ECMWF Feature article

METEOROLOGY

Recent developments in the automatic checking of Earth system observations
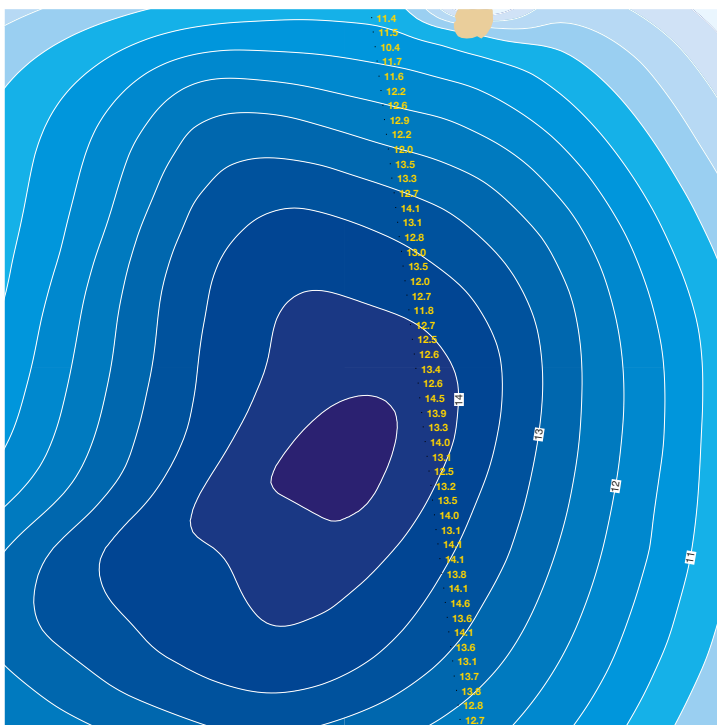
Recent developments in the automatic checking of Earth system observations

Mohamed Dahoui, Niels Bormann, Lars Isaksen, Tony McNally

For the last few years, an automatic data checking system has been used at ECMWF to monitor the quality and availability of observations processed by the atmospheric data assimilation system (Dahoui et al., 2014). Recently the system has been upgraded to add support for all Earth system observations processed by data assimilation systems running at ECMWF. This includes the 4D-Var system for the atmosphere, land surface data assimilation (LDAS) and ocean data assimilation (OCEAN5). The new framework also makes it easier to cross-check warnings from all components of the observing system and to robustly distinguish between data issues and limitations related to the model and/or data assimilation. Given the increasing size and diversity of the observing system, this tool is playing an essential role in flagging up observation issues and enabling the timely triggering of mitigating actions. On many occasions, the warnings generated by the automatic data checking system have been useful in highlighting unusual events and model or data assimilation limitations. An example is shown in Figure 1. In this article, we describe the most important features of the new system and provide a list of supported observation datasets. The automatic data checking system is partially supported by the EUMETSAT Satellite Application Facility for numerical weather prediction (NWP SAF). The system is chiefly used internally at ECMWF but important notifications are shared with selected users from EUMETSAT and the NWP SAF consortium.
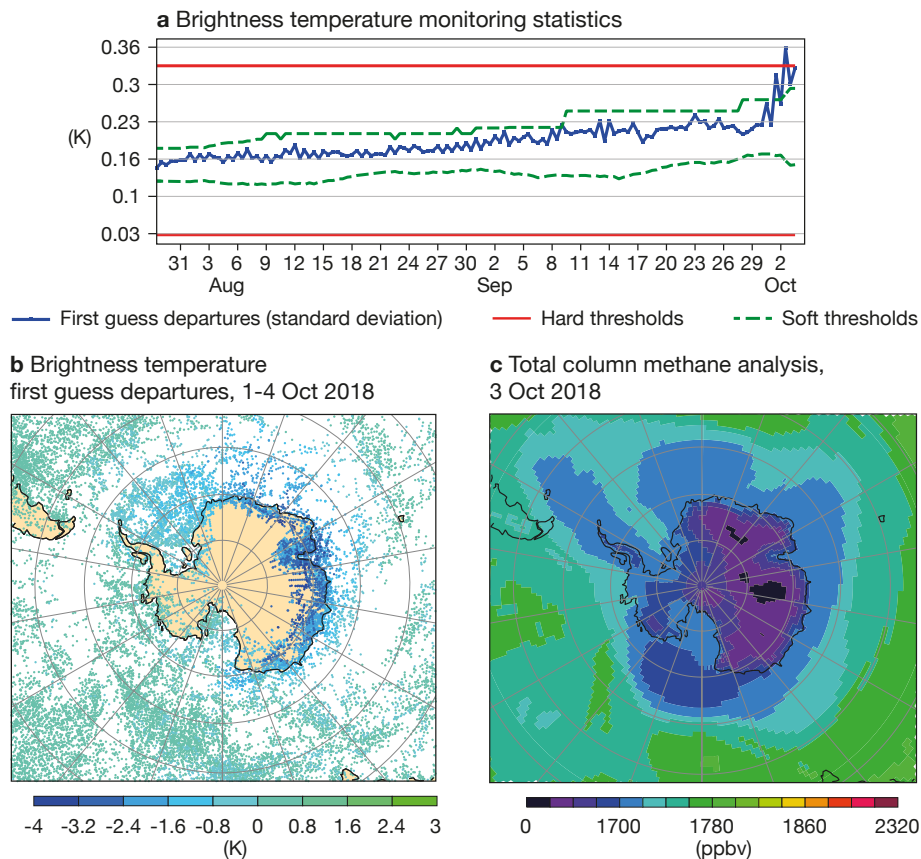

