ESA CONTRACT REPORT

Contract Report to the European Space Agency

Operational Assimilation of Space-borne Radar and Lidar Cloud Profile Observations for Numerical Weather Prediction

WP-3000 report: Assimilation system development for cloud radar and lidar observations

M. Janisková, M. Fielding, M. Crepulja, D. Vasiljevič, T. Král and P. Lean

ESA ESTEC contract 4000116891/16/NL/LvH

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen terme

Series: ECMWF ESA Project Report Series

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/en/research/publications

Contact: library@ecmwf.int

Contract Report to the European Space Agency

Operational Assimilation of Space-borne Radar and Lidar Cloud Profile Observations for Numerical Weather Prediction

WP-3000 report: Assimilation system development for cloud radar and lidar observations

Authors: M. Janisková, M. Fielding, M. Crepulja, D. Vasiljevič, T. Král and P. Lean

ESA ESTEC contract 4000116891/16/NL/LvH

October 2017

ABSTRACT

This report details the developments to the assimilation system at European Centre for Medium Range Weather Forecasts (ECMWF) in preparation for the inclusion of cloud radar and lidar data directly into the Four-Dimensional Variational (4D-Var) system. The work is divided into three sections. Firstly, the off-line observation handling of raw satellite data into Binary Universal Format (BUFR) is outlined. The new variable names required for the Observation Data Base (ODB) are also described. Secondly, the extensive developments to the assimilation system are reported. Finally, the system is tested using CloudSat and CALIPSO radar reflectivity and attenuated backscatter respectively.

Contents

1 Introduction

The operational assimilation of Earth, Clouds, Aerosols and Radiation Explorer (EarthCARE; Illingworth et al., 2015) cloud radar and lidar observations into the ECMWF Four-Dimensional Variational (4D-Var) assimilation system poses many scientific challenges. In addition to tackling these challenges, significant technical developments are also required to transform the raw observations received from the satellite into a format that is understood and useful. In particular, while the profiling nature of the observations will provide a wealth of information on the vertical structure of clouds to the numerical weather prediction (NWP) model used at ECMWF, such an observation type has not been assimilated previously.

The necessary technical developments can be roughly separated into two sections. The first involves 'off-line' data handling where routines to convert the raw satellite data into the format recognised by the system need to be written. The second set of developments involve coding changes and additions to the 4D-Var data assimilation system itself. The Integrated Forecasting System (IFS) at ECMWF uses a 4D-Var data assimilation technique (Rabier et al., 2000) based on the incremental formulation (Courtier et al., 1994) to reduce computational cost. In the incremental approach the highest possible resolution is used for the computation of the model trajectory, whereas a lower-resolution model (its adjoint and tangent linear) is used for the iterative and relatively costly computation of analysis increments. Updates and modifications will be required at all of these steps. Once both streams of developments are complete, all the modifications will subsequently need to be tested thoroughly.

As EarthCARE observations will not be available until after the satellite is launched, the initial data handling and testing will use CloudSat (NASA's cloud radar mission; Stephens et al., 2002) and CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations; Winker et al., 2009). Although in a constellation rather than on-board the same satellite, the CloudSat radar and CALIOP (Cloud-Aerosol Lidar with Orthogonal Polarization) lidar are sufficiently similar to the EarthCARE CPR (Cloud Profiling Radar) and ATLID (ATmospheric LIDar) to be used for testing and feasibility studies. The technical developments specific to the EarthCARE CPR and ATLID will be made in WP5000.

This document outlines the technical developments required for the assimilation of EarthCARE-like observations of cloud radar and lidar. In Section 2, the technical pre-processing and handling developments are defined. Section 3 details the developments of the 4D-Var assimilation system. Section 4 gives examples of the testing of the system. A summary concludes the report in Section 5.

2 Pre-processing and handling developments

All observational data to be assimilated in the ECMWF operational system follows the same procedures. As the data arrives, it is converted to Binary Universal Form (BUFR) and stored locally. The level of processing from its source format is typically limited to initial quality checks and the precision may be reduced to save storage space. Once the cut-off time for observations to be included in the analysis has passed, the observations required for assimilation are then included in the Observation Data Base (ODB). The ODB is unique to each analysis or experiment. The role of the ODB is to provide fast I/O to the assimilation system on all observation related data, including the errors and bias correction.

2.1 BUFR definitions

BUFR is a WMO standard for transmitting and storing observed all kinds of meteorological data. Its flexibility lies with its use of 'data descriptors', which are used to access data values. Metadata is stored in external table files. Variables stored within BUFR must be selected from a finite pre-existing list of observation types. New variable types must be approved by WMO. The list of data descriptors for each observation type is known as a 'BUFR sequence'.

Table 2.1 shows the BUFR sequence created for CloudSat observations. Many of the descriptors are generic, such as the time and geolocation descriptors, however the more specific descriptors, such as 'Cloud radar reflectivity', are new variable types and will need to be approved by WMO for operational use. For feasibility studies, the codes can be used locally with no loss of functionality. The BUFR sequences have also been chosen to be applicable to both CloudSat and EarthCARE; the satellite identifier descriptor can be used to infer the origin of the data and processing flags are included to differentiate between the level of processing of the data.

The corresponding BUFR sequence for CALIPSO is showing in Table 2.2. There are many similarities to the CloudSat BUFR sequence, with the differences predominantly related to the observation variable itself (i.e. attenuated backscatter). To allow the BUFR sequence to be used for aerosol observations, the data classification type includes the option for aerosol and also 'Extinction coefficient' is included as a descriptor.

2.2 ODB definitions

While BUFR is extremely efficient for storing data, a different format with fast I/O is required for operational data assimilation. At ECMWF, the ODB is in-house data storage software to allow the 4D-Var system within IFS to store and access data. ODB is formulated on fast and efficient Structured Query Language (SQL) to define and retrieve observational data. There is no unique centralized ODB database: a new ODB is created each analysis time a 4D-Var analysis is made. Each ODB is stored locally in the ECFS (ECMWF's File Storage System) for post-processing and evaluation. Any new ODB variable name codes must be approved by the ODB Governance to ensure consistency between ECMWF, member states and collaborators.

The ODB contains all the input data that is needed by the data assimilation system. Therefore, in addition to the observation values, information such as the corresponding observation errors and bias correction need to be stored. The actual observed values of radar reflectivity and lidar backscatter are averaged to the nearest model grid box in a process known as 'SuperObbing' (see Sec. 3.2.4). These averaged values are stored in the 'obsvalue' variable (see e.g., Table 2.3). The dimension of 'obsvalue' is fixed to the number of model levels, although not every model level is likely to contain a valid observation.

The total observation error is stored in 'final_obs_error', but a breakdown of the error into representativity error, measurement error and forward model error is made possible by the inclusion of 'obs_error' and

Code	Description	Scale	Ref.	Width	Units	Comment
0 01 007	Satellite identifier	0	0	10		satID=788
0 02 019	Satellite instruments	0	0	11		instrumentID=309
3 01 011	Year, month, day					
3 01 013	Hour, minute, second					
3 01 021	Latitude / Longitude (high accuracy)					
0 10 033	Altitude (Platform to Ellipsoid)	1	0	27	m	
0 25 182	L1 processing flag					
0 25 181	L2 processing flag					
0 21 194	Data classification type	0	0	4	CODE TABLE	0 Surface 1 Cloud likely 2 Cloud probable 3 Cloud possible 4 Unclassified 15 Missing value
0 33 003	Quality information				CODE TABLE	
0 08 049	Number of observations					
0 21 195	Cloud fraction	3	0	11		
0 31 000	Delayed replication factor					
0 02 153	Satellite Channel Centre Frequency	-8	0	26	Hz	94 GHz
0 21 197	Height	0	-1000	17	m	
0 21 192	Cloud radar reflectivity	2	-9000	15	dBZ	Range: -90.00 to 237.00 dBZ
0 21 193	Cloud radar reflectivity uncertainty	2	0	9	dB	Range: 0 to 50.00 dB

Table 2.1: BUFR sequence for CloudSAT observations.

Code	Description	Scale	Ref.	Width	Units	Comment
0 01 007	Satellite identifier	0	0	10		satID=787
0 02 019	Satellite instruments	0	0	11		instrumentID=303
3 01 011	Year, month, day					
3 01 013	Hour, minute, second					
3 01 021	Latitude / Longitude (high accuracy)					
0 10 033	Altitude (Platform to Ellipsoid)	1	0	27	m	
0 25 182	L1 processing flag					
0 25 181	L2 processing flag					
0 21 194	Data classification type	0	0	4	CODE TABLE	0 Surface 1 Cloud 2 Aerosol 3 Unclassified 15 Missing value
0 33 003	Quality information				CODE TABLE	
0 08 049	Number of observations					
0 21 195	Cloud fraction	3	0	11		
0 31 000	Delayed replication factor					
0 02 153	Satellite Channel wavelength	9	0	16	m	
0 21 197	Height	0	-1000	17	m	
0 15 063	Attenuated backscatter	2	-9000	15	$m^{-1}\,sr^{-1}$	Range: 0 to 0.1 $m^{-1}\,sr^{-1}$
0 15 064	Uncertainty in attenuated backscatter	2	0	9	$m^{-1}\,sr^{-1}$	Range: 0 to 0.1 $m^{-1}\,sr^{-1}$
0 15 067	Extinction coefficient	9	0	30	m^{-1}	Range: 0 to 0.1 m^{-1}
0 15 068	Uncertainty in extinction coefficient	9	0	30	m^{-1}	Range: 0 to 0.1 m^{-1}

Table 2.2: BUFR sequence for Calipso observations.

Varname	Parent	Dimension	Description
obsvalue	body	model_levels (vertco_type 5)	Radar reflectivity averaged to model grid and level
biascorr	body	model_levels (vertco_type=5)	offline bias correction to be applied to radar reflectivity
datum_status	body	model_levels (vertco_type=5)	flag for storing screening/blacklisting/quality control information
standard_deviation	superob	model_levels (vertco_type=5)	Standard deviation of radar reflectivity within model grid and level
n_obs	superob	model_levels (vertco_type=5)	number of samples used to compute obsvalue and standard_deviation
cloud_fraction	superob	model_levels (vertco_type=5)	number of cloudy points defined by cloud mask divided by number of samples
repres_error	errstat	model_levels (vertco_type=5)	Flow dependent representativity error
obs_error	errstat	model_levels (vertco_type=5)	Measurement error
final_obs_error	errstat	model_levels (vertco_type=5)	Combination of meaurement, representativity and forward model error
surface_pressure	modsurf	scalar	background surface pressure from model
lat	hdr	scalar	Average latitude of observations
lon	hdr	scalar	Average longitude of observations
stalt	hdr	scalar	Height of satellite above sea level

Table 2.3: Selected ODB variable definitions for CloudSAT observations.

