPDv2.2	'1607'	e2 (22)	oper (1025)	ofb (263)
ICOADS2.5	'1608'	e2 (22)	oper (1025)	ofb (263)
ERA-20C	'0001'	e2 (22)	oper (1025)	ofb (263)

Table 2: Mars keys and ODB columns that identify a product

ICOADS data sets from their native format into ODB.

2.6 Organization in the MARS archive

The location of data in the MARS archive is determined by a set of keys. The values of these keys are also present in the archived data themselves. For GRIB fields, e.g., these are in the headers. For ODB files, they are represented by ODB columns, whose names and values do not always correspond to the MARS keys themselves. Also, for data that has passed the ECMWF assimilation system, the table name of these columns is usually omitted (i.e., for ERA-20C, but not for ISPDv2.2 and ICOADSv2.5). The MARS keys are expver, class, stream, type, date, time, obsgroup and reportye.

2.6.1 expver, class, stream, type

These keys, see Table 2, allow for the identification of a specific product. For the OFA's presented here, the class, stream, and type are all the same, indicating that they were produced within 20th century reanalysis class, and relate to ODB feedback, rather than BUFR or GRIB. Column `expver` is of type STRING, while the other three are of type INTEGER and contain the internally used numerical values for the corresponding mars keys (between parentheses).

2.6.2 date, time

These two keys specify an assimilation window, which groups together all observations that were considered during such a window. The corresponding ODB columns (both INTEGER, see Table 3) are `andate` and `antime` rather than date and time, since, confusingly, these latter indicate the actual timing of the observation. The ERA-20C reanalysis, (Poli *et al.*, 2013), uses an assimilation window of 24 hours. It is always labelled by `antime`=0 and contains data from 9 UTC (exclusive) the previous day until 9 UTC (inclusive). The ISPDv2.2 and ICOADSv2.5 are not the result of an in-house assimilation, but were acquired as analysis input for ERA-20C. For such data sets the 'analysis window' is by convention 6 hours long, indicated by 0, 6, 12, and 18, and containing data from 3 hours before (exclusive) and three hours after (inclusive) these time stamps. This also corresponds to the 6-hour assimilation windows as used in the 20CR reanalysis. The format of `antime` is HHMMSS, while it is HH, or HH:MM:SS for the mars key.

2.6.3 obsgroup, reportype

These two mars keys determine the type of report. Different obsgroups (`groupid` in ODB) have different specialized ODB tables (like conventional data versus satob satellite winds). All data considered here are conventional (obsgroup=conv, which relates to `groupid`=17). The reportype is a relatively

Data set	date (`andate@desc`)	time (`antime@desc`)	window
ISPDv2.2	17680101 - 20101231	0, 6(0000), 12(0000), 18(0000)	(-3h,+3h]
ICOADSv2.5	16621015 - 20111231	0, 6(0000), 12(0000), 18(0000)	(-3h,+3h]
ERA-20C	19000101 - 20101231	0	(-15h,+9h]

Table 3: Mars keys and ODB columns that identify the assimilation window

new attribute that was introduced to facilitate the archiving in MARS. Previously, a type of report was characterized by a number of quantities which either related to in-house classification (`obsgroup`, `codetype`, ..) or BUFR keys (`bufrtype, subtype`, ..). Each unique combination of such quantities has now been assigned its own reportype. Therefore, the values of a reportype dictates the value of all these other quantities, which are still being used inside the 4D-Var assimilation code.

For conventional data all reportypes are of the form 160XX. The reportypes as used in this report are summarized in Tables 8 and 16. The (evolving) full list of reportypes and associated columns is available at the ODB-Governance portal (http://data-portal.ecmwf.int/odbgov/FullReportType/).

2.7 Feedback information

ODB has dedicated columns that contain feedback information. This embraces collocated model information, applied bias corrections, assumed observation errors and flags that indicate how the data was used in the assimilation system. The subset of columns that apply to the OFA's presented in this document are listed in Table 4, where only for ERA-20C all columns are available. Bits in columns of type BITFIELD can be set independently. The resulting numerical value is given in binary representation, where the first bit listed is the least significant one (e.g. 2^0).

The most important feedback quantities are the departures (`fg_depar, an_depar`), the status flags (`report_status, datum_status`) and observation bias estimate (`biascorr`, whose unfortunate name choice could incorrectly suggest minus the bias estimate).

The OFA's for the ISPDv2.2 and ICOADSv2.5 were not processed within a 4D-Var assimilation, but were ingested directly from the original data format (HDF5 and IMMA, respectively). Therefore, these ODB's do not contain feedback information from an ECMWF (re)analysis. The ISPDv2.2 OFA contains feedback information from the 20CR reanalysis as available in the HDF5 format. The ICOADS data set does not contain any feedback information. For this OFA feedback was created at ECMWF by collocating observations with available 20CR model fields. This has been done for a number of geophysical quantities, as described below in Section 4.8. One should be aware, though, that this 'off-line' collocation is less accurate than the information obtained in an assimilation suite, and choices made for certain geophysical quantities are not ideal. Therefore, the *'offline'* determined feedback for ICOADS *should be treated with care* since it is not directly available from ICOADS nor the 20CR reanalysis, and reflects the choices made in the ECMWF collocation procedure between the two, instead. The ERA-20C OFA, embracing the usage of the ISPDv3.2.6 and ICOADSv2.5 does contain the feedback information as produced by this analysis.

Although the ISPDv2.2, ICOADSv2.5 and ERA-20C OFA's contain many overlaps, their attached feedback differs.

OFA	Column@table	Type	Description
1,2,3	report_status@hdr	BITFIELD	active, passive, rejected, blacklisted
1,3	report_event1@hdr	BITFIELD	no_data, all_rejected, bad_practice, rdb_rejected redundant, stalt_missing, qc_failed, overcast_ir
1,2,3	timeseries_index@hdr	INTEGER	introduced for historical data, see Section 2.8.2
2,3	lsm@modsurf	REAL	model land-sea mask (fraction)
1,2,3	orography@modsurf	REAL	model orography (m)
2,3	seaice@modsurf	REAL	model sea-ice fraction
2,3	t2m@modsurf	REAL	model first-guess 2-metre temperature (K)
2,3	windspeed10m@modsurf	REAL	model first-guess 10-metre wind speed ($m\,s^{-1}$)
2,3	u10m@modsurf	REAL	model first-guess 10-metre zonal wind ($m\,s^{-1}$)
2,3	v10m@modsurf	REAL	model first-guess 10-metre meridional wind ($m\,s^{-1}$)
1,2,3	datum_status@body	BITFIELD	active, passive, rejected, blacklisted
1,3	datum_event1@body	BITFIELD	see IFS documentation
1,2,3	fg_depar@body	REAL	bias-corrected observation minus model first guess
1,2,3	an_depar@body	REAL	bias-corrected observation minus model analysis
3	biascorr_fg@body	REAL	first-guess bias estimate, *not* bias correction
1,3	biascorr@body	REAL	analysis bias estimate, *not* bias correction
3	varbc_ix@body	REAL	connects observations within one bias group
1,2,3	bias_volatility@body	REAL	introduced for historical data, see Section 2.8.3
3	an_sens_obs@body	REAL	measure of influence in the assimilation
3	qc_pge@body	REAL	see IFS documentation
3	hires@update_2	REAL	see IFS documentation
1,3	obs_error@errstat	REAL	assumed observation error
3	final_obs_error@errstat	REAL	modified observation error
3	fg_error@errstat	REAL	model first-guess error at obs location
3	eda_spread@errstat	REAL	model first-guess ensemble spread at obs location

Table 4: ODB columns related to feedback information available for 1) ISPDv2.2, 2) ICOADSv2.5 and 3) ERA-20C. For BITFIELD the description lists the bit flags as addressable in SQL, e.g., report_status.active. The IFS documentation is available from www.ecmwf.int.

Column name	Type	Description
source@hdr	STRING	specifies the historical data set as a whole
collection_identifier@conv	INTEGER	specifies a particular collection, source, deck in a data set
unique_identifier@conv	INTEGER	specifies a particular report
timeseries_index@conv	INTEGER	effort to connect reports between different sensing times
bias_volatility@body	REAL	break-point indicator estimated from feedback information

Table 5: ODB columns created for historical data sets

2.8 New ODB columns for historical data sets

2.8.1 Traceability

A desirable property for a historical data set is traceability. For this reason three new ODB columns were introduced, that uniquely specify data up to the report level. At the top level, a column `source@hdr` describes a data set as a whole. This string ('ISPDv2.2', 'ISPD3.26' and 'ICOADS25' for the data sets described here) allows for the separation of sets in case they are mixed together (like in ERA-20C). Typically, each data set consists of a number of collections, sources or also called decks (ICOADS), each representing a subset delivered from one particular source (such as "KNMI stations 1911-2006" within ISPD). Since data within each subset has been typically collected in a similar fashion, it is very useful to be able to separate them at high level. For this the column `collection_identifier@conv` was introduced (The `conv` table is an example of a report-level table that specifically applies to conventional data, but not to other data, such as from satellites). Finally, it is desirable to be able to pin-point individual reports. When, e.g., a problem is encountered (e.g., alleged digitization error) this highly facilitates the tracing back to the actual data sheet from which that report was obtained. All three data sets provide such unique labels, which have been incorporated in the new ODB column `unique_identifier@conv`.