Figure 1 These three charts illustrate how the automatic checking system flagged up a case of unusually low concentrations of methane over Antarctica combined with warmer stratospheric conditions. They show (a) a time series of the standard deviation of first guess departures for brightness temperatures from the S-NPP CrIS satellite instrument, channel 789 (7.6 microns), (b) a map of first guess departures for the same channel averaged over the period from 1 to 4 October 2018, and (c) the analysis of total column of methane (in ppbv) for 00 UTC on 3 October 2018 provided by the Copernicus Atmosphere Monitoring Service implemented by ECMWF. The large first guess departures are due to low methane concentrations combined with relatively high stratospheric temperatures in Antarctica.

Design of the new system

This significant upgrade of the automatic data checking system had three main objectives: to extend support to all Earth system observations (atmosphere, ocean and land surface); to improve the reliability of the system (by reducing the false alarm rate and maximising the detection efficiency); and to modernise the underlying software to enhance modularity and simplify maintenance. The new system is based on a mixture of Python and shell scripting. The modular design of the previous system has been preserved and improved. It is based on four main components: data retrieval; threshold computation; data checking; and warning notifications (see Figure 2). Warnings are archived in an event database that is subsequently used for web publishing and blacklist generation. The system interfaces directly with the ODB (ECMWF Observations Database) for inputs. Intermediate statistics and thresholds are also based on the ODB format. This enables the software to use ECMWF ODB-API tools for encoding and filtered decoding. Most of the settings are controlled by a few configuration files, which makes it easy to maintain the system and to add support for new data types.
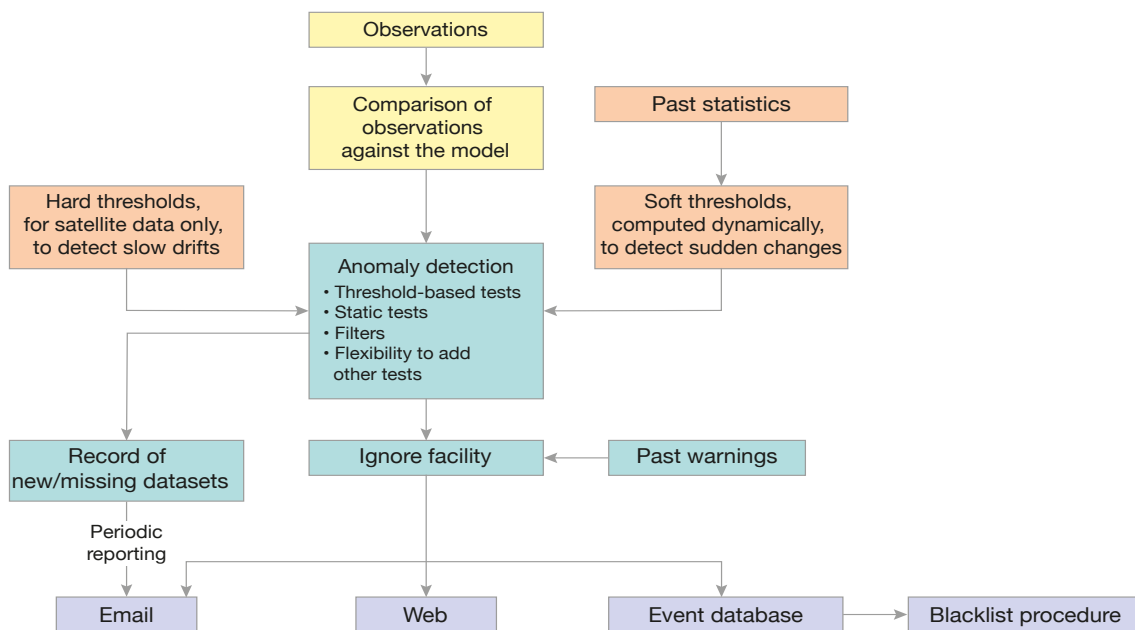