'repres_error', which will be useful for the analysis of experiments. Also useful for analysis are the 'standard_deviation','n_obs' and 'cloud_fraction' variables. Note that much of the ODB variable definitions framework for CALIPSO observations (Table 2.4) is identical to that for CloudSat (Table 2.3), the interpretation of some variables differs, such as the 'datum_status' variable, which will contain different flags.

Varname	Parent	Dimension	Description
obsvalue	body	model_levels (vertco_type=5)	Lidar backscatter averaged to model grid and level
biascorr	body	model_levels (vertco_type=5)	offline bias correction to be applied to lidar backscatter
datum_status	body	model_levels (vertco_type=5)	flag for storing screening/blacklisting/quality control information
standard_deviation	superob	model_levels (vertco_type=5)	Standard deviation of lidar backscatter within model grid and level
n_obs	superob	model_levels (vertco_type=5)	number of samples used to compute obsvalue and standard_deviation
cloud_fraction	superob	model_levels (vertco_type=5)	number of cloudy points defined by cloud mask divided by number of samples
repres_error	errstat	model_levels (vertco_type=5)	Flow dependent representativity error
obs_error	errstat	model_levels (vertco_type=5)	Measurement error
final_obs_error	errstat	model_levels (vertco_type=5)	Combination of meaurement, representativity and forward model error
surface_pressure	modsurf	scalar	background surface pressure from model
lat	hdr	scalar	Average latitude of observations
lon	hdr	scalar	Average longitude of observations
stalt	hdr	scalar	Height of satellite above sea level

Table 2.4: Selected ODB variable definitions for Calipso observations.

3 4D-Var assimilation system developments

3.1 Description of the 4D-Var system

The current operational data assimilation system at ECMWF is an incremental version of 4D-Var (Courtier et al., 1994; Rabier et al., 2000). 4D-Var seeks an optimal balance between observations and the dynamics of the atmosphere by finding a model trajectory $\mathbf{x}(t)$ that is closer, in a least-square sense, to the observations available during a given time period $[t_0, t_n]$. The model trajectory $\mathbf{x}(t)$ is completely defined by the initial state $\mathbf{x}_0$ at time t_0.

The mis-match to a given set observations $\mathbf{y}^o$ and to an *a-priori* model state $\mathbf{x}^b$ called the background, which is usually provided by a short-range forecast, is measured by an objective (cost) function. The cost function effectively penalises both differences between the state of the model $\mathbf{x}_0$ and the background $\mathbf{x}_0^b$, and differences between the observations and model-equivalent observations. These differences are weighted according to the inverse of their expected error. An additional constraint to the cost function $\mathscr{J}^c$ is used in 4D-Var to control fast gravity waves using the digital filter approach developed by Gauthier and Thépaut (2001).

Using the incremental approach, 4D-Var can be approximated to the first order by finding the analysis increment $\delta\mathbf{x}_0$ at initial time t_0 which minimizes the following cost function $\mathscr{J}$:

$$\mathscr{J}(\delta\mathbf{x}_0) = \underbrace{\frac{1}{2}(\delta\mathbf{x}_0)^T\mathbf{B}^{-1}(\delta\mathbf{x}_0)}_{\mathscr{J}^b} + \underbrace{\frac{1}{2}\sum_{i=0}^{n}\left(H_i'\delta\mathbf{x}_i - \mathbf{d}_i\right)^T\mathbf{R}_i^{-1}\left(H_i'\delta\mathbf{x}_i - \mathbf{d}_i\right)}_{\mathscr{J}^o} + \mathscr{J}^c \tag{3.1}$$

where at any time t_i,

- $\delta\mathbf{x}_i = \mathbf{x}_i - \mathbf{x}_i^b$ is the analysis increment and represents the departure of the model state ($\mathbf{x}$) with respect to the background ($\mathbf{x}^b$) which consists of temperature, humidity, vorticity, divergence and surface pressure in the current 4D-Var system;
- H_i' is the linearized observation operator providing the model equivalent to the observations and it also includes the spatial interpolations to observation locations as well as the propagation of the initial state to each observation time using the forecast model;
- $\mathbf{d}_i = \mathbf{y}_i^o - H_i(\mathbf{x}^b{}_i)$ is the so-called innovation vector providing the departure of the model background equivalent from the observation ($\mathbf{y}_i^o$);
- $\mathbf{R}_i$ is the observation error covariance matrix (including measurement and representativeness errors);
- $\mathbf{B}$ is the background error covariance matrix of the state $\mathbf{x}^b$ and is based on a wavelet formulation (Fisher, 2004) to introduce regime-dependent error statistics.

A key advantage of the incremental approach is to reduce computational cost. Figure. 3.1 provides a schematic description of the incremental 4D-Var solution algorithm (IFS-Documentation, 2015). A high resolution version of the model is used for the computation of the model trajectory, and for calculating the departures between observations and model, whereas a lower-resolution model (its adjoint and tangent linear) are used for the iterative and relatively costly computation of analysis increments. The lower-resolution iterations (the inner-loops) can optionally be nested within a set of outer-loop iterations at full resolution. Apart from the resolution, the cost of the inner-loops will depend also upon the complexity of the inner-loop model, e.g. the use of simpler

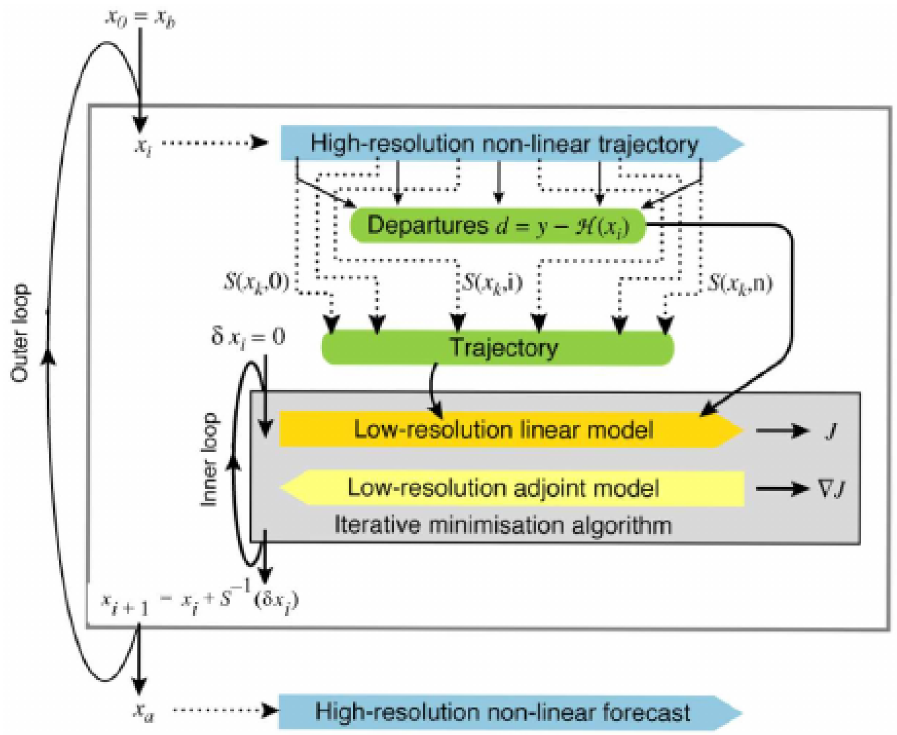

Figure 3.1: Schematic of the 4D-Var solution algorithm at ECMWF (IFS-Documentation, 2015). Outer loops are performed at high resolution using the full non-linear model. Inner iterations are performed at lower resolution (increased with each iteration) using the tangent-linear and adjoint versions of the forecast model, linearised around a 12-hour succession of model states ('the trajectory') obtained through interpolation from high resolution (S denotes the truncation operator, J the cost function and x the atmospheric state vector).

or more complete representations of the physical processes (Janisková and Lopez, 2013). The inner-loop resolution is increased with each iteration of the outer-loop using 'multi-resolution' extension to the incremental method (Veerse and Thépaut, 1998).

The current length of the 4D-Var assimilation window at ECMWF is 12 hours, running from 09 UTC to 21 UTC to produce the 12 UTC analysis and forecast products, and from 21 UTC to 09 UTC for the 00 UTC production. Observations for assimilation system are organized in time-slots (currently half-hourly). Timeslots are subdivisions of the 4D-Var window into which observations are grouped for processing. For interpolation from model to observation space, one timestep of the model is chosen to represent the model state in that timeslot. Timeslots are useful to isolate the observation grouping from the model time resolution which varies within 4D-Var.

3.2 Structure of 4D-Var tasks

3.2.1 Overview of required developments

The above brief description of the incremental 4D-Var system used at ECMWF indicates that the system is rather complex and it consists of several different steps, which require modifications and/or developments to account for the new observation types of cloud radar and lidar. The particular steps of 4D-Var solution are summarized here and in Fig. 3.2, with indications (*in italic* in text and by pink colour in figure) where modifications and developments were required for the new cloud radar and lidar observations:

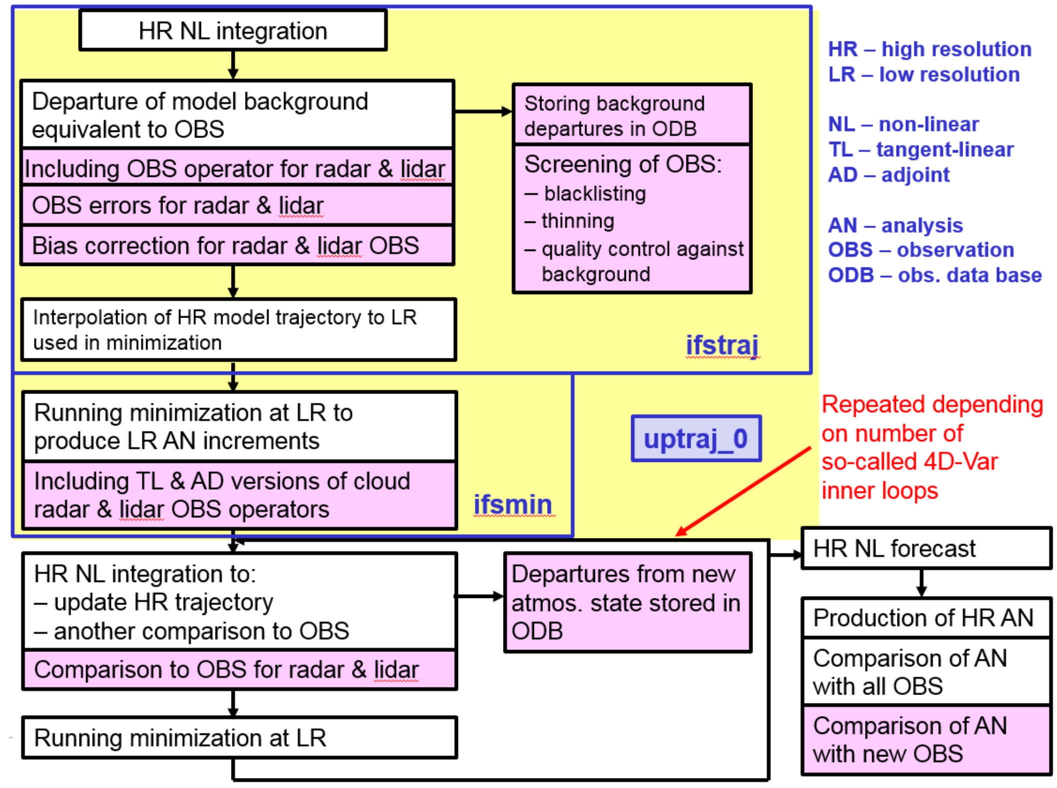

Figure 3.2: Simple flow diagram providing overview of the required developments and modifications (indicated by pink colour filled boxes) for including new cloud radar and lidar observations into the 4D-Var data assimilation system.