Note that different data sets use different labelling for collections and unique identifiers. Therefore, only the joining together with the source provides a 'universally' unique description of those quantities.

2.8.2 Linking reports

To facilitate the usage of a bias-correction scheme when presented to a data assimilation suite, like ERA-20C, two more columns have been introduced. The contents of these columns *does not* directly originate from the original data sets, but has been determined in a subjective way at ECMWF, based on information that is available in these sets. They can be regarded as some form of feedback.

One column, called `timeseries_index@conv`, aims to connect reports between different dates and times. Where unique station names are known, such labelling is trivial. However, for historical data such info is not always available. For instance, a very popular station identifier for early maritime data (the Call Sign) is SHIP, which actually embraces a large number of different ships. In order to accommodate for this, a simplistic scheme was designed that tries to separate physically different stations/ships. Each such subset was given its own `timeseries_index`.

This algorithm works as follows. In ODB the information on station identifier or Call Sign is placed into the 8-character column `statid@hdr`. First the entire archive is divided into subsets of reports with the same `statid` and reportype. The reason for the stratification according to reportype is to avoid the grouping together of e.g., ships with the names of cities with stations in those cities. Different collections

are grouped together, since overlap between them does occasionally occur.

For each such subset it is investigated into how many subgroups the individual reports could be divided, each, hopefully characterizing a separate physical platform. The algorithm scans the reports in chronological order, and starts with one subgroup containing the earliest report, only. For each subsequent report it is determined for which subgroup the distance which its latest containing report is minimal and for which subgroup the average speed would be minimal (according to the great circle distance; land/sea masks are not taken into consideration). In case the minimal distance is below 200 km the report is assigned to that subgroup, regardless of the required speed to arrive there. Reason for this tolerance is the fact that the position of early maritime data can be available up to 2 degrees, only. This avoids an artificial splitting of subgroups due to the limited level of accuracy. If the minimal distance does exceed 200 km it is tested whether the minimal required average speed is below $50\,\mathrm{m\,s^{-1}}$. If that is the case, the report is assigned to the corresponding subgroup. If this condition is not satisfied a new subgroup is created.

After this analysis the subgroups are reordered for each reportype independently, according to the volume of reports they contain and are assigned a value of `timeseries_index@hdr` with 1 for the most populated, 2 for the second-most populated, etc. Therefore, the `timeseries_index@hdr` should be combined with `reportype@hdr` to obtain candidates for physically different platforms.

The algorithm employed here is rather simplistic and ad-hoc, and will certainly contain many misconnections. A more thorough grouping, however, would require an in depth analysis, which, given the amount of different statid's (over one million for maritime data) was found infeasible. Also, the `timeseries_index` is just a label to connect reports. It was, e.g., determined independently for the ISPDv2.2 and ISPDv3.2.6 data sets, and, therefore the actual values of this quantity will not coincide. The ordering was applied to facilitate the identification of the most populated stations.

2.8.3 Bias volatility for pressure, wind and temperature observations

The second column was introduced to provide proxy information on the likelihood for a break, e.g. due to developing instrument problems or a change in station location. Such metadata is not always available, and it is desirable to develop a method that automatically estimates the likelihood of a break. A quantity called `bias_volatility@body` was determined for this purpose, which is determined per observation. and is based on feedback departure statistics from the 20CR reanalysis.

For a given observation measuring a certain geophysical quantity, let a set for a number of preceding observations for given `timeseries_index` and reportype be labelled by 1, a set after this observation by 2, and their union by 0. If one defines the number, mean and standard deviation of 20CR first-guess departures (excluding the applied observation bias correction) for set $i = 0,1,2$ by n_i, μ_i and σ_i, respectively, then the bias volatility is defined as:

$$b = S/b_n, \tag{1}$$

where

$$S = \frac{(\mu_1 - \mu_0)^2 + (\mu_2 - \mu_0)^2}{\sigma_0^2} \tag{2}$$

and

$$b_n = \frac{n_2}{n_1} + \frac{n_1}{n_2} \tag{3}$$

is a normalization factor. It is related to the standard normal homogeneity test (SNHT, Haimberger (2007), Alexandersson (1986)). It is important that the observation bias corrections, as applied in 20CR, are removed from the departures, since otherwise many breaks would be potentially masked.

In the appendix it is shown that b can also be written as:

$$b = \frac{(\mu_2 - \mu_1)^2}{(\mu_2 - \mu_1)^2 + \left(\frac{n_0^2}{n_1 n_2}\right)\sigma_m^2}, \tag{4}$$

where

$$\sigma_m^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} \tag{5}$$

is the weighted average of the standard deviations of the two individual subsets. From this it directly follows that:

$$0 \leq b \leq 1.$$

The bias volatility is equal to zero when sets 1 and 2 have the same mean departure (regardless their individual variance), and it is only equal to 1 in the case of two constant time series, but with different mean. This latter is the clearest break imaginable, and, indeed, is the situation where the bias volatility has its largest value.

The value of the bias volatility obviously depends on the length of the sets 1 and 2. For each observation, these were chosen as the latest and next 365 days with data, respectively. The aim is to filter out seasonal variations, which could lead to spurious break signals (see e.g., Haimberger (2007)). Only when data is available every day, this relates to a one-year period. If for one day there would be no data, the period is 366 days, etc. In the case of a data void period, the bias volatility, therefore, ignores the gap and just gathers the number of data-days up to 365. This simple approach may not be optimal for some stations with a (irregular) long periods of data gaps, such as in polar regions. In such cases the strategy by Haimberger (2007) to delete data from a respective month in an interval when it was missing in other intervals, may be preferred, though was found more difficult to implement for the full archive.

For (both versions of) ISPD bias volatility was determined for each observation of mean sea level pressure and surface pressure and where 20CR feedback was available. For ICOADS it was in addition determined for wind speed, wind direction, sea-surface temperature, 2-metre temperature and pressure tendency. For all other geophysical quantities it was set to missing.

Technically this was achieved in the following way. First, the data was read in chronological order, where for each timeseries index, reportype and geophysical quantity considered, the number, mean and standard deviation for each of the latest 365 data days was kept in memory (by flushing the earliest day whenever a new day was encountered). This enabled the determination of n_1, μ_1, σ_1 for each observation, which was stored in auxiliary ODB files. Next the archive was scanned in reverse order to determine n_2, μ_2, σ_2. Finally the information in the forward and backward auxiliary ODB files were combined to determine the bias volatility from Eq. 4, and was inserted back into the original ODB files.

In ERA-20C the bias volatility determined from 20CR feedback served as input into the assimilation system. It allowed to vary the confidence in the first-guess bias estimate. The larger the bias volatility (from 20CR), the quicker the observation bias estimate (in ERA-20C) was allowed to change. An example for a land station in ISPDv3.2.6 (assigned a timeseries index of 9) is displayed in the middle panel of Fig. 1.