Figure 2 High-level diagram of the automatic checking system.

Computing the thresholds

Warnings may be issued when the observations are found to cross certain thresholds. There are two types of thresholds: flexible or 'soft' thresholds, which are used to detect sudden changes in the data, and 'hard' thresholds, which are used to detect slow drifts in the data. The computation of soft thresholds is based on the past 10 days, excluding the last two days. Outliers are excluded by default. Experience shows that this is best achieved by keeping data between the 10th and 80th percentiles. The soft thresholds can optionally be dependent on the analysis cycle. This is currently applied to data types for which the diurnal cycle is important, or data availability is time dependent. As in the previous system, the soft thresholds are bound by hard thresholds. The computation of hard thresholds is performed on demand when needed and automatically on a three-monthly basis to adjust to seasonal variability. As in the previous system, hard thresholds are not used for in-situ data, whether from individual stations or area-based statistics. This choice was initially made to simplify the maintenance of the system. However, there are plans to introduce hard thresholds for in-situ data over geographical domains to persist warnings due to soft thresholds being exceeded, as the soft thresholds adjust quickly to persistent changes in the data.

One of the new aspects of the system is the possibility to automatically monitor satellite data over specific geographical areas in addition to the global domain. This has the advantage of distinguishing between data-related issues and model or data assimilation limitations. Problems flagged up over all areas are clearly due to anomalies with the data. Warnings limited to one domain or two neighbouring ones are indicative that the root cause of the deviations is not data related. The new system also monitors selected satellite data over land and sea separately in addition to all surface types combined. Currently warnings specific to land or sea areas are ignored pending an improved tuning of filters, to avoid too many insignificant alarms. For all data types, the Ensemble of Data Assimilations (EDA) spread is systematically checked. Doing so helps to distinguish between data issues on the one hand and model or data assimilation issues on the other. Increased EDA spread is usually a sign that large departures between observations and the short-range forecasts used in the data assimilation system (first guess departures) are due to large model uncertainty. For all data apart from satellite observations, the probability of gross error (PGE) is systematically checked in addition to other observation diagnostics (such as first guess departures). PGE is available for all in-situ data used in the atmospheric 4D-Var data assimilation system. For the other data assimilation systems (OCEAN5 and land surface data assimilation), PGE is computed by the automatic checking system using first guess departures and prescribed observation errors. When the root-mean-square of differences between observations and the analysis (analysis departures) is higher than that of first guess departures, a test is performed to check if this is usual behaviour (as is the case for some non-used satellite data) before generating a warning. A test is also performed to assess the magnitude of the analysis departures. A warning is triggered only if the departures are estimated to be large.

For all data types and for all observation quantities involved, the main test is based on the exceedance of computed thresholds. Such a test inevitably generates several small non-persistent deviations that need to be filtered out, given the large number of checked datasets. To make this possible, several additional static tests are added to either consolidate, discard or temporarily ignore insignificant warnings. The tests are designed based on our experience of running the system and they are expected to evolve with time.

Delivering warnings

If the thresholds are exceeded, a severity level is determined by computing the ratio of the difference between the checked statistics and the mean to the difference between these statistics and the closest threshold (upper or lower).