i. Comparison of the observations with the background is done at high resolution to compute the innovation vector providing the departure of the model background equivalent to the observation. This procedure involves using appropriate observation operators which need to be properly included in the 4D-Var system. *The direct, non-linear, versions of cloud radar and lidar observation operators had to be included for that.*

The background departures must then be stored in the ODB for later use in the minimization. This job step also performs screening (i.e. blacklisting, thinning and quality control against the background) of observations. The screening determines which observations will be passed to the main minimization. Very large volumes of data are present during the screening run only, for the purpose of data monitoring. *These steps had to be implemented for the new observations as well.*

The model trajectory is subsequently interpolated to the resolution of the next job step and written out.

ii. The first minimization at low resolution is used to produce preliminary low-resolution analysis increments, using simplified tangent-linear and adjoint physics, and tangent-linear and adjoint observation operators. *Therefore the tangent-linear and adjoint versions of cloud radar and lidar observations were included to the system.*

iii. First update of the high-resolution trajectory to take non-linear effects partly into account applies the analysis increments obtained in the first minimization and performs another forecast integration with comparison to observations. Departures from this new atmospheric state are stored in the ODB. *This procedure also involves the new observations.*

The analysis problem is then re-linearized around the updated model state which provides a new linearization state for the next minimization. The updated model trajectory is then interpolated to the resolution of the next job step and written out.

iv. The second main minimization is performed at increased resolution with a more complete representation of tangent-linear and adjoint physics. *The tangent-linear and adjoint versions of cloud radar and lidar observation operators are also used in this minimization.*

Steps (*iii*) and (*iv*) are then repeated again in the current operational data assimilation system.

v. The production of the high-resolution analysis is finally carried out by adding the low-resolution increments to the background (at initial time), and integrating to the analysis times. The comparison of the analysis with all observations (including those not used by the analysis, but also those used for diagnostic purposes) is performed. *This step had also to be implemented for the new observations.*

3.2.2 Overview of observation processing

The handling of observations is rather complex part of the IFS data assimilation system and it happens at many different stages. Observation processing involves (i) ingestion of raw data (as BUFR files), i.e. observations themselves, to the archiving of ODBs, (ii) running the observation operators in the data assimilation, (iii) screening, (iv) quality control, (v) assigning observation errors and (vi) applying bias corrections.

In the first, preparation stage, all the parameters needed to run the observation operator in the assimilation system are generated. This stage also includes format conversions (e.g. BUFR to ODB). It may further perform simple thinning (such as discarding some observations) or averaging observations to the model horizontal resolution and/or vertical matching of observations. This stage also sets up observation errors and other parameters as the settings for quality control.

Screening decisions whether or not to use a certain data are made throughout the observation processing chain. They includes blacklisting and other control decisions that can only be made when a model first guess is available, or in the presence of all observations. Parameters that control all these decisions can be pre-set during pre-processing and stored in the ODB for use during the assimilation.

Observation errors are also created during observation processing. Errors, such as instrument ones, are set at the pre-processing stage as they do not depend on any model information. Other types of observation errors are created when the observation operator is called for the first time, particularly if the observation error is situation-dependent. The observation errors are then stored in the ODB for later use by the assimilation system.

Some bias correction can be performed as part of pre-processing, but most is now handled inside the 4D-Var minimization using variational bias correction (VarBC). For new observations from cloud radar and lidar, pre-computed biases will be applied during the 4D-Var minimization.

For computational reasons it can be helpful to reduce observation numbers early in processing chain, before too much computational time has been spent on unwanted data. However, much of screening needs to be left until later stages because it is often necessary to have access to: (i) the first guess (FG) model state, (ii) the first guess departures (i.e. difference between observations and their model equivalents) and/or (iii) the results of screening decisions from other observations. In the data assimilation system, there is a distinction between 'independent' and 'dependent' screening decisions. The dependent decisions are such as checking for redundant observations and thinning, where a selection is made among all the observations that passed quality control and blacklisting.

In general, observation preparations may take place before the main 4D-Var calculation (in the so called pre-processing stage), or inside the main calculations. The pre-processing stages are identified as 'state-independent' since the only place that has access to the atmospheric state is the main part. In this part, the sequence of jobs starts with the first (high resolution) trajectory run. During this run the model counterparts for all the observations are calculated through the non-linear observation operators, and the observation minus model differences (the departures) are calculated. At this stage, 'state-dependent' observation processing is performed.

Technically, the final result of the observation screening is a pair of ODBs. The original extended observation data base, ECMA (Extended CMA, Central Memory Array) ODB, contains observations complemented by the background departures, together with quality control information for most of the observations. This ECMA remains on disc for later use in feedback creation at the end of the whole 4D-Var calculation. The compressed ODB, the CCMA (Compressed CMA), contains a subset of the original observations, and is passed for the subsequent minimization job. It contains only those observations, so called active observations, that are to be used in the minimization.

3.2.3 Structure of 4D-Var assimilation tasks

Figure 3.3 provides an overview of task groups used by the 4D-Var assimilation system. Call tree of tasks there and in the following subsections comes from ecFlowview - a graphical user interface used at ECMWF to display graphically the status of tasks within experiments that are in-progress. The running of a large number of programs, as required for data assimilation, is enabled by ECMWF's work-flow manager (ecFlow).

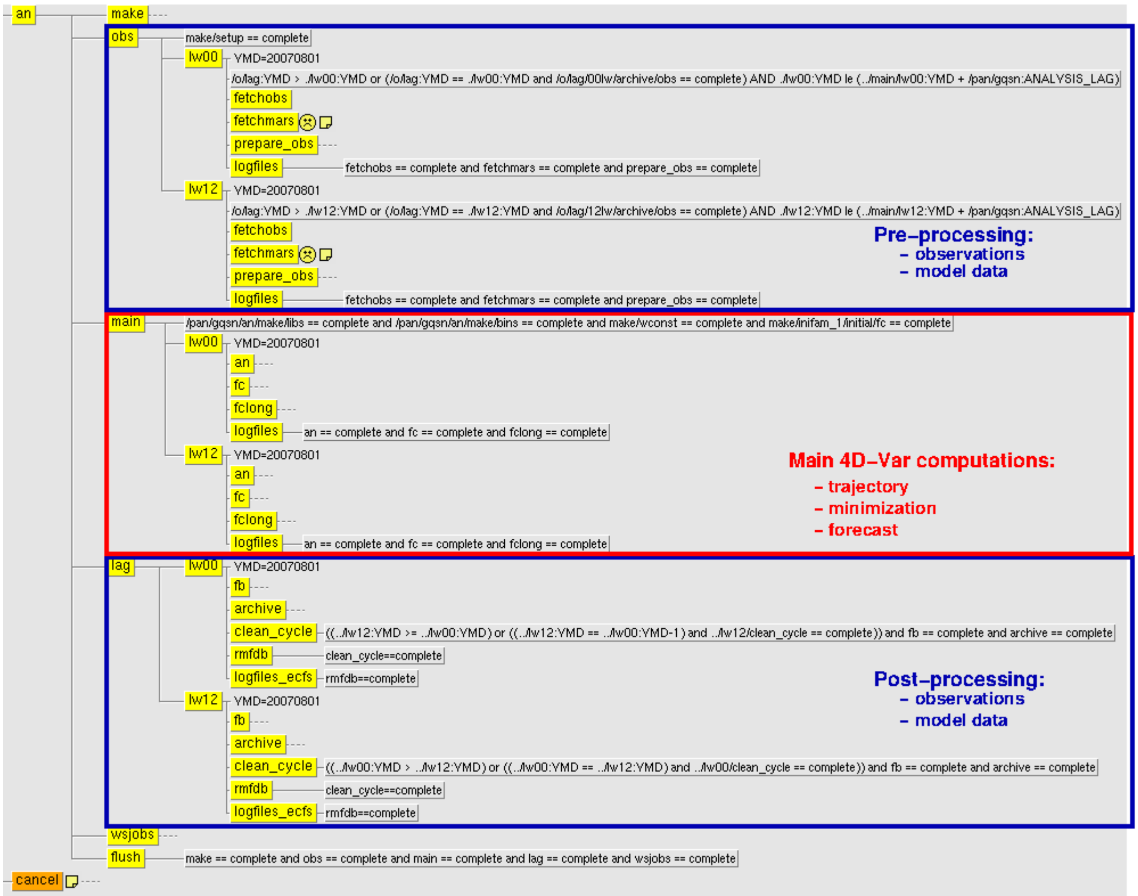

Figure 3.3: Structure of the ECWMF 4D-Var assimilation computation split to three task groups: **obs** *group - tasks for observation pre-processing;* **main** *group - tasks performing main 4D-Var computations such as high resolution trajectory run, minimization and forecast run;* **lag** *group - tasks for post-processing and archiving observations and model data. Tasks groups* **lw00** *and* **lw12** *are for the 00 UTC and 12 UTC analysis production, respectively.*

The current length of the 4D-Var assimilation window at ECMWF is 12 hours, running from 21 UTC to 09 UTC to produce the 00 UTC analysis and forecast products (**lw00** task groups in Fig. 3.3), and from 09 UTC to 21 UTC for the 12 UTC production (**lw12** task groups).

The first group of tasks under **obs** is devoted to different pre-processing computations (as described in subsection 3.2.4) for the subsequent assimilation calculations. At this stage, observations and model data are fetched from the ECMWF archive. Observations are then pre-processed and converted from BUFR to ODB.

The main 4D-Var computations are performed under the second group of tasks, **main**, where high resolution trajectory, minimization and forecast runs are done. More detail explanation is provided in subsection 3.2.5.