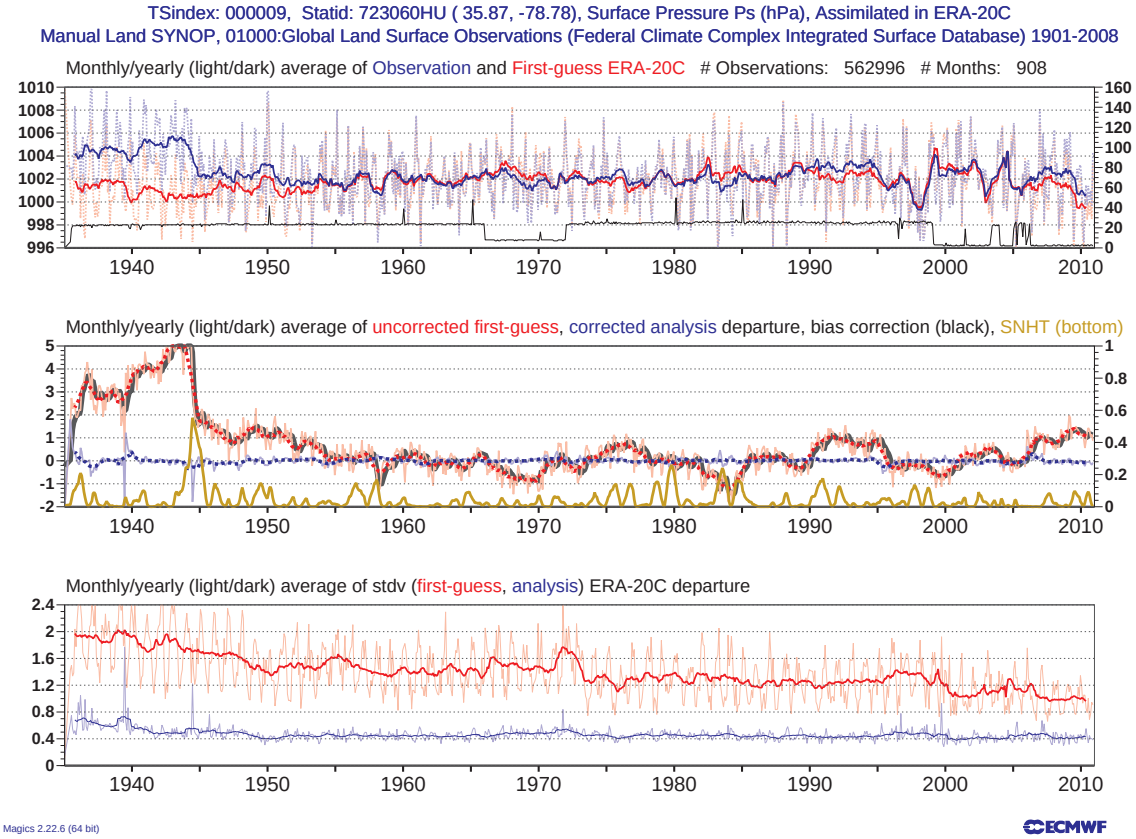

Figure 1: Assimilation of surface pressure in ERA-20C for data at station 723060HU as provided by ISPDv3.2.6. Top panel shows the monthly/yearly mean (light/dark colours) for the observed (blue) and ERA-20C model (red) pressure (hPa). The black curve indicates the number of observations assimilated per day. The middle panel displays the monthly/yearly mean (light/dark) departures for the uncorrected ERA-20C first guess (red) and bias corrected analysis (blue), based on the dynamically evolved bias estimate (black curve; not time averaged). Its adaptivity depends on the value of the bias volatility (gold) as determined before hand from 20CR departures. The lower panel shows monthly/yearly means (light/red) for the stdv of the bias corrected first-guess (red) and analysis (blue) departures.

3 The International Surface Pressure Databank version 2.2 with feedback from 20CR

The International Surface Pressure Databank v2.2 (Compo *et al.* (2010), Cram *et al.* (2015)) is the world's largest collection of pressure observations, as prepared in 2009. It has been gathered through international cooperation with data recovery facilitated by the ACRE Initiative and the other contributing organizations and assembled under the auspices of the GCOS Working Group on Surface Pressure and the WCRP/GCOS Working Group on Observational Data Sets for Reanalysis by NOAA Earth System Research Laboratory (ESRL), NOAA's National Climatic Data Center (NCDC), and the Climate Diagnostics Center (CDC) of the University of Colorado's Cooperative Institute for Research in Environmental Sciences (CIRES). The ISPDv2 consists of three components: station, marine, and tropical cyclone best track pressure observations. The station component is a blend of many national and international collections. Prior to ERA-20C, 20CR was the first reanalysis that had used this data set.

Table name	Used attributes
Common	
MetaData	unoc, id, year, month, day, hour, minute, second, lat, lon, elv
Observations	slp, sfp
ObsType	ispdbcid
Tracking	sname
Marine	
SrcMarine	pt
20CR	
NewFeedBack	epvims, bias, sfsfp, ai, uc, bcf, qc, mpmemfg, mpmema, melv

Table 6: Part of the HDF5 data structure for ISPD that was used in this document. No information from dedicated Land and Hurricane tables were used.

3.1 Data Format and volume

The data is available in HDF5 format. For ISPDv2.2, yearly tar files, containing hourly files have been obtained from NCAR (http://rda.ucar.edu/datasets/ds132.0/). A manual describing the data contents can be found at this location as well. A backup of all files was placed in the ECMWF File Storage system (ECFS) at location ec:/ERAS/observations//ispd/ispdv2.2.tar/yyyy.tar. Data is organized in dedicated tables into three main components (Land, Maritime and Hurricanes). Feedback from 20CR is available in the NewFeedBack table. For convenience, the table structure and the names for those attributes that were used in the ODB conversion is summarized in Table 6. Naming follows that of the C++ reader that is provided at the NCAR site as well (in h5f_reader.tar under the Software tab). Details on full table names, attributes and their contents can be found in the online HDF5 documentation.

The total ISPDv2.2 database embraces 1,396,997,564 reports, distributed over 57 collections from 1768 to 2010. No data is present for the years 1794, 1795, 1797, 1798, 1799, 1803, 1806, 1808 and 1814.

3.2 Import into ODB

A C program using the available C++ readers was written that translates the hourly HDF5 data files into 6-hourly ODB files which were archived in MARS (expver=1607). It is not always clear what type of information from ISPD should belong where in ODB. For quantities like geo-location, date and time this is not an issue, but for flags regarding data usage, and keywords that classify reports, e.g., it is more difficult to map these onto the ECMWF-based design. For this reason the choices made may not always be optimal. Also, not all available meta data could be translated. However, the usage of the unique identifier should facilitate the access to such quantities, if necessary. An overview of the ODB columns that were created for ISPD is given in Table 7.

3.3 Source, collection, unique identifier and observation time

The source of this data set was chosen to be 'ISPDv2.2'. Data has been gathered from 57 collections, labelled by the attribute ObsType.ispdbcid (International Surface Pressure Data Bank Collection ID). The full list and specification can be found in the HDF5 documentation and in Cram *et al.* (2015). Examples are pressure observations from 'ICOADS release 2.5' (0105) and 'International Best Track Archive for Climate Stewardship IBTrACS' (8001). The ODB column `collection_identifier` adopts this

column@table	Type	Relation to the ISPD HDF5 Table
expver@desc	STRING	'1607'
type@desc	INTEGER	263
class@desc	INTEGER	22
stream@desc	INTEGER	1025
andate@desc	INTEGER	17680101 - 20101231
antime@desc	INTEGER	0, 60000, 120000, 180000
seqno@hdr	INTEGER	1 to number of OrigMetaData in 6-hour window
reportype@hdr	INTEGER	see Table 8
bufrtype@hdr	INTEGER	see Table 8
subtype@hdr	INTEGER	see Table 8
groupid@hdr	INTEGER	17, i.e. Conventional data
obstype@hdr	INTEGER	see Table 8
codetype@hdr	INTEGER	see Table 8
stalt@hdr	REAL	MetaData.elv, RUNDEF if 9999
statid@hdr	STRING	see Section 3.6
date@hdr	INTEGER	100*(100*MetaData.year+MetaData.month)+MetaData.day
time@hdr	HHMMSS	100*(100*MetaData.hour+MetaData.minute); see Section 3.3
report_status@hdr	BITFIELD	see Table 10
report_event1@hdr	BITFIELD	see Table 11
lat@hdr	REAL	MetaData.lat
lon@hdr	REAL	MetaData.lon, cast into (-180,180] interval
source@hdr	STRING	'ISPDv2.2'
orography@modsurf	REAL	NewFeedBack.melv, RUNDEF if 9999
timeseries_index@conv	INTEGER	see Section 2.8.1
collection_identifier@conv	INTEGER	ObsType.ispdbcid, IUNDEF if -99999
unique_identifier@conv	INTEGER	MetaData.unoc; see Section 3.3
entryno@body	INTEGER	1 or 2
obsvalue@body	REAL	100*(Observations.slp, Observations.sfp)
varno@body	INTEGER	110 (ODB convention for pressure observation)
vertco_type@body	INTEGER	2
vertco_reference_1@body	INTEGER	0: MSLP, stalt@hdr: Surf Press.
datum_status@body	BITFIELD	see Table 10
datum_event1@body	BITFIELD	see Table 12
biascorr@body	REAL	100*NewFeedBack.bias, RUNDEF if >1.E+20
bias_volatility@body	REAL	see Section 2.8.3
an_depar@body	REAL	100*NewFeedBack.mpmema, RUNDEF if 9.9E+31 or 9.95E+15
fg_depar@body	REAL	100*NewFeedBack.mpmemfg, RUNDEF if 9.9E+31 or 9.95E+15
obs_error@errstat	REAL	100*NewFeedBack.epvims, RUNDEF if -9.99
ppcode@conv_body	INTEGER	0: MSLP, 1: Surf Press.