All flagged warnings are processed by an ignore module designed to filter out known issues. In the previous automatic checking system, the ignore module was limited to a static list updated manually. In the new system, three forms of ignore lists are available:

- Manual ignore list: this is updated manually for known issues (e.g. expected outages).

- Static ignore list: this is generated alongside the hard thresholds and reflects the long-term behaviour of the data. Data usually present in only one analysis cycle are not reported as missing for the other analysis cycles (for example in the case of radiosondes reporting once a day). Data persistently present in small numbers or intermittently present are ignored when their counts are below certain values. Data quality warnings are always reported.

- Dynamic ignore list: this is based on recent availability and usage of the data. The list depends on the analysis cycle. Warnings related to recently blacklisted or missing data will cease after a few days. If the data are reactivated their quality will be checked and reported.

The system keeps records and reports about new/missing data (even for in-situ data from individual stations). The list is continuously made available to selected users. Time series are generated for all data affected by warnings (even missing datasets). For in-situ data, a map is generated to show the geographic distribution of affected stations. When warnings are generated, a module processes them to reduce where possible the number of events to be communicated to users and to alter the severity levels in certain conditions:

- If warnings are generated for the global domain and other geographical domains, then only global warnings are retained.

- If warnings are affecting all surface types as well as sea and land separately, then only warnings affecting all surface types are retained.

- If the same warnings are affecting many channels, then only one message is retained with an indication of the number of channels affected.

- Warnings are ranked by their severity.

- 'Severe' warnings are escalated to 'severely persistent' if the same warning occurred more than six times during the past ten days.

- Some warnings are ignored because they are not severe enough, but if they are persistent then the system will eventually communicate them to users. This approach helps to limit the number of warnings communicated every day, especially for individual in-situ stations.

The delivery of warnings also depends on the choices made by users of the system. Users can for example specify data types of interest and the severity levels to be applied for warnings to be delivered to them. For most users, the best option is to receive severe warnings only.

Supported data types

Adding support for new observation types is relatively simple. Raw sea-surface temperature (SST) and sea ice data from the UK Met Office's OSTIA product (i.e. data as received) are checked differently from other observations. The check is based on day-to-day variability in the OSTIA fields themselves instead of differences between those fields and the model. Observations supported by the system are summarised in Table 1.

Data type	Geophysical parameter	Vertical resolution	Geographical domains	Surface types
Radiances (all satellites used)	Clear-sky and all-sky radiances	Channels	Global and 5 large domains	Land; sea; and all surface types combined
Wind (all satellites used)	Wind vector difference	Pressure layers every 300 hPa	Global and 5 large domains	All surface types combined
Surface wind (all satellites used)	Wind vector difference	Surface	Global and 5 large domains	Sea
GPS radio occultation	Bending angles	Every 2 km	Global and 5 large domains	All surface types combined
Ozone (all satellites used)	Ozone	Total column and pressure layers (every 20 hPa)	Global and 5 large domains	All surface types combined
Aircraft, Radiosondes, Pilot sondes and profilers	Wind vector difference, temperature and specific humidity (if available)	Pressure layers with 300 hPa binning	Global and 14 large domains	All surface types combined
SYNOP weather stations, METARs (Meteorological Aerodrome reports), SHIP and Buoys	Surface pressure, 2 m temperature (LDAS) and 2 m humidity (LDAS)	Surface	Global, 16 large domains and all WMO blocks	
NexRad	Precipitation	Surface	Global	Land
SNOW	Snow depth	Surface	Global, 5 large domains and all WMO blocks	Land
ARGO floats, Mooring buoys, CTD devices and XBT thermographs	Salinity and Potential temperature (OCEAN5)	Ocean depth with 100 m binning		Sea
Altimeters	Sea-level anomaly (OCEAN5)	Surface	Global and 14 large domains	Sea
OSTIA sea ice	Sea-ice concentration (OCEAN5)	Surface	North Pole, South Pole, HUDSON, Great Lakes, BALTIC	Sea
OSTIA SST (raw data)	SST	Surface	Large number of small domains	Sea
OSTIA sea-ice (raw data)	Sea-ice concentration	Surface	North Pole, South Pole, HUDSON, Great Lakes, BALTIC	Sea

Table 1 List of datasets supported by the automatic checking system.