In the last group, **lag**, post-processing and archiving of observations and model data are done (subsection 3.2.6).

3.2.4 Pre-processing observations

The first phase for preparation of observations to be used in the data assimilation system is the processing required before a normalised first guess departure can be computed. It combines the 'pre-processing' tasks and the preparations that takes place within the main run during the initial screening run.

All satellite data goes through the long-established path of BUFR preprocessing (the **prepare_obs** tasks, Fig. 3.4) and BUFR to ODB conversion (the **bufr2odb**, Fig. 3.5 tasks). In this area, data undergo some rudimentary quality controls, such as a check for the observation format and position, and for the climatological limits.

The new cloud radar and lidar observations undergo a pre-screening process (tasks **preclrad** and **precllid**, respectively in Fig.3.4). This is used to reduce the data volume (thus the computational cost) of the main screening and also to reject observations that fail to contain crucial information. The observation screening begins with a preliminary check of the completeness of the reports, such as observed value, observation error and vertical coordinate of observations. At this stage, the so-called superobbing is also performed. Observations are first placed onto some grid with resolution depending on the level of smoothing desired. For initial experiments this will be the model resolution of TCo639, but in theory any grid can be chosen. The resulting average forms each superob. Vertical matching with the model vertical resolution is also done at this stage.

Once the superobs have been created, there are additional pre-screening tasks, which include threshold checks on: the observed cloud fraction, the number of observations forming the superob, the standard deviation of the observations forming the superob and the mean observation value itself. Additionally, for radar, there is a check for multiple scattering by applying a threshold check on the integrated radar reflectivity. After all the checks are complete, the superobs are written to a temporary BUFR file.

The pre-processed data for cloud radar and lidar observations, still in BUFR format, are converted to ODB in the **b2o_clrad** and **b2ocllid** tasks, respectively (Fig. 3.5). The temporary BUFR files have been constructed in a way such that the mapping of superobs contained within the temporary BUFR file to the ODB variables results in a one-to-one mapping.

3.2.5 Main 4D-Var computations

The ECMWF 4D-Var data assimilation system uses an incremental approach in the minimization scheme. Observations are read into 4D-Var from the Observation Data Base (ODB) in the **main** task group (Fig. 3.6). The sequence of jobs starts with the first (high resolution) trajectory run, **ifstraj**. During this run the model counterparts for all observations are calculated through the non-linear observation operators, and the observation minus model differences (departures) are calculated and stored in the ODB. These departures are an important input to the data selection procedures as many quality control decisions depend on the magnitude of the departure. Once all background departures are available, the screening is performed. The purpose of the observation screening is to select the best quality observations, to detect duplicates and reduce data redundancy through

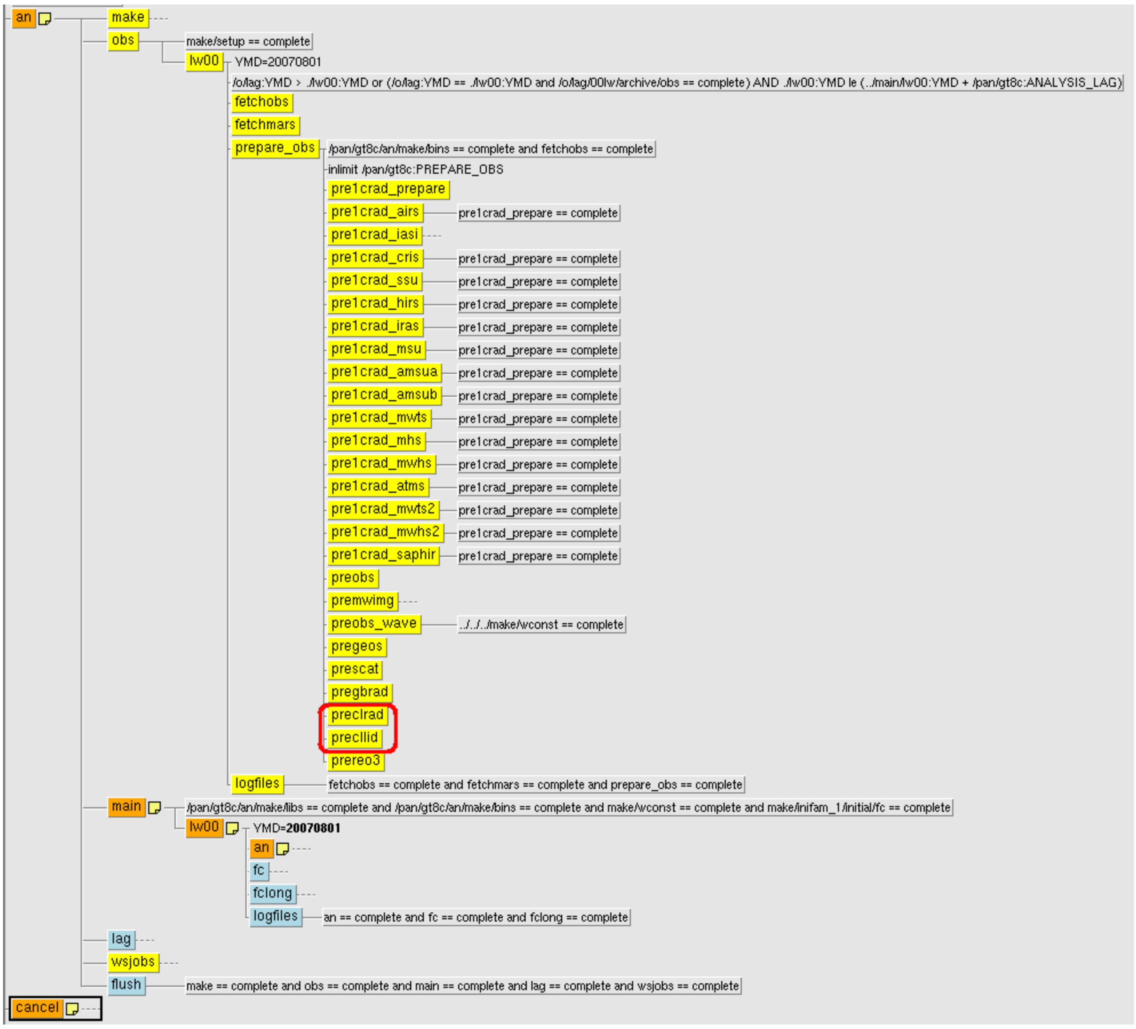

Figure 3.4: Observation pre-processing tasks in 4D-Var cycling of the IFS - tasks **preclrad** *and* **precllid** *(marked in red) perform some pre-screening for cloud radar and lidar observations, respectively.*

thinning. Any observations with too large departures (i.e. exceeding the predefined limits) are rejected at this stage.

Next, the observations are scanned through for blacklisting during which the selection of variables for assimilation is specified. This controls which observation types, variables, vertical ranges etc. will be selected for the assimilation. Some more complicated decisions are also performed through the data selection file; for instance, an orographic rejection limit is applied in the case of the observation being too deep inside the model orography. The first type of decisions made to use or not use a particular observations is an a priori type decision which takes no account of the actual value of the observations. The second type of test is based on the observed values (or observation departure from the background). Blacklist-rejected data are subsequently excluded and will not be present in the minimisation job steps.

Subsequently, the dependent observation screening decisions are performed. In this group, there are such decisions as the vertical-consistency check for multilevel reports, removal of duplicated reports or horizontal thinning when necessary.

After performing screening, the original ODB is complemented by the background departures together with quality information for observations. The compressed ODB, the CCMA, contains only those observations that

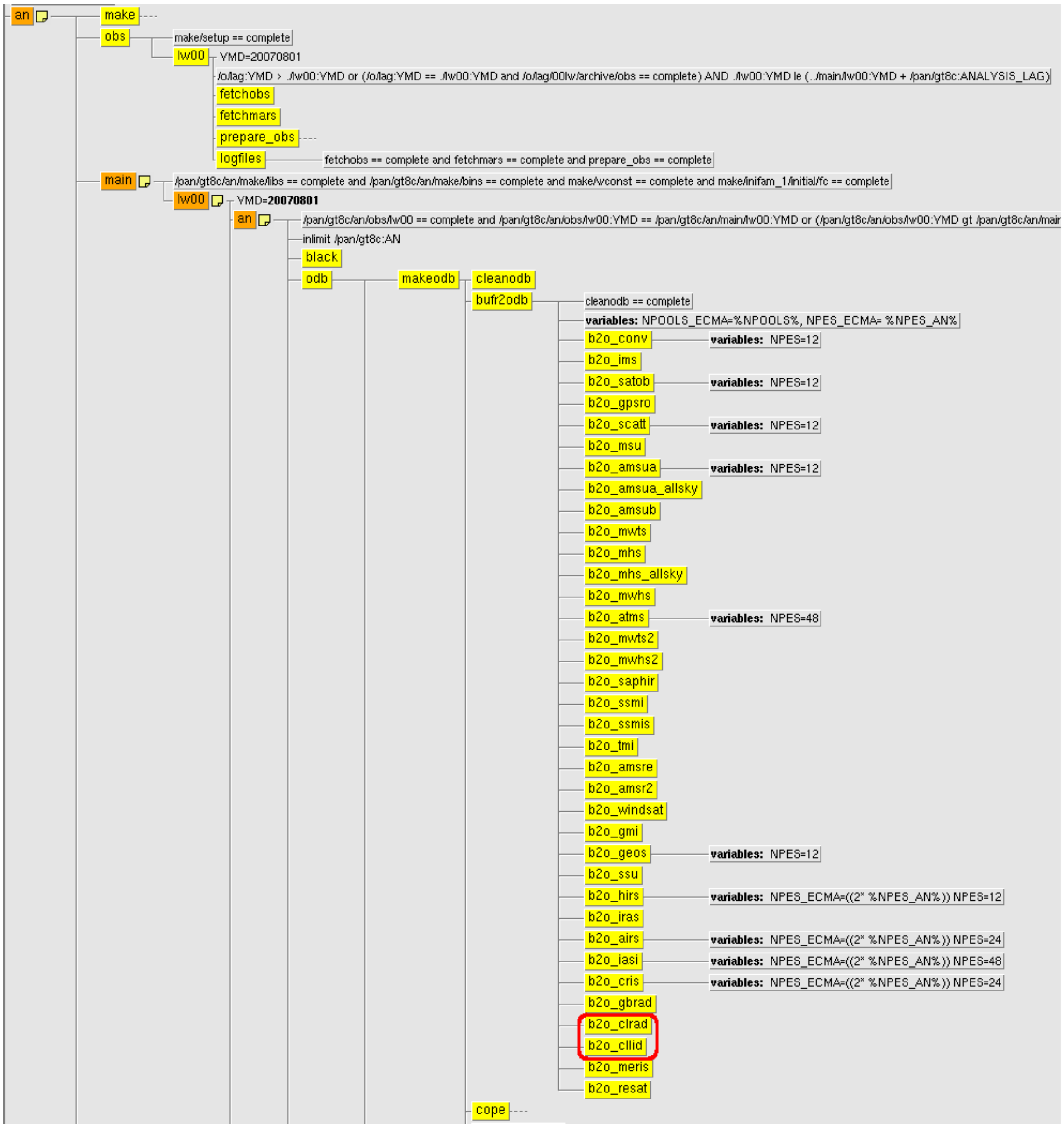

Figure 3.5: Conversion of BUFR to ODB in the pre-processing stage of 4D-Var cycling tasks - in **b2o_clrad** *and* **b2o_cllid** *(marked in red), the pre-processed data are converted to ODB for cloud radar and lidar observations, respectively.*

passed screening and that are to be used in the minimization (**ifsmin** tasks).