Table 7: ODB layout for ISPDv2.2 data. `vertco_reference_1` *was encoded as height (m), rather than geopotential height (*m^2s^{-2}*), which latter would have been the correct ODB unit.*

HDF5 table	Reportype	Description	B	S	O	C
Land	16002	Manual Land SYNOP	0	1	1	11
Maritime	from SrcMarine.pt as in Table 16					
Hurricane	16060	Tropical cyclone bogus	1	19	1	23

Table 8: Reportypes assigned in ISPD and derived bufrtype (B), subtype (S), obstype (O) and codetype (C).

numbering.

The ISPD data set contains a 7-digit unique identifier, represented by the attribute MetaData.unoc. However, it is reset for each minute of data. Therefore, the time and date of a report is also required to uniquely label it. For Land, Maritime and Hurricane reports, the first digit is always 0, 1 and 8, respectively.

Date and time information is provided in the MetaData table. For maritime data, however, the minute information is often missing (MetaData.minute=99). For those cases the sensing time was assumed to be on the rounded hour (i.e., minute=0).

3.4 Observables and reports

The HDF5 format embraces one report per row, containing up to two observations:

- mean sea level pressure MSLP (Observations.slp)
- surface pressure SP (Observations.sfp)

In ODB each row contains one observable. Therefore, for those reports where both quantities are available, the report extends over two rows, otherwise one. Reports can be separated by the auxiliary column `seqno@hdr`. Alternatively, `unique_identifier` can be used as well. The auxiliary column `entryno@body` also runs from 1 to up to 2. Its value doesn't ambiguously link with MSLP or SP, though. This is usually taken care of by `varno@body`. In this case, however, both MSLP and SP regard a pressure observation, and, therefore, share the same value `varno=110`. To accommodate for this, an additional ODB column, called `ppcode@conv_body` exists, which is 0 for MSLP, and 1 for SP. Pressure is converted from hPa (ISPD convention) to Pa (ODB convention) and placed in the column `obsvalue@body`.

The type of report (`reportype`) depends on the component (see Table 8). For land data it is always assumed to be a 'Manual Land SYNOP', for Hurricane the new reportype 'Tropical cyclone bogus' was created, while for Maritime data (pressure information originating from ICOADS) the reportype varies according to the platform type as specified by the SrcMarine.pt attribute (see Table 16). Unfortunately, reportype does not distinguish between drifting and moored buoy (both using 16005). The columns `bufrtype`, `subtype`, `obstype`, and `codetype` are dictated by the reportype. Their actual values are not related to BUFR data at all (i.e., data format is HDF5, not BUFR), and the only reason why these have been added to the archive is that, internally, the ECMWF assimilation system still uses them for stratification.

3.5 Vertical coordinate

In ODB the vertical coordinate is contained in the `vertco_reference_1@body` column. For ISPD, this quantity is provided in meters (for SP, 0 for MLSP), which means that the column that indicates

the type of the coordinate, `vertco_type@body`=2 (for height; 1 would indicate pressure coordinate). Actually, the value of 2 indicates geopotential height (in m^2s^2), i.e., not height (m). So unfortunately this has been *encoded incorrectly in the ISPD ODB*. On report level, station height is also stored in the `stalt@hdr` column.

3.6 Station identifier

ISPD uses a 13-character string (Metadata.id, first character always blank) to specify a fixed station for land data. In addition, a 31-character string (Tracking.sname) is available. Although each of them is available for each report, one of them often contains missing values.

For ODB, this information had to be condensed into a single 8-character string `statid@hdr`, since that is the only column dedicated to station names. In total 1,166,843 different station names are identified.

The following rules were applied:

- the choice which attribute is used depends on the collection.
  - Metadata.id for collections 1000, 1002, 1003, 2000, 2001, 3002, 3004, 3005, 3006, 3008, 3009, 3013, 3014, 3015, 4004, 10000 to 19999,
  - Tracking.sname for collections 104, 105, 300, 400, 1005, 1006, 1007, 1011, 1012, 3007, 3010, 3011, 3012, 3016, 4000, 4001, 4002, 5002, 5006, 8000, 8001,
  - both for collection 4003 (see below).
- Only letters a-z, A-Z and numbers 0-9 are accepted; all other characters are stripped, starting from the left: This means that all leading blanks are removed and any strange characters (like $\backslash, \#, @, !$), which likely indicate incorrect encoding, are removed.
- All letters are converted to upper case. The inconsistent usage of upper/lower cases for a number of stations is now automatically addressed.
- Any remaining characters beyond the first 8 permissible characters cannot be incorporated.
- Any resulting blank string is converted to ODB missing value: '????????'.

Special rules are applied for the following collections:
ICOADS data (collections 104, 105, 300, 400)
Attribute Tracking.sname contains a combination of a 1-2 digit ID indicator, followed by the Identification/Call Sign. The ID indicator is stripped, to only retain the Identification/Call Sign.

Federal Climate Complex Integrated Surface Database (collection 1000)
Attribute Metadata.id is composed as: XXXXXX-YYYYY, where XXXXXX=USAF, YYYYY=WBAN, all numbers. Usually one is missing, occasionally both are present. When USAF is missing the 5 digits of WBAN are used, else, the 6 digits of USAF. If in addition WBAN is non-missing, the remainder of WBAN is transformed with respect to base-26 into a 2-character string AA to ZZ, and appended to the 6 USAF digits. This last rule is necessary, since a few USAF id's occur several times. The most extreme example is USAF=949999, which has USAF's from 00001 to 00370; each belonging to a different location.

Metadata.id	Tracking.sname	statid@hdr
03895	JERSEY AIRPORT	03895
999999999999	JERSEY CHANNEL ISLAND G AND A	GANDA
999999999999	JERSEY CHANNEL ISLAND LANGLOIS	LANGLOIS
999999999999	JERSEY CHANNEL ISLAND ST AUBIN	STAUBIN
999999999999	JERSEY CHANNEL ISLAND ST LOUIS	STLOUIS
999999999999	JERSEY MAISON ST LOUIS	MALOUIS

Table 9: Statid for Jersey Island

NewFeedback	Meaning	Add to flag	Description
ai=1	The observation was assimilated	1 (set bit 0)	Active
qc=1	The observation was rejected	4 (set bit 2)	Rejected

Table 10: Contents of the `report_status@hdr` *and* `datum_status@body` *columns*

Russian Land Surface Observations (collection 1003)
Format of Metadata.id is: XXX000YYYYY. The middle 000 is removed, so statid=XXXYYYY.

Jersey Island (collection 4003)
This involves 6 stations, which are encoded according to Table 9.

3.7 20CR feedback Information

Feedback information from the 20CR reanalysis is available in the HDF5 NewFeedBack table. First-guess, analysis departures, bias estimates, and observation errors are found from the mpmemfg, mpmema, bias and epvims attributes, respectively. All units are converted from hPa to Pa. These quantities all relate to the ensemble mean of 20CR, i.e., *not* to individual members.

Each Numerical Weather Prediction centre (NWP) uses its own convention for flagging data. This complicates the translation of the 20CR reanalysis flags into the ECMWF-based ODB columns. An attempt has been made for the ODB columns `report_status`, `datum_status`, `report_event` and `datum_event1`(see Tables 10, 11 and 12). Status flags indicate how an observation (or entire report) was used in the assimilation system, while event flags provide information on why status flags were set.

The ISPDv2.2 HDF5 product contains information on the 20CR ensemble spread at observation location, which, in hindsight could have been included in the `eda_spread` column (see Tables 4). Unfortunately this quantity was not included in the current ODB conversion (i.e., `expver='1607'`).

NewFeedback	Meaning	report_event1	Description
uc=1	usable	0	
uc=0	not usable	1024 (set bit 9)	Redundant Report

Table 11: Contents of the `report_event1@hdr` *column*

NewFeedback	Meaning	Add to datum_event1	Description
uc=1	usuable	0	
bcf=1	Obs outside the ens spread + obs error	1024 (set bit 9)	Too big fg depar
uc=0	Observation not usable	8192 (set bit 12)	Redundant datum

Table 12: Contents of the `datum_event1@body` *column*

4 The International Comprehensive Ocean-Atmosphere Data Set release 2.5 with collocated feedback from 20CR

The International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Woodruff *et al.*, 2010) is a very mature collection of marine surface observations, containing data as far back as the 17th century. There have been several major releases since the first one in 1985, as well as near-real time (monthly) updates. At the time of writing, the latest official release was version 2.5. However, access was also provided to a version 2.5.1 (up to 2007) whose data contents is identical to version 2.5, but where each report was now assigned a unique identifier.