Ongoing developments

Given that most of the checking tests are based on first guess departures, and considering model limitations and the effects of atmospheric variability, the warnings that are generated are not necessarily related to observation problems. This means that, in principle, the automatic checking system can fulfil two functions. Its main function is, of course, to prevent the data assimilation system from using poor-quality observations. But the system can also help to identify unusual events and systematic problems in the model or in the data assimilation system. The current implementation of the automatic checking system can partially achieve this objective thanks to the inclusion of the EDA spread and area-based checks. Work is ongoing to exploit redundancies and similarities in the observing system (similar instruments, instruments sensitive to conditions in similar atmospheric layers, etc.) in order to improve the filtering of warnings. For example, a module is being added to cross-check warnings from all components of the observing system against each other using a decision tree algorithm (Figure 3). The system could also exploit information on the past frequency of similar events.
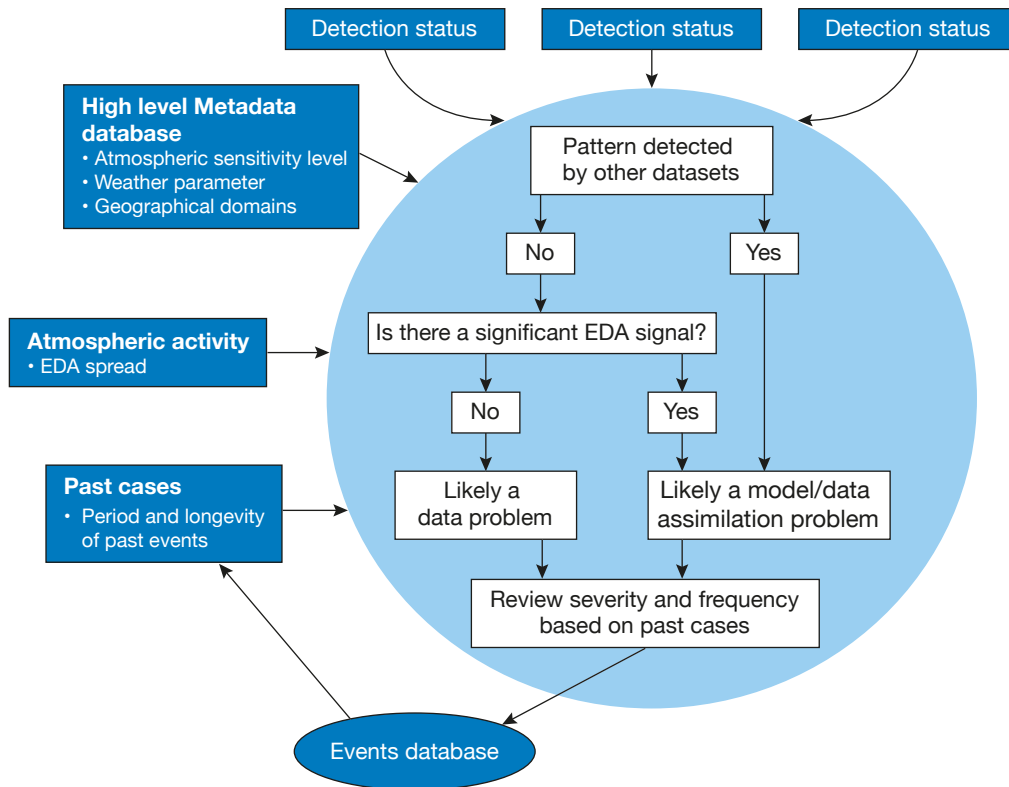
Figure 3 Decision tree for the cross-checking of warnings against other datasets.

Conclusion

The automatic checking system plays an essential role in protecting the data assimilation system from using poor-quality observations. The recent upgrade of the system has extended support to all Earth system observations used at ECMWF. Further improvements of the system are expected to make it more useful for diagnostic purposes by flagging atmospheric patterns systematically triggering warnings (e.g. onset of sudden stratospheric warming events, orographic gravity waves, mesoscale convective systems, etc.).

Further reading

Dahoui, **M.**, **N. Bormann** & **L. Isaksen**, 2014: Automatic checking of observations at ECMWF. *ECMWF Newsletter* **No. 140**, 21–24.