The following computations are performed under the **4dvar** task group (Fig. 3.6):

- The so called 'screening run' described above is done in the first trajectory run (**ifstraj** in **uptraj_0**) which includes a model integration, comparison to observations, and observation screening (quality control and data selection).
- The first minimization (**ifsmin** in **uptraj_0**) typically runs at low resolution. This step includes estimation of analysis and forecast error variances, and calculation of Hessian eigenvectors for pre-conditioning of subsequent minimisation(s).
- The first trajectory update (**ifstraj** in **uptraj_1**) applies the analysis increments obtained in the first minimisations and performs another forecast integration with comparison to observations. This provides a

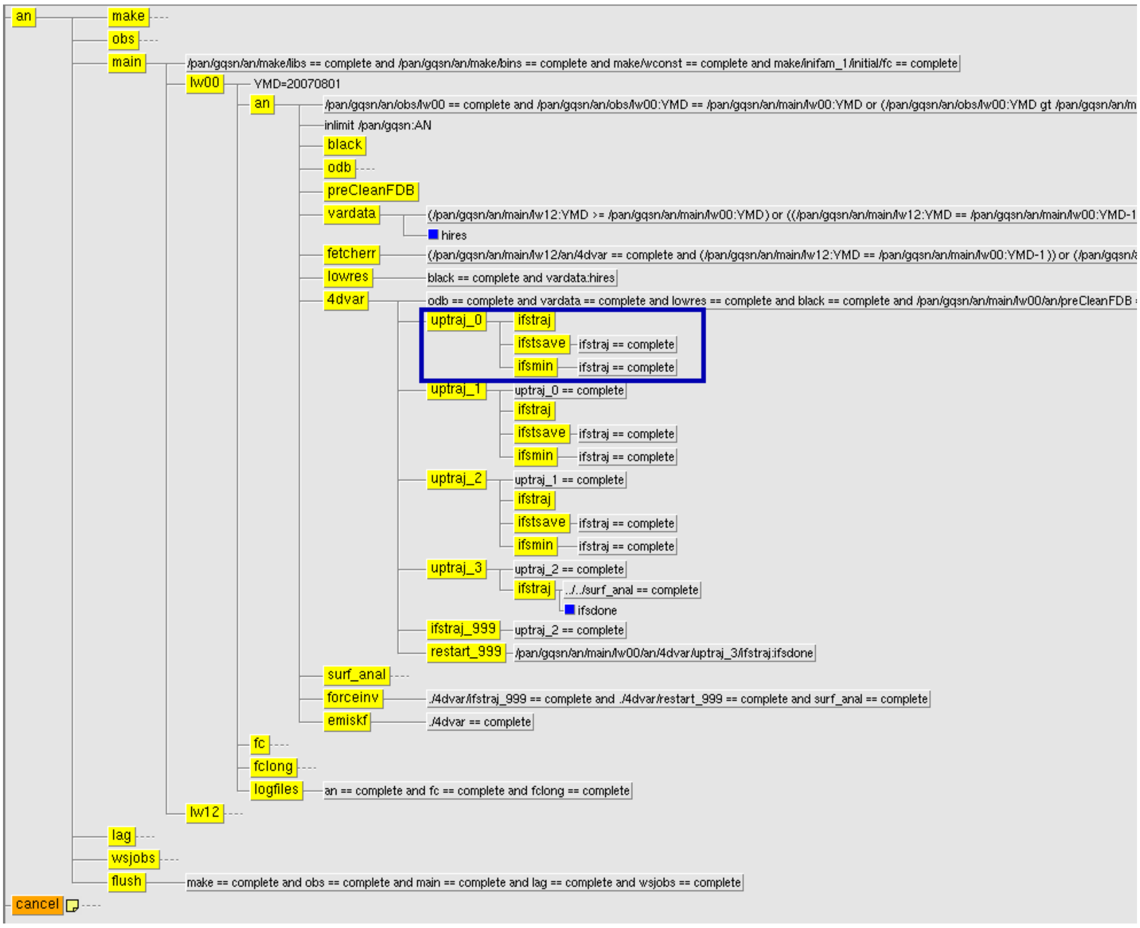

Figure 3.6: Tasks of 4D-Var cycling as used by the IFS at ECMWF - main 4D-Var computations. The group of tasks in blue rectangle corresponds to yellow highlighted part of the flow diagram in Fig. 3.2.

new linearisation state for the next minimisation.

- The second (**ifsmin** in **uptraj_1** and subsequent in **uptraj_2**) minimization uses higher resolution increments.
- The final trajectory runs (**ifstraj** in **uptraj_3**) carries out verification screening, which involves a comparison between the observations and the final analysis. In the final trajectory the final analysis is formed by adding the low-resolution increment to the background (at initial time), and integrating to the analysis times.

At the end of **an** task group, departures of observations from the background and the analysis, and quality information are written to the ODB.

3.2.6 Post-processing of observations

After completing all the tasks in the **main** 4D-Var calculations, when the high-resolution analysis is produced, post-processing of observations is performed. This is done in two stages (Fig. 3.7). First, an ODB file is created from ECMA (Extended CMA) containing information about all observations, not only the active ones. This is done in the task group **archive_prepare**, where the new tasks **convert_clrad** and **convert_cllid** had to be added for cloud radar and lidar observations. ODB files created by **convert_**$*obsgroup* tasks are then archived in

MARS (Meteorological Archiving and Retrieval System of ECMWF). This is done by the tasks **archive_clrad** and **archive_cllid** for cloud radar and lidar observations, respectively.

Figure 3.7: Post-processing tasks in 4D-Var cycling: (left) creation of ODB database with direct column access under **archive_prepare** *task group - tasks* **convert_clrad** *and* **convertcllid** *(marked in red) used for the new cloud radar and lidar observations; (right) archiving ODB database under* **archive_odb** *task group - tasks* **archive_clrad** *and* **archive_cllid** *(marked in red) for the new observations.*

3.3 Observation operators

The master routine dealing with all different types of observations in the IFS is called **hop** (Fig. 3.8). Computation of the model equivalent to observations is done at this level. Currently it is assumed that each observation equivalent can be computed from a single vertical profile of model data. After obtaining these equivalents, their departures to observations and the observation cost function are computed. Bias correction is also carried out at this point by subtracting the bias estimate from the calculated departure. Finally the departure is divided by the observation error to form the normalized departure.

The routine for computation of observation cost function also stores the resulting effective departure in ODB, for reuse as the input to the adjoint. The effective departure is the normalized departure after the effects of (vertical) observation error correlation and quality control have been taken into account.

Under the **hop** routine, the subroutine for all computations related to cloud radar and lidar observations, **obsop_clradlid**, is called. At this point, having access to the model state information, computation of situation dependent observation errors, screening for high observation errors as well as departures and assigning bias correction are performed. Since some of those requires knowledge of the model equivalent to the observations, the above observation processing is only done once such equivalent is available after calling **clradlid_wrapper** routine. In this routine, firstly observational data are obtained from ODB using **clradl_get** routines (suffix **_tl** and **_ad** are used for the tangent-linear, TL, and adjoint, AD, versions of those routines). Observation screening depending on model situation is done in routine **clradlid_screen**. The model equivalents to cloud radar and lidar observations are obtained by performing calculations in the routines **clradlid_obsop** (suffix **_tl** and **_ad** for TL and AD versions, respectively). More detailed structure of the subroutines for the observation operators is below (Fig. 3.9 and 3.10). At the end of **clradlid_wrapper** routine, observation related information is stored to ODB in the **clradlid_put** and **clradlid_put_tl** routines.

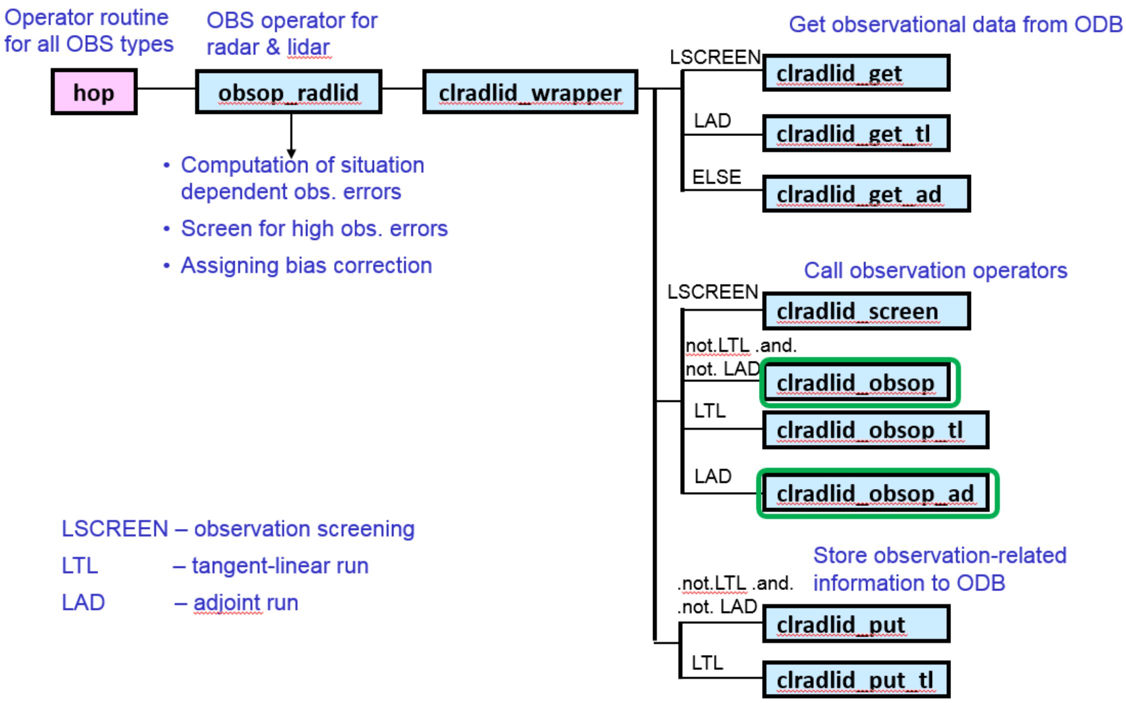

Figure 3.8: Schematic diagram describing inclusion of the observation operators for computing model equivalent to cloud radar and lidar observations, as well as getting/storing these observational data in ODB.