4.1 Data Format

ICOADS data is encoded in the IMMA format, which is ASCII based, containing one report per row. Each row is composed of a number of 'columns', each embracing a prescribed number of ASCII characters, which contain information on geo-location, observation date/time, a range of geophysical observables, and meta data. Missing values are indicated by blanks. To facilitate the decoding process, each 'column' is labelled by a meaningful key word, such as YR ('column' 1, 4 characters) for year, SLP ('column' 25, 5 characters) for sea level pressure, HOA ('column' 193, 3 characters) for height of anemometer, etc. The full list can be found at the online IMMA Documentation, page 24-30, which is available from http://icoads.noaa.gov/.

IMMA data is presented in compressed (.Z) monthly files, which are grouped into yearly or multi-yearly tar files. Both versions 2.5. and 2.5.1 were obtained from NCAR (http://rda.ucar.edu/datasets/ds540.0/). At ECMWF, backups were stored at ec:/ERAS/observations/icoads/.

4.2 Import into ODB

A dedicated Fortran90 program was developed, based on a Fortran-77 reader, rdimma0.f, as available from http://icoads.noaa.gov/software/rdimma0. It converts the monthly IMMA files into 6-hourly ODB files, where reports (one per IMMA row) are expanded over multiple ODB rows (one per observable). The program also performs a collocation with 20CR model fields, as obtained from NCAR (http://rda.ucar.edu/datasets/ds131.1/, Compo *et al.* (2009); details are given in Section 4.8. The resulting ODB feedback archive (expver=1608 in MARS) is based on IMMA_R2.5.1 where available (1662 to 2007), and augmented by IMMA_R2.5 from 2008 to 2011.

It is not always clear what type of information from ICOADS should belong where in ODB. Difficulties encountered relate to units for some geophysical quantities, and the translation of the available 'enhanced trimming flags' into ECMWF-based conventions. For this reason the choices made may not always be optimal. Also, not all available meta data could be incorporated. However, the usage of the unique identifier should facilitate the access to such quantities, if necessary. An overview of the ODB columns

Column@table	Type	Relation to ICOADS IMMA key words
expver@desc	STRING	'1608'
type@desc	INTEGER	263 (ofb)
class@desc	INTEGER	22 (e2)
stream@desc	INTEGER	1025 (da, a.k.a oper)
andate@desc	INTEGER	YR, MO, DY
antime@desc	INTEGER	HR
seqno@hdr	INTEGER	entry in monthly IMMA file
reportype@hdr	INTEGER	see table 16
bufrtype@hdr	INTEGER	see table 16
subtype@hdr	INTEGER	see table 16
groupid@hdr	INTEGER	17, i.e. Conventional data
obstype@hdr	INTEGER	see table 16
codetype@hdr	INTEGER	see table 16
stalt@hdr	REAL	HOP (height of visual observation platform)
statid@hdr	STRING	ID (Identification/Call Sign)
date@hdr	INTEGER	YR, MO, DY
time@hdr	INTEGER	HR
report_status@hdr	BITFIELD	enhanced trimming/NCDC flags
lat@hdr	REAL	LAT
lon@hdr	REAL	LON
source@hdr	STRING	'ICOADS25'
lsm@modsurf	REAL	field LAND_sfc, from 20CR
seaice@modsurf	REAL	field ICEC_sfc, from 20CR
orography@modsurf	REAL	field HGT_sfc, from 20CR
t2m@modsurf	REAL	field TMP_2m, from 20CR
windspeed10m@modsurf	REAL	field UGRD_10m, VGRD_10m, from 20CR
u10m@modsurf	REAL	field UGRD_10m, from 20CR
v10m@modsurf	REAL	field VGRD_10m, from 20CR
timeseries_index@conv	INTEGER	see Section 2.8.1
collection_identifier@conv	INTEGER	DCK (Deck)
unique_identifier@conv	INTEGER	see Section 4.3
anemoht@conv	REAL	HOA (height of anemometer)
baroht@conv	REAL	HOB (height of barometer)
station_type@conv	INTEGER	PT (Platform Type)
entryno@body	INTEGER	1 to number of observables in report
obsvalue@body	REAL	see Table 14
varno@body	INTEGER	see Table 14
vertco_type@body	INTEGER	2
vertco_reference_1@body	REAL	HOP, HOA, HOB, HOT, DOS
datum_status@body	BITFIELD	enhanced trimming/NCDC flags
bias_volatility@body	REAL	see Section 2.8.3
an_depar@body	REAL	from 20CR model fields
fg_depar@body	REAL	from 20CR model fields
ppcode@conv_body	INTEGER	0

Table 13: ODB columns for ICOADS2.5. The description of the IMMA keywords can be found in the online IMMA documentation (from http://icoads.noaa.gov/). The 20CR forecast fields are specified in Table 19.

that were created for ICOADS is given in Table 13.

4.3 Source, collection, unique identifier and observation time

The source of this data set was chosen as 'ICOADS25'. ICOADS consists of 134 different collections, called decks. These 3-digit labels run from 110 to 999, and were assigned to the `collection_identifier` column. A list and description of all decks can be found in the online IMMA documentation.

IMMA_R2.5.1 contains a unique identifier, which is a 6-character alpha-numeric string encoded as a base-36 number, starting from 0 for the first report (1662-10-15, 12 UTC) up to 294,720,587 (2007-12-31, 21 UTC). In ODB this number was used for `unique_identifier`. From 2008 onwards, only IMMA_R2.5 is available. As an alternative `unique_identifier` was assigned the reports entry number in the monthly file. Therefore, for this period the value of the month is required as well to pin-point a specific report.

In IMMA time information is condensed into one quantity (HR), digitized up to 0.01 hour. For ODB it was converted to the closest minute, i.e., `date=HHMM00`.

4.4 Observables and reports

ICOADS contains a wide range of geophysical observables. In total 28 were imported in ODB, each identifiable by their own value of `varno`. The full list is given in Table 14, which states on what IMMA key it was based, and what unit conversion was applied. For some quantities this involved the conversion from WMO tables. In such cases the centre values of specified intervals were chosen. Details are provided in Table 15. For wind, U and V components were added as well, and calculated from speed (W) and direction (D, meteorological convention) where available.

For a given report (i.e., one IMMA row) only those observables were placed in ODB where an observed value was available. Nevertheless, one report can extend over a number of rows and its length varies from case to case. Reports can be separated by the auxiliary column `seqno@hdr` or also the `unique_identifier`. The auxiliary column `entryno@body` runs from 1 up to the number of available observables, which varies. Its value does not ambiguously link to geophysical quantities.

The reportype is based on the ICOADS Platform Type (PT). The link is provided in Table 16. Unfortunately, drifting and moored buoys share the same reportype (16005). The value of PT has been stored in the column `station_type`, which therefore allows for a more direct link. The columns `bufrtype`, `subtype`, `obstype`, and `codetype` are dictated by the reportype. Their actual values are not related to BUFR data at all (i.e., data format is IMMA, not BUFR), and the only reason why these have been added to the archive is that, internally, the ECMWF assimilation system still uses them for stratification.

4.5 Vertical coordinate

In ODB the vertical coordinate is contained in the `vertco_reference_1` column. For ICOADS this quantity is always provided in meters (specified according to Table 14) which means that the column that indicates the type of the coordinate, `vertco_type@body`=2 (for height; 1 would indicate pressure coordinate). Actually, the value of 2 indicates geopotential height (in m^2s^2), i.e., not height (m). So this has been *encoded incorrectly in the ICOADS ODB*. Fortunately, it didn't affect the assimilation in ERA-20C. All data relate to observations around sea level, which means that `pp_code=0`. On report

varno	ICOADS Description	ICOADS unit	ODB unit	IMMA key	vertco	QC flag
83	Ship course	WMO 0700	degree	DS	HOA	2
82	Ship speed	WMO 4451	m/s	VS	HOA	2
110	Pressure	hPa	Pa	SLP	HOB	6
30	Pressure tendency	hPa/3h (>0)	$\pm$ Pa/3h	PPP	HOB	14
39	2m air temperature	Celsius	Kelvin	AT	HOT	7
40	Dew point temperature	Celsius	Kelvin	DPT	HOT	9
111	Wind direction	degree	degree	D	HOA	2
112	Wind speed	m/s	$\mathrm{m\,s^{-1}}$	W	HOA	2
41	U Wind component	m/s	$\mathrm{m\,s^{-1}}$	W, D	HOA	2
42	V Wind component	m/s	$\mathrm{m\,s^{-1}}$	W, D	HOA	2
130	Char. of Press. tendency	WMO 0200	WMO 0200	A	HOB	14
62	Visibility	WMO 4377	WMO 4377	VV	HOP	3
61	Present weather	WMO 4677	WMO 4677	W1	HOP	4
60	Past weather	WMO 4561	WMO 4561	WW	HOP	5
160	Past weather 2	WMO 1860	WMO 1860	W2	HOP	5
12	Sea-surface temperature	Celsius	Kelvin	SST	DOS	10
91	Total cloud amount	WMO 2700	WMO 2700	N	HOP	11
67	Low cloud amount	WMO 2700	WMO 2700	NH	HOP	11
65	Low cloud type	WMO 0513	WMO 0513	CL	HOP	11
64	Middle cloud type	WMO 0515	WMO 0515	CM	HOP	11
63	High cloud type	WMO 0509	WMO 0509	CH	HOP	11
66	Cloud height	WMO 1600	m	H	HOP	11
84	Significant wave height	0.5 m	m	WH	HOP	12
85	Significant wave period	s	s	WP	HOP	12
86	Significant wave direction	10 degree	degree	WD	HOP	12
76	Ice accretion rate	WMO 3551	WMO 3551	RS	HOP	-
77	Ice thickness	cm	m	ES	HOP	-
78	Ice accretion	WMO 1751	WMO 1751	IS	HOP	-
79	Duration period of Precip	WMO 4019	h	TR	HOP	-
80	Amount of Rain	WMO 3590	$\mathrm{kg\,m^{-2}}$	RRR	HOP	-

Table 14: Geophysical quantities imported from ICOADS. The column vertco indicates what IMMA quantity was used for the vertical coordinate `vertco_reference_1`*. It was encoded as height (m), rather than geopotential height ($\mathrm{m^2 s^{-2}}$), which latter would have been the correct ODB unit. The QC flags are listed in Table 18.*

Observable	Code Table		Converted value	Unit
Ship course	0700	0-9	0, 45, 90, 135, 180, 225, 270, 315, 0, RMISS	degrees
Ship speed (before 1968)	UKMO	0-9	0, 2, 5, 8, 11, 14, 17, 20, 23, 24 knots	$\mathrm{m\,s^{-1}}$
(1968 onwards)	4451	0-9	0, 3, 8, 13, 18, 23, 28, 33, 38, 43 knots	$\mathrm{m\,s^{-1}}$
Cloud height	1600	0-9	25, 75, 150, 250, 450, 800, 1250, 1750, 2250, 2500	m
Duration of precip	4019	1-9	6, 12, 18, 24, 1, 2, 3, 9, 15	h
Amount of Rain	3590	0-999	face value; 0-0.9 from 990-999	$\mathrm{kg\,m^{-2}}$

Table 15: Conversion of (WMO) code tables to ODB units.

Platform Type	Reportype	Description	B	S	O	C
6, 7	16005	Dribu (includes moored buoys)	1	21	4	165
0,1,4,5, UNDEF	16008	Ship	1	11	1	21
18	16011	Dribu-tesac	1	23	4	64
3	16049	Ocean station vessel on station	1	11	1	21
2	16050	Ocean station vessel off station	1	11	1	21
9	16051	Station or ship on ice	1	11	1	21
10	16052	Ocean bottle and low-resolution conductivity temperature depth CTD and XCTD	1	11	1	21
11	16053	Mechanical or digital or micro bathythermograph MBT	1	11	1	21
12	16054	Expandable bathythermograph XBT	1	11	1	21
13	16055	Coastal-marine automated network CMAN	1	11	1	21
14	16056	Coastal or island station	1	11	1	21
15	16057	Fixed ocean platform or rig	1	11	1	21
17	16061	High Resolution Conductivity Temperature Depth CTD and XCTD	1	11	1	21
19	16062	Undulating Oceanographic Recorder UOR	1	11	1	21
20	16063	Autonomous Pinniped Bathythermograph APBT	1	11	1	21
21	16064	Underwater Ocean Glider	1	11	1	21

Table 16: Relation between IFS reportype, derived bufrtype (B), subtype (S), obstype (O), codetype (C), and ICOADS platform type

Nobs	Call Sign	statid	First date	Lat	Lon	Last date	Lat	Lon
852	John D.BR	John DBR	18840114	-23.55	70.39	18840326	41.49	-69.26
752	James S.S	James SS	18800121	-40.77	41.92	18800401	21.42	118.78
1025	James S.L	James SL	18850108	43.38	-69.80	18850531	30.38	-72.83
7372	John D.Br	John DBr	18820502	40.08	-72.26	18850805	-41.79	48.54

Table 17: Conversion from ICOADS Call Sign to ODB statid for deck 704

level, information on station (`stalt@hdr`), anemometer (`anemoht`) and barometer height (`baroht`) has been imported as well.

4.6 Station Identifier

In ICOADS information on station name is provided by the Call Sign. For ODB, this 9-character attribute is to be translated into a 8-character string (`statid`). The following conversion rule is applied:

- Call Sign is first left-adjusted, i.e. all trailing blanks from the left are removed
- `statid` uses the first 8 left-adjusted characters
- when Call Sign does contains 9 characters, and the 8th character of statid is a blank, it is replaced by the 9th character of Call Sign.
- when Call Sign has less than 8 characters, statid is right-filled with blanks.

Index	Description	NCDC flag	Trimming
1	incorrect ship position		ZZQF
2	wind direction/speed		WZQF
3	visibility	BNC	
4	present weather	XNC	
5	past weather	YNC	
6	sea level pressure		PZQF
7	air temperature		AZQF
8	wet bulb temperature		RZQF
9	dew point temperature		RZQF
10	sea surface temperature		SZQF
11	cloud	CNC	
12	wave	ENC	
13	swell	FNC	
14	pressure tendency	TNC	

Table 18: Indicators used for flagging data. Quantities in the enhanced trimming column correspond to those of the trimqc0.f routine from http://icoads.noaa.gov/software/, while NCDC flag refers to IMMA key words.

There are some special cases. For deck 704 (US Marine Meteorological Journals Collection (1878-94)) the deletion of the excess character leads to ambiguities for 4 Call Signs. They are converted according to Table 17. For Deck 780 the blank at location 3 is removed, and any sequence of ”******” is replaced by blanks.

As a result, 1,759,044 different combinations of statid, deck, and platform type remain.

4.7 Quality control flags

ICOADS provides several quality indicators. One example is the set of NCDC flags that provide a range of confidence levels for a number of geophysical quantities. Information from those flags was imported in ODB, and the `datum_status.rejected` flag was set whenever the corresponding NCDC flag exceeded the value 3 (correctable). For quantities for which no NCDC flag is available, the enhanced trimming flag was used, instead. Strictly speaking, this flag only tests whether an observation is inside a plausible climatological range. It is to be determined from a separate routine *trimqc0.f*, which is available from http://icoads.noaa.gov/software/.

Details on what method was used for what quantities is given in Tables 14 and 18. Whenever the rejected bit for `datum_status` was set for any observable within a report, `report_status.rejected` was set as well.

4.8 The collocation of 20CR model values

Other than quality control indicators as discussed in Section 4.7, ICOADS currently does not contain added information from, e.g., a reanalysis. One of the objectives of the ingestion of ICOADS in ODB was to facilitate its usage in the ERA-20C reanalysis. For that purpose a prior break-point analysis, as discussed in Section 2.8.3 is desirable. For this reason it was decided to collocate ICOADS data with 20CR model fields, where available, i.e. from November 1869 until end 2009.

Name	step (h)	Description
LAND_sfc	6	forecast land-sea mask (fraction)
HGT_sfc	6	forecast orography (m)
TMP_sfc	6	forecast SST (K)
UGRD_10m	6	forecast U-comp 10m wind (m/s)
VGRD_10m	6	forecast V-comp 10m wind (m/s)
ICEC_sfc	6	forecast sea ice (fraction)
TMP_2m	3, 6	forecast 2m temperature (K)
PRES_sfc	3, 6	forecast surface pressure (Pa)
HGT_sfc		analysis orography (m)
UGRD_sigma		analysis U-comp sigma=0.995 wind ($m\,s^{-1}$)
VGRD_sigma		analysis V-comp sigma=0.995 wind ($m\,s^{-1}$)
TMP_sigma		analysis Temp at sigma=0.995 (K)
PRMSL_msl		analysis mean sea level pressure (Pa)

Table 19: Used 20CR model fields

Such global synoptic fields were obtained from NCAR (http://rda.ucar.edu/datasets/ds131.1/). These comprise ”Yearly Time Series Analysis Fields”; 6-hourly analyses on a 2x2 degree lat/lon grid, and ”Yearly Time Series First Guess Forecast Fields”; 3 and 6-hour forecasts from the 6-hourly analyses on a T62 Gaussian grid. These fields all relate to the 20CR ensemble mean, i.e., not individual members. They are organized into yearly GRIB files with names ”pgrbanl_mean_*yyyy_param*.grib” and ”sflx-grbfg_mean_*yyyy_param*.grib” for analysis and forecast fields, respectively, and where *param* specifies the geophysical parameter. The list of parameters that were used is given in Table 19. Mostly only the step 6 forecast fields were regarded. For tendencies, step 3 was used as well. At ECMWF, these 20CR surface fields have (together with a number of upper-air fields) recently been archived in MARS (class=nr).