Figures 3.9 and 3.10 provide the detailed structure of the observation operator subroutines for the direct non-linear and adjoint integrations, respectively. There are two versions of the direct observation operator. The first one is **clradlid_monit** which is complex and computationally demanding and therefore will only be used for monitoring and model validation. The second, simpler version of observation operator for cloud radar and lidar observations, **clradlid_assim** is computationally efficient and will be used for data assimilation.

A brief description of the contents of all the observation operator subroutines for cloud radar and lidar observations follows:

clradlid_monit and **clradlid_assim** are drivers to all routines for computation of overlap (cloud and precipitation related quantities), cloud radar reflectivity and lidar backscatter. For the complex **clradlid_monit** scheme

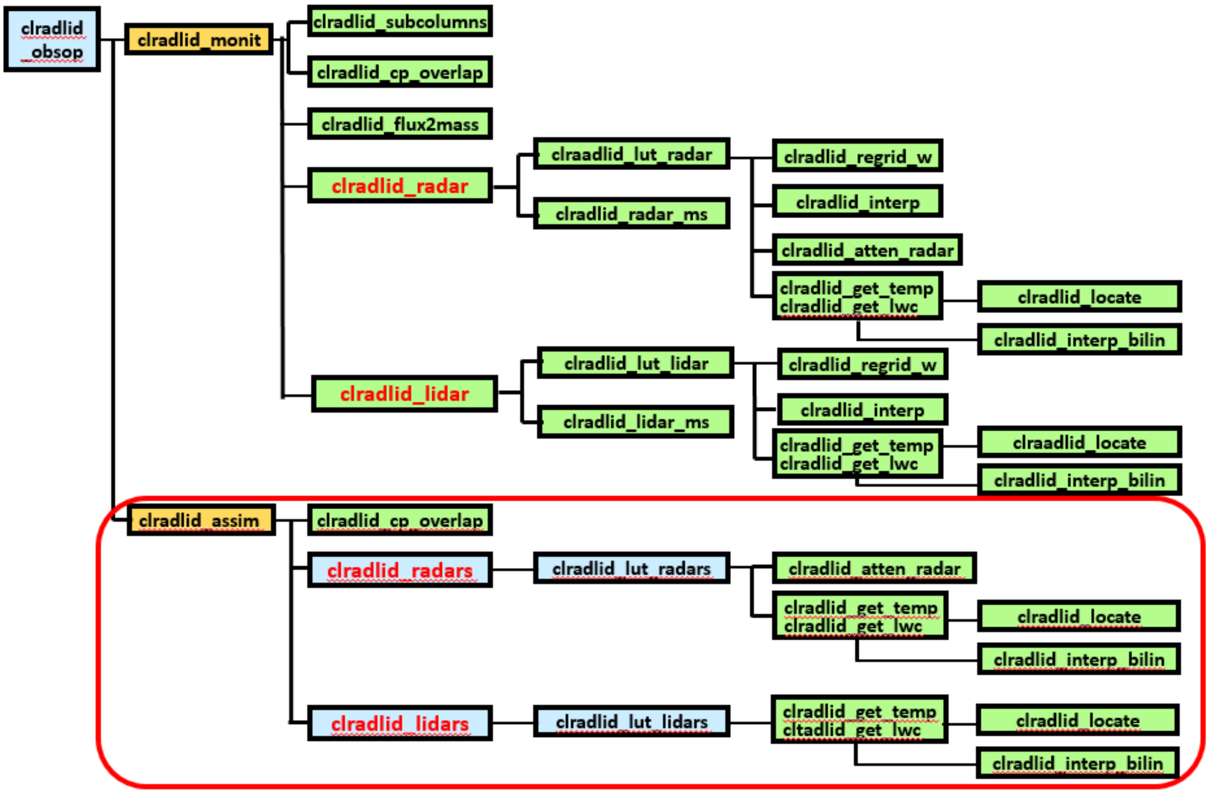

Figure 3.9: Call tree of routines for direct observation operators used for cloud radar and lidar observations. Two groups of operators are called under **clradlid_obsop***:* **clradlid_monit** *- complex version of operators, computationally demanding, to be used just for monitoring and model validation;* **clradlid_assim** *- simple version, computationally efficient, to be used for data assimilation.*

using multicolumn approach, **clradlid_subcolumns** routine is also called.

clradlid_overlap is used for the computation of overlap for cloud and precipitation related quantities.

clradlid_radars is used to compute of radar reflectivity and extinction from the input values (temperature, specific humidity, rain, snow, cloud, ice) using a look-up table of radar reflectivity and extinction values. In this routine, cloudy sky attenuation is also given by the sum of clear sky attenuation and the hydrometeor extinction.

clradlid_lut_radars serves for localization of temperature and hydrometeor intervals and searching the reflectivity and extinction tables. It then calls routine for interpolation according to the temperature and the water content. The scaling of extinction and reflectivity to obtain in cloud values is also performed there.

clradlid_lut_lidars provides the same as **clradlid_lut_radars**, but for the lidar backscatter and extinction.

clradlid_get_temp performs searching to obtain specified model temperature values from the lookup tables.

clradlid_get_lwc performs searching to obtain specified model water content values from the lookup tables.

clradlid_locate is used for locating position in lookup table based on prescribed temperature or water content intervals and model values.

clradlid_interp_bilin performs bilinear interpolation using the model temperature and in-cloud water content to obtain unattenuated radar reflectivity, lidar backscatter or extinction for each hydrometeor type from the pre-computed lookup-tables.

clradlid_atten_radar computes attenuation of the signal along the radar beam. Clear sky attenuation due to gases is calculated using the model of Liebe (1985); Liebe et al. (1992).

clradlid_lidars computes lidar backscatter and extinction from the input values (temperature, specific humidity, rain, snow, cloud, ice) using a look-up table of lidar backscatter and extinction values.

Figure 3.10 provides a call tree of adjoint version of the routines for observation operators. Adjoint routines have the same name as nonlinear, but suffix **_ad** is added.

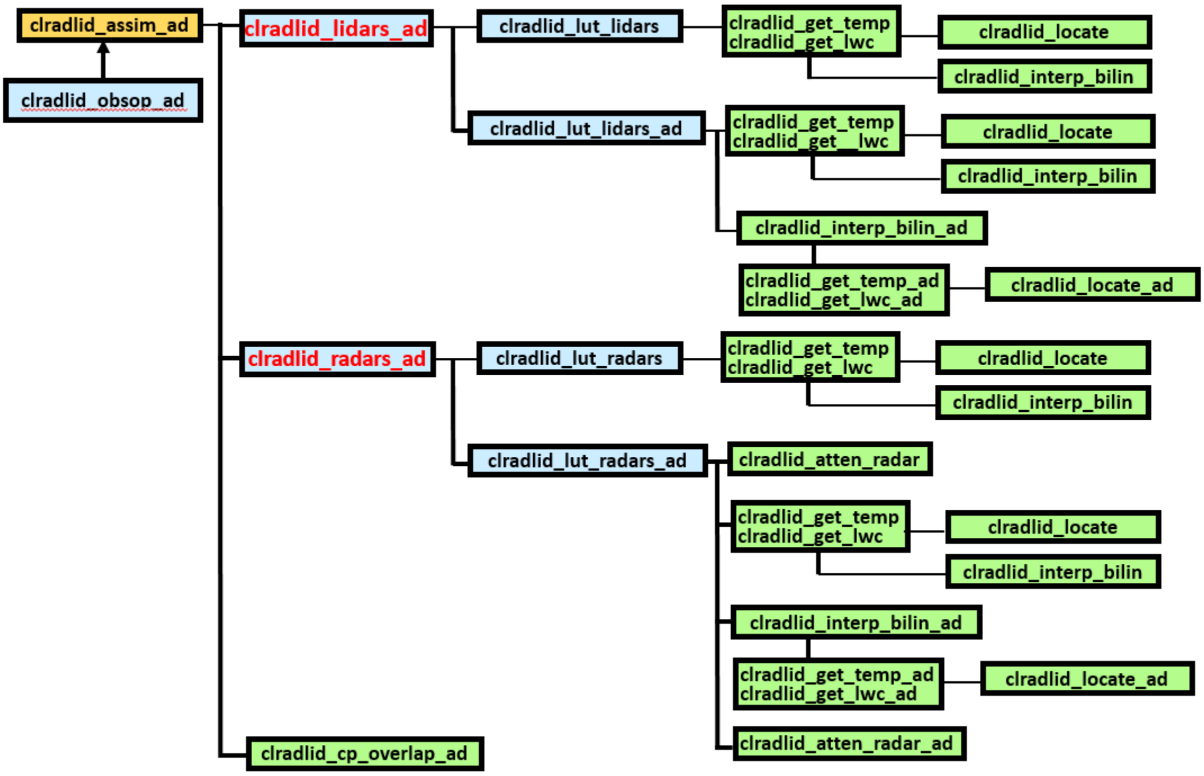

Figure 3.10: Call tree of routines for adjoint versions of observation operators to be used in 4D-Var for assimilation of cloud radar and lidar observations.

4 Testing

In this section we detail the testing of the all the technical developments outlined in Sections 2 and 3. Technical testing is required to ensure that the new observations are recognized correctly and are fed through to the ODB properly. The observation equivalents generated by the observation operators passing to the ODB must also be verified. Once all the data stored in the ODB is known to be correct, a basic test of the 4D-Var minimization can be made, where the convergence of the system with the new observations and operators can be checked. All testing is performed with a single transect of CALIPSO L2 C-Pro data (Winker et al., 2009) and CloudSat L2B data (Marchand et al., 2008), see Table 4.1.