Collocation was based on a Fortran module *interpol*, originating from a collocation code of wind and surface parameters for scatterometer data (Lars Isaksen, private communication). It was adapted to GRIB_API standards and made more modular such that it can handle the interpolation of different grids (like 2x2 lat/lon and T62 Gaussian) on different temporal resolutions simultaneously, and provides direct access to individual GRIB fields in large files to boost performance. Fields are interpolated bi-linearly in space, and in time the closest field is chosen. For the current application it means that the maximum collocation error for the 20CR model fields is 3 hours.

Interpolated 20CR model values have been determined for two classes of ODB columns, departures, and 7 columns for model collocated values (`@modsurf`). Departures were estimated for mean sea level pressure, pressure tendency, 2-metre temperature, wind, and sea-surface temperature (see Table 20). Since no bias estimates were available, the departures are based on uncorrected observed values. The fields as used for the `@modsurf` columns (all 6-hour forecasts) are indicated in table 13.

Besides interpolation, some adjustments to model values were made as well, which were applied to the collocated output values from *interpol*. Requirements were not always transparent and some non-optimal choices may have been made. Details are as follows.

For the 20CR forecast fields, mean sea level pressure MSLP is not available. It was estimated from available forecast surface pressure PRES_sfc, instead, as:

$$\mathrm{MSLP} \approx \mathrm{PRES\_sfc} \times \exp\left(\frac{9.806}{287.04}\frac{\mathrm{HGT\_sfc}}{\mathrm{TMP\_2m}}\right). \qquad (6)$$

Observable	varno	fg_depar	an_depar
Mean sea level pressure	110	PRES_sfc	PRMSL_msl
Pressure tendency	30	PRES_sfc, PRES_sfc-3	PRMSL_msl, PRMSL_msl-6
2m air temperature	39	TMP_2m	TMP_sigma
U Wind component	41	UGRD_10m	UGRD_sigma
V Wind component	42	VGRD_10m	VGRD_sigma
Wind direction	111	UGRD_10m, VGRD_10m	UGRD_sigma, VGRD_sigma
Wind speed	112	UGRD_10m, VGRD_10m	UGRD_sigma, VGRD_sigma
Sea-surface temperature	12	TMP_sfc	

Table 20: Specification of the 20CR model fields that were used for the construction of first-guess and analysis departures.

Reason for this transformation is that orography HGT_sfc can vary substantially over the ocean surface for Gaussian grids at lower resolutions ('spectral ripples' as for T62), especially west of the Andes. For analysis departures, surface analysis MSLP is available (PRMSL_msl), so no conversion was applied here.

Pressure tendencies regard to the change in pressure over the previous 3 hours. For first-guess departures the difference between the 6 and 3-hour forecast from the same analysis was used (each subject to transformation 6). For analysis departures half the difference with the surface pressure for the analysis 6 hours before was used. This may occasionally be quite inaccurate, since part of such differences may reflect jumps due to the assimilation of the latest observations.

For surface winds, spectral ripples can play a role as well. Forecast winds are available at 10-metre height. But they were adapted according to a logarithmic wind profile whenever the local orography was below sea level:

$$(\mathrm{U10},\mathrm{V10})_{\mathrm{fc}} = b_n \times (\mathrm{UGRD\_10m},\mathrm{VGRD\_10m}), \tag{7}$$

where

$$b_n = \log(10/z_0)/\log(z_h/z_0), \qquad z_h = \max(10, 10 - \mathrm{HGT\_sfc}), \tag{8}$$

and z_0 =2.E-4 m is the roughness length at the sea surface for winds of about $8\,\mathrm{m\,s^{-1}}$.

Analysis winds are available at the 0.995 sigma level. Conversion to 10m-height was

$$(\mathrm{U10},\mathrm{V10})_{\mathrm{an}} = b'_n \times (\mathrm{UGRD\_sigma},\mathrm{VGRD\_sigma}), \tag{9}$$

where

$$b'_n = \log(10/z_0)/\log(z'_h/z_0), \qquad z'_h = \max\left(10, 0.005\left(\frac{287.04}{9.806}\right)\mathrm{TMP\_sigma} - \mathrm{HGT\_sfc}\right). \tag{10}$$

Surface wind speed and direction (meteorological convention) were calculated from the U and V components.

For the remaining quantities, i.e., land-sea mask, sea-ice fraction, 2-metre and sea-surface temperature, no transformations were applied. Also, no height corrections were applied for departures regarding observations of temperatures not at 2-metre, and winds not at 10-metre height, respectively.

Report level			Observation level
seqno@hdr	stalt@hdr	source@hdr	
reportype@hdr	statid@hdr	timeseries_index@conv	vertco_type@body
bufrtype@hdr	date@hdr	collection_identifier@conv	obsvalue@body
subtype@hdr	time@hdr	unique_identifier@conv	varno@body
groupid@hdr	lat@hdr	anemoht@conv	bias_volatility@body
obstype@hdr	lon@hdr	baroht@conv	ppcode@conv_body
codetype@hdr		station_type@conv	

Table 21: ODB columns in ERA-20C that were imported from ISPD and ICOADS. For ISPD `anemoht`, `baroht` *and* `station_type` *are not available, and are set to NULL, instead. Whenever* `stalt` *is missing it was replaced by a collocated value from a T3999 reduced-Gaussian (∼ 5 km) model-orography field.*

5 The ERA-20C observation feedback archive

Besides initializing the creation of observation feedback archives for historical data sets at ECMWF, the ISPD and ICOADS ODB's formed a convenient starting point for their ingestion into the ERA-20C reanalysis (Poli *et al.*, 2013). At ECMWF input data conventionally originates from BUFR. Therefore, the part of the suite that deals with such extraction, was extended to allow for the inclusion of ODB files from MARS. Only a required subset of ODB columns is selected (see Table 21) and the structure of reports (i.e., the number and order of rows within each report) is expanded to the strict schema imposed by IFS. Next the new-style ODB files are converted to the old-style ODB hierarchal representation (as required inside the assimilation) and are combined with ODB data originating from BUFR, if any. From that point onwards, no further adaptations are required. Data is screened and assimilated as usual, resulting in new-style ODB files on output containing feedback information from ERA-20C. These are subsequently archived into MARS using the expver of the ERA-20C suite rather than that of the input ISPD and ICOADS archives. Only pressure observations and marine winds were assimilated with weights (observation error) that were prescribed inside IFS; all other observables were blacklisted. Details on the assimilation may be found in Poli *et al.* (2013).

Just prior to the production of ERA-20C a new version of ISPD had become available. This set, ISPDv3.2.6 included several new collections, such as some data digitized during the ERA-CLIM project and marine pressure observations from the Oldweather.org initiative. It was decided to use this version of ISPD, rather than ISPDv2.2. It was converted to ODB similar to ISPDv2.2 (but `source`='ISPD3.26') and archived into MARS (expver=1722). One practical difference is that no feedback from 20CR was available. For that reason, departures were constructed by collocation with 20CR model fields, based on the method as used for ICOADS (see Section 4.8). Not all columns as in Table 7 were produced, i.e. no status and event flags, and no `biascorr` column. It does have one extra column, `lsm@modsurf`, containing the collocated land-sea fraction from 20CR. Bias volatility was created on the basis of the constructed first-guess departures. A comparison for a number of stations that are also available in ISPDv2.2 showed very similar results.

The ERA-20C OFA embraces all columns as listed in Tables 4, 7 and 13, with the exception of `anemoht` and `baroht`. In addition, five more columns were stored (see table 22).

The traceability columns as set in the ISPD and ICOADS input archives allow for a separation between source ('ISPD3.26' or 'ICOADS2.5'), collections and individual reports in the ERA-20C OFA as well.

As can be seen from Table 21, any feedback information from ISPD and ICOADS is ignored on input. Such columns (see Table 4) are filled with information gathered during the assimilation in ERA-

Column@table	Type	Description
enda_member@desc	INTEGER	0
numtsl@desc	INTEGER	49, number of time slots in assimilation window
timeslot@timeslot_index	INTEGER	1-49, time slot in assimilation window
sonde_type@conv	INTEGER	0 (type of radiosonde)
solar_elevation@conv	REAL	solar elevation (degrees)
datum_qcflag@conv_body	REAL	1.0: indicates constant pressure series

Table 22: ODB columns for ERA-20C that are not listed in Tables 4, 7, or 13.

20C (flags, bias corrections, departures, collocated first-guess model values). Details can be found in the IFS documentation (from www.ecmwf.int/en/research/). Exception are `timeseries_index` and `bias_volatility` which are read on input and used inside the ERA-20C variational bias correction scheme for grouping and adaptivity of pressure observations, respectively.