4.1 BUFR pre-processing

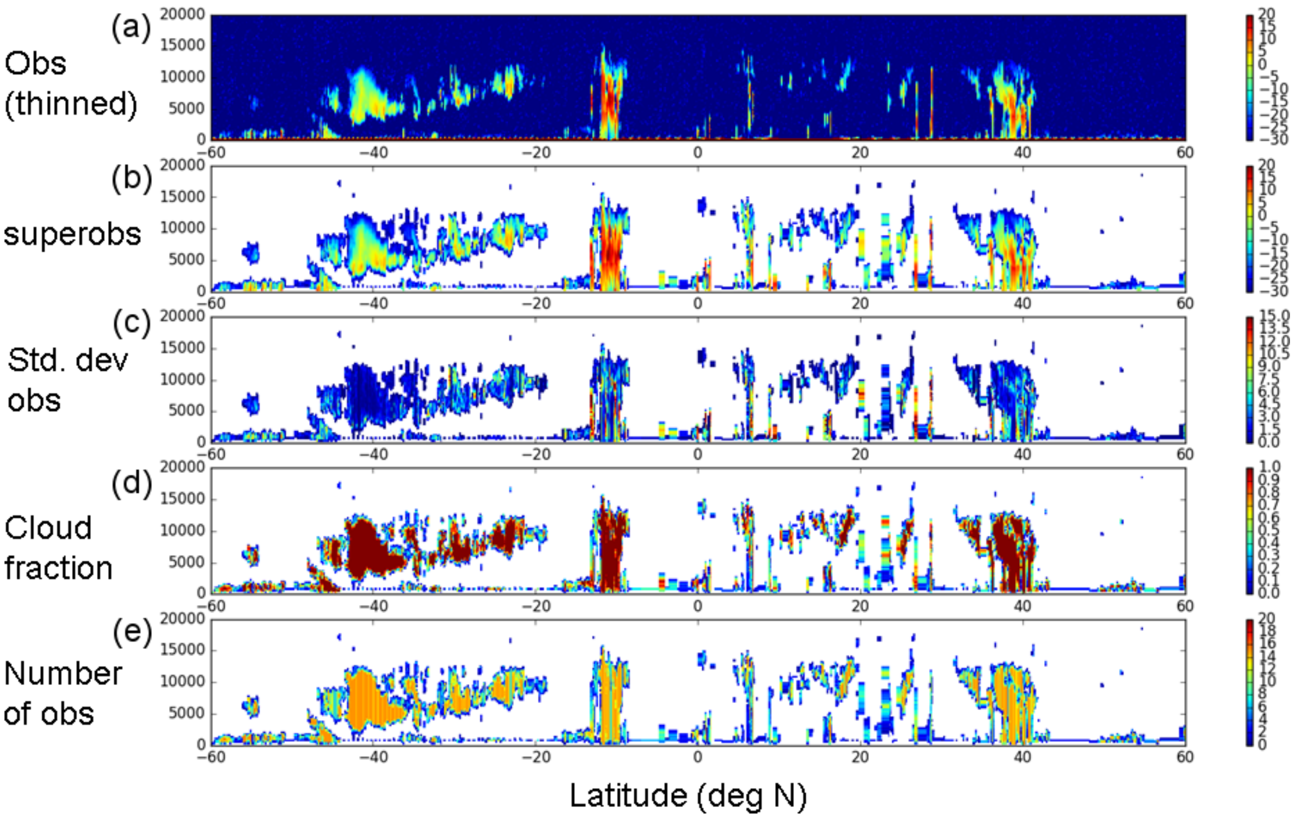

Figure 4.1: Direct output of the pre-processing of the CloudSat test data. Panels show: (a) native resolution radar reflectivity from input BUFR test file, (b) Superobbed radar reflectivity to native model resolution (TCo639), (c) standard deviation of radar reflectivity within each superob, (d) cloud fraction of superob (number of cloudy observations divided by total number of observations), (e) total number of observations within each superob.

As explained in Sec. 3.2.4, an initial processing of the data is performed to prepare the raw BUFR input for conversion to ODB. Figure 4.1a shows the raw CloudSat radar reflectivity converted from NetCDF (Network Common Data Form) to BUFR following the procedures defined in Sec. 2.1. Due to the high volume of data, only every 10th profile is plotted. The data is then matched to the TCo639 model grid with 137 vertical levels and then averaged to create the superobs shown in Fig. 4.1b. Due to fine resolution of the TCo639 data (approximately 18 km grid spacing), only smaller scale features are smoothed from the raw BUFR data. Averaging the data is preferable to simple thinning as it reduces noise, therefore providing a more accurate estimate of the grid-box reflectivity.

Superobbing to TCo639 leads to typically 15 native radar volumes or observations per superob (see Fig. 4.1e). The standard deviation of radar observations per superob is shown in Fig. 4.1c. This is necessary to include in the BUFR file (and hence ODB) as it is used to estimate the representativity error (a component of the total observation error). Finally cloud fraction (number of cloudy observations divided by the total number of observations) is shown in Fig. 4.1d, which is used for screening and also in calculating the observation error.

Figure 4.2 shows the output of testing for the CALIPSO test BUFR data. The transect is from a different time period to the CloudSat test data and shows a region of deep convection between 0 and 20 °N, stratiform precipitation between -60 and -40 °N and ice cloud between 40 and 60 °N. Superobbing the attenuated backscatter to TCo639 model resolution shown in Fig. 4.2b behaves as expected, with small scale features smoothed while larger scale features are retained. The number of observations included in each superob (Fig. 4.2e) is greater than for CloudSat because of CALIPSO's finer native vertical resolution.

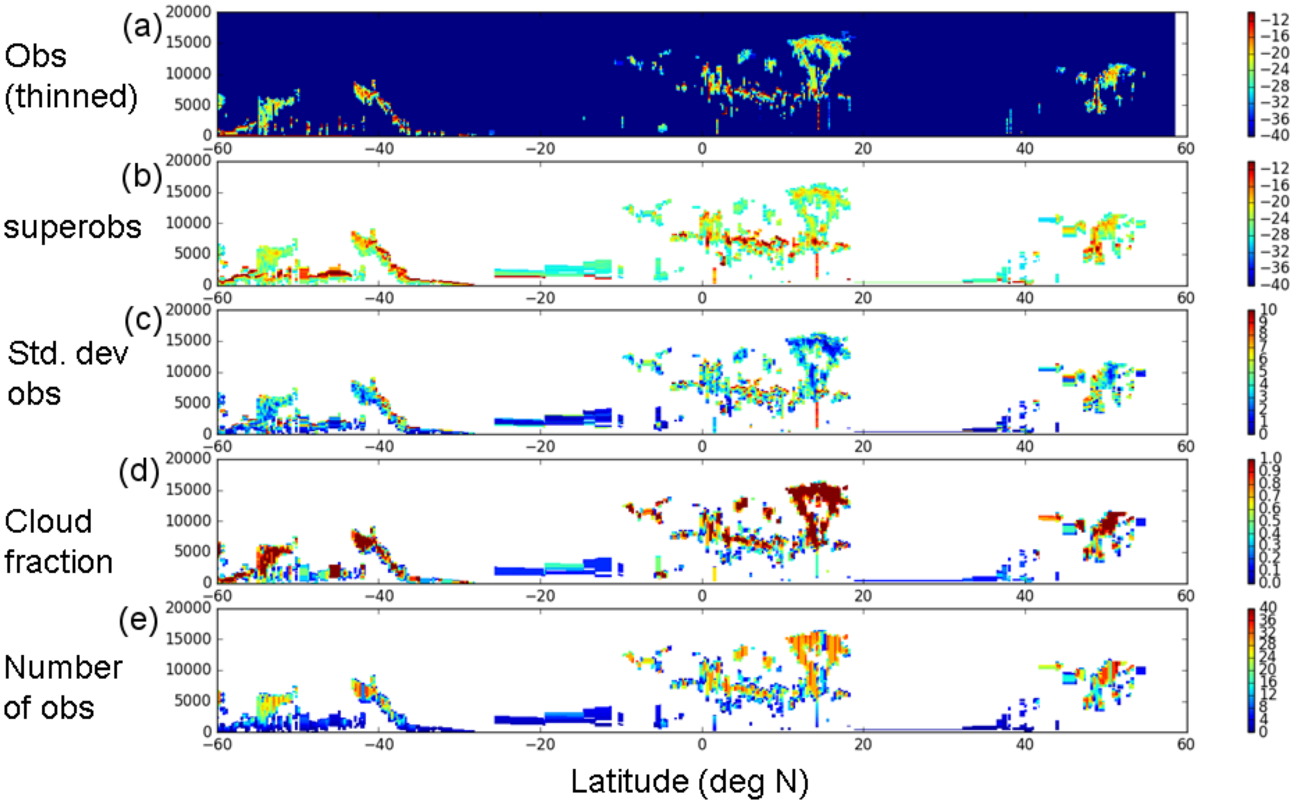

Figure 4.2: Same as Fig. 4.1, but for CALIPSO attenuated backscatter ($dB\beta$).

Instrument	**Filename**
CloudSat	2007213002658_06694_CS_2B-GEOPROF_GRANULE_P_R04_E02.hdf
CALIPSO	CAL_LID_L2_05kmCPro-Prov-V3-01.2007-08-01T00-10-17ZN.nc

Table 4.1: List of filenames for cloud radar and lidar observations used in testing.

4.2 BUFR to ODB

All information relating to observations needed by the assimilation system is stored in the ODB, so its correct functioning is crucial and must also be tested. Figure 4.3 provides a visual check on the contents of the CloudSat related entries to the ODB. The observed radar reflectivity stored in 'obsvalue' is plotted in Fig. 4.3a and is identical to the superobs shown in Fig. 4.1b. Also stored in the ODB is the bias correction, total observation error, first guess departures and screening flag. The areas in black in Fig. 4.3e show the superobs that passed screening, while white areas indicate areas that will not be assimilated.