In addition, `vertco_reference_1` was not ingested in ERA-20C. It was calculated during the screening in ERA-20C from `stalt`, by multiplication with $g_0 = 9.80665\,\mathrm{ms}^{-2}$ to obtain geopotential height. Therefore, the error made in the ICOADS and ISPD OFA', where height (in m) was used, did not affect the assimilation of surface pressure data. Whenever `stalt` was not available on input, a collocated value from a T3999 reduced-Gaussian ($\sim 5\,\mathrm{km}$) model-orography field was used.

The column `datum_qcflag@conv_body` has value 1.0 if surface pressure observations are found to be constant. This is assessed on a station basis, and reporting practice basis (ie., treating separately sea-level reporting and station-level reporting), over a period of 5 days (± 2 days centered around the day of interest), when a minimum of 4 observations are reported. Using the information contained in this column to filter the surface pressure observation feedback data reveals for example that station "CAYLOBOS" (Cay Lobos, a lighthouse off the coast of Cuba in the Caribbean) in ISPD3.2.6 reported 36 times a constant pressure of 1027.77 hPa between 7 and 13 March 1900. Such information could be used to isolate suspicious data and further enhance ISPD and ICOADS."

6 Concluding remarks

The ISPD2.2, ICOADSv2.5 and ERA-20C data sets are the first data sets that were ingested into the observation feedback archive and made publicly available on the ECMWF public data server. The OFA data model, ODB is very flexible and generic since its lowest hierarchal level is based on single observations rather than entire reports, The model feedback contains important information on data quality.

For ISPD and ICOADS, the conversion from the native format did incur some challenges. These were not related to the translation between formats, but to the translation of information. It was not always clear what piece of information should go where in ODB. This mainly applied to meta data and the difference in conventions used for quality indicators. Translation of geo-location and observed values was more straightforward.

For the historical data sets some new attributes (ODB columns) were created that facilitate traceability and grouping of reports. An attribute bias volatility was introduced as well which, for the sets considered here is based on 20CR feedback, and could facilitate an automatic system for break detection. It was used as input to the ERA-20C variational bias correction scheme.

The three data sets presented in this document form the start for a longer list of observation feedback

archives. One example is the conversion of the ERA-Interim reanalysis data archive from the original ODB representation, which is currently in progress. Another activity is the ingestion of (early) upper-air data sets, and a number or reprocessed satellite data sets from their native formats, where the conversion to BUFR is troublesome, and which will serve as input to the next ECMWF reanalysis ERA5.

7 Acknowledgements

The work described here is part of the ERA-CLIM project, which was funded by the Seventh Framework Programme of the European Union under grant agreement no. 265229. Support for the International Surface Pressure Databank is provided by the U.S. Department of Energy, Office of Science Innovative and Novel Computational Impact on Theory and Experiment (DOE INCITE) program, and Office of Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration Climate Program Office. The International Comprehensive Ocean-Atmosphere Data Set has been jointly developed by NOAA's Earth System Research Laboratory, NOAA's National Climate Data Center, and the UCAR's National Center for Atmospheric Research. We thank Gil Compo, Chesley McColl, Thomas Cram and Jeffrey Whitaker for their invaluable help and provision of the ISPDv2.2 and ISPDv3.2.6 data sets and 20CR reanalysis fields. We thank Scott Woodruff and Steve Worley for their essential help regarding ICOADSv2.5.1, and support on the retrieval of all data sets from the NCAR data portal. We thank Leo Haimberger, Gil Compo and Manuel Fuentes for proofreading this document. We would also like to thank Anne Fouilloux and Peter Kuchta for their essential help regarding ODB, and Paul Dando and Xavier Abellan Ecija for their always excellent solutions to computer-related problems.

A Rewriting of the bias volatility

Relation (4) is obtained by expressing μ_0 and σ_0 in terms of quantities of the two sub sets 1 and 2. For μ_0 this is rather trivial:

$$\mu_0 = \frac{n_1\mu_1 + n_2\mu_2}{n_1 + n_2}.$$

From this it follows that:

$$\mu_0 - \mu_1 = \frac{n_1\mu_1 + n_2\mu_2 - (n_1 + n_2)\mu_1}{n_1 + n_2} = \frac{n_2}{n_1 + n_2}(\mu_2 - \mu_1),$$

similarly for $\mu_0 - \mu_2$, so:

$$(\mu_1 - \mu_0)^2 + (\mu_2 - \mu_0)^2 = \frac{n_1^2 + n_2^2}{(n_1 + n_2)^2}(\mu_2 - \mu_1)^2. \tag{11}$$

For the standard deviation σ_0 some more work is required. For any standard deviation σ one has,

$$\sigma^2 = s/n - \mu^2, \quad \text{or} \quad s = n\sigma^2 + n\mu^2,$$

where s is the sum of squares. Then:

$$\begin{aligned} n_0^2\sigma_0^2 &= n_0 s_0 - (n_0\mu_0)^2 = n_0(s_1 + s_2) - (n_1\mu_1 + n_2\mu_2)^2 \\ &= n_0(n_1\sigma_1^2 + n_2\sigma_2^2) + n_0 n_1\mu_1^2 + n_0 n_2\mu_2^2 - (n_1\mu_1 + n_2\mu_2)^2 \\ &= n_0^2\sigma_m^2 + n_2 n_1\mu_1^2 + n_1 n_2\mu_2^2 - 2n_1 n_2\mu_1\mu_2 \\ &= n_0^2\sigma_m^2 + n_1 n_2(\mu_1 - \mu_2)^2 \end{aligned}$$

Or:

$$\sigma_0^2 = \sigma_m^2 + \frac{n_1 n_2}{(n_1+n_2)^2}(\mu_2 - \mu_1)^2. \tag{12}$$

So σ_0^2 is *not* just the weighted average of the standard deviations of the individual time series but has an additional term that depends on the difference in the means.

Division of (11) by (12) and (3) leads to (4).

References

Alexandersson, H. (1986). A homogeneity test applied to precipitation. *IJC*, **6**, 661675.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., ø Nordli, Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J. (2011). The twentieth century reanalysis project. *Quart. J. Roy. Meteor. Soc.*, **137**, 1–28, doi:10.1002/qj.776.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Bronnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, O., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J. (2010). International surface pressure databank (ispdv2). URL http://dx.doi.org/10.5065/D6SQ8XDW.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Bronnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, O., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D. and Worley, S. J. (2009). Noaa cires twentieth century global reanalysis version 2. URL http://dx.doi.org/10.5065/D6QR4V37.

Cram, T. A., Compo, G. P., Yin, X., Allan, R. J., McColl, C., Vose, R. S., Whitaker, J. S., Matsui, N., Ashcroft, L., Auchmann, R., Bessemoulin, P., Brandsma, T., Brohan, P., Brunet, M., Comeaux, J., Crouthamel, R., Gleason, B. E., Groisman, P. Y., Hersbach, H., Jones, P. D., Jónsson, T., Jourdain, S., Kelly, G., Knapp, K. R., Kruger, A., Kubota, H., Lentini, G., Lorrey, A., Lott, N., Lubker, S. J., Luterbacher, J., Marshall, G. J., Maugeri, M., Mock, C. J., Mok, H. Y., Nordli, Ø., Rodwell, M. J., Ross, T. F., Schuster, D., Srnec, L., Valente, M. A., Vizi, Z., Wang, X. L., Westcott, N., Woollen, J. S. and Worley, S. J. (2015). The international surface pressure databank version 2. *Submitted to Geoscience Data Journal*.

Dee, D., Balmaseda, D. M., Balsamo, G., Engelen, R. and Simmons, A. (2012). Toward a consistent reanalysis of the climate system. *Ecmwf research department memorandum*, ECMWF, Shinfield Park, Reading, 687.

Haimberger, L. (2007). Homogenization of radiosonde temperature time series using innovation statistics. *JCLI*, **20**, 1,377–1,403.

Poli, P., Hersbach, H., Tan, D., Dee, D., Thépaut, J.-N., Simmons, A., Peubey, C., Laloyaux, P., Komori, T., Berrisford, P., Dragani, R., Trémolet, Y., Holm, E., Bonavita, M., Isaksen, L. and Fisher,

M. (2013). The data assimilation system and initial performance evaluation of the ecmwf pilot reanalysis of the 20th-century assimilating surface observations only (era-20c). *Ecmwf era report series*, ECMWF, Shinfield Park, Reading, 14.

Saarinen, S. (2004). Odb user guide. *www.ecmwf.int*, ECMWF, Shinfield Park, Reading.

Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Freeman, J. E., Berry, D. I., Brohan, P., Kent, E. C., Reynolds, R. W., Smith, S. R. and Wilkinson, C. (2010). Icoads release 2.5: extensions and enhancements to the surface marine meteorological archive. *International Journal of Climatology*, doi:10.1002/joc.2103.