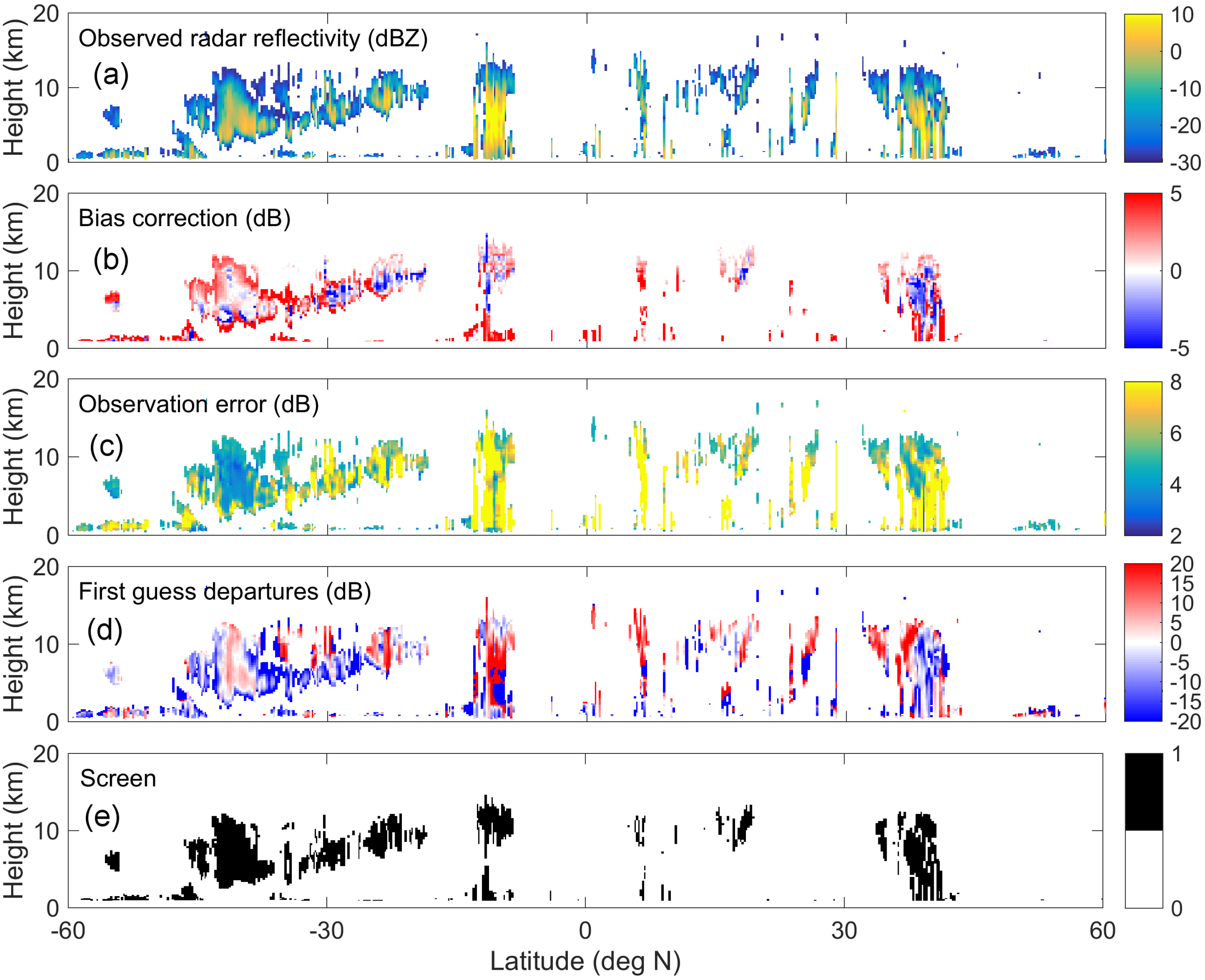

Figure 4.3: Example of observations stored in the ODB related to the CloudSat test data. Panels show: (a) Superobbed radar reflectivity to native model resolution (TCo639; dBZ), (b) bias correction to be applied to model equivalent (dB), (c) total observation error of each superob (dB), (d) observation minus background first guess departures (dB), (e) screening value, where '1' is pass and '0' is reject.

Figure 4.4 shows the equivalent testing of the ODB interface for the CALIPSO test dataset. Again, Fig. 4.4a shows the superobs that were generated in the pre-processing step (Fig. 4.2b). Screening tends to be stricter for the CALIPSO data (Fig. 4.4e) than for the CloudSat data, with a significant amount of observations below 3 km removed. These black-listed observations often correspond to observations with large first guess departures (Fig. 4.2d) or large observation errors (Fig. 4.2c), so are likely to have a negative impact on assimilation if they were included.

The integration of the observation operators with the ODB, specifically the inclusion of the model equivalents, has been tested and is evident in the first guess departures plotted in Fig. 4.3d and Fig. 4.4d. The validity of the tangent linear and adjoint codes has also been tested and is reported in Fielding and Janisková (2017).

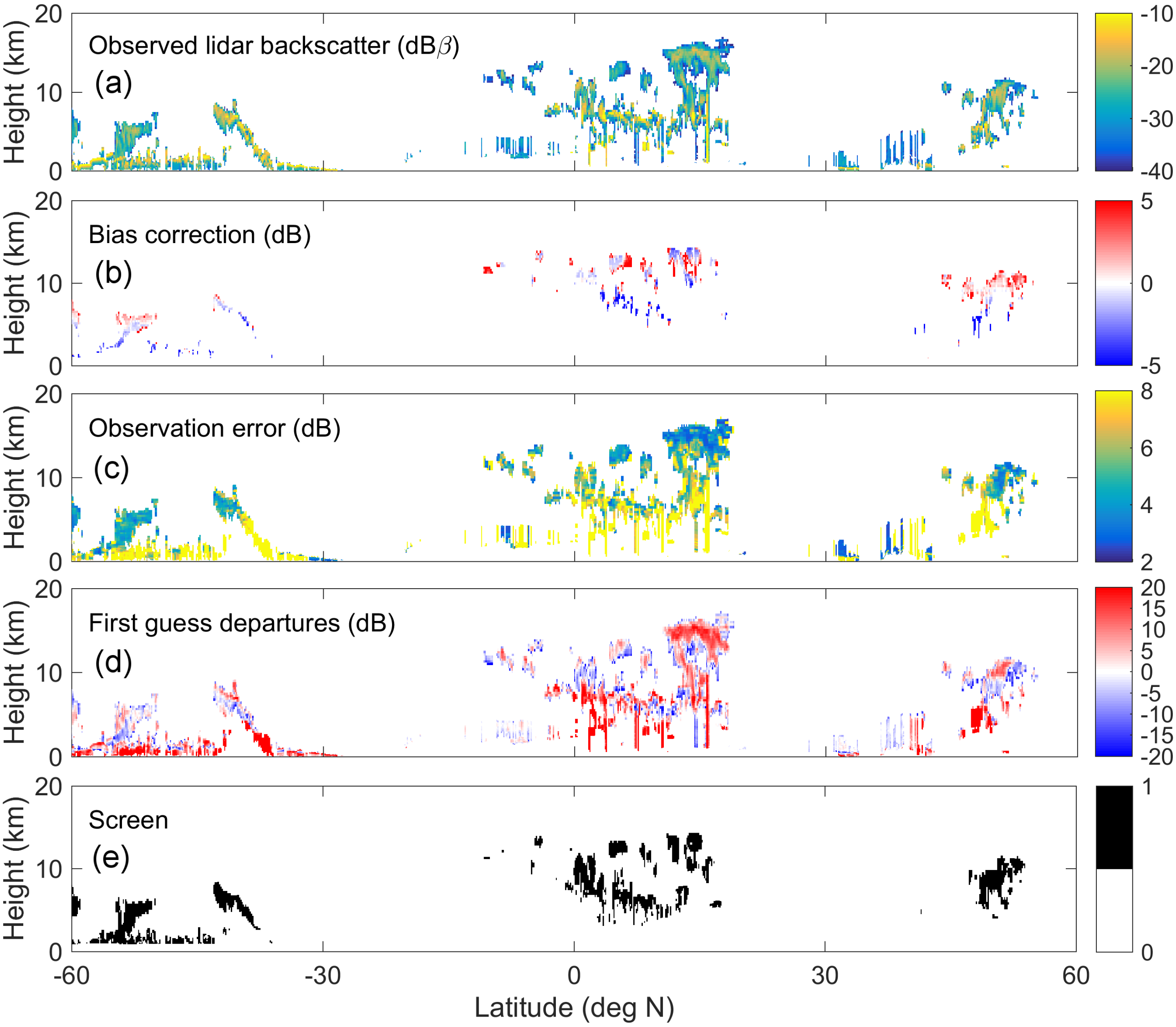

Figure 4.4: Same as Fig. 4.3, but for CALIPSO attenuated backscatter (dBβ).

5 Summary

In this report, the extensive developments to the operational data assimilation at ECMWF in preparation for the ingestion of cloud radar and lidar observations have been documented. The report contains three sections. Firstly, the off-line observation handling of CloudSat and CALIPSO into BUFR format are detailed, along with the necessary ODB developments. The second part of the report details all the coding developments that have been made to the assimilation system, the inclusion of the observations operators and the interface to the ODB. Finally the report concludes with an initial testing of the system using transects of CloudSat and CALIPSO data. All the above developments allow for the start of feasibility studies of direct 4D-Var assimilation of cloud radar and lidar observations using CloudSat and CALIPSO.

Acknowledgments

The authors would like to thank Philippe Lopez and Alan Geer for their help in deciphering existing IFS code. Thanks also to Enrico Fucile and Roberto Ribas for their help in preparing the pre-processing routines related to BUFR. Finally, the authors would like to thank Robin Hogan for invaluable discussions.

The NASA CloudSat Project is kindly acknowledged for providing the CloudSat data. The authors are also grateful to the NASA Langley Research Center - Atmospheric Science Data Center for making the CALIPSO data available.

List of Acronyms

4D-Var	Four-Dimensional Variational Assimilation
AN	Analysis
ATLID	ATmospheric LIDar
BUFR	Binary Universal Form for the Representation of meteorological data
C-PRO	Cloud profiling radar PROcessing
CALIOP	Cloud-Aerosol Lidar with Orthogonal Polarization
CALIPSO	Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation
CloudSat	NASA's cloud radar mission
CCMA	Compressed CMA (used for active observations after screening)
CMA	Central Memory Array (used for observations at ECMWF)
CPR	Cloud Profiling Radar
EarthCARE	Earth, Clouds, Aerosols and Radiation Explorer
ECFS	ECMWF's File Storage system
ecFlow	ECMWF's work-flow manager enabling to run large number of programs
ecFlowview	graphical user interface to display the status of experiment tasks
ECMA	Extended CMA (used for all observations before screening)
ECMWF	European Centre for Medium Range Weather Forecasts
ESA	European Space Agency
ESTEC	European Space Research and Technology Centre
FG	First Guess
HR	High Resolution of the model
IFS	Integrated Forecasting System of ECMWF
LR	Low Resolution of the model
MARS	Meteorological Archiving and Retrieval System
NASA	National Aeronautics and Space Administration
NetCDF	Network Common Data Form
NL	Non-Linear
NWP	Numerical Weather Prediction
OBS	OBServations
ODB	Observation Data Base
SQL	Structured Query Language
STSE	Support-to-Science-Element
TCo639	Model cubic octahedral grid with spectral truncation T639
TL	Tangent Linear
UTC	Universal Time Coordinated
VarBC	Variational Bias Correction
WMO	World Meteorological Organization
WP	Work Package

References

Courtier, P., J.-N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach, *Q. J. R. Meteorol. Soc.*, **120**, 1367–1387.

Fielding, M. D. and M. Janisková, 2017: Observation quality monitoring and pre-processing, WP-2000 report for the project Operational Assimilation of Space-borne Radar and Lidar Cloud Profile Observations for Numerical Weather Prediction, 4000116891/16/NL/LvH, pp.

Fisher, M., 2004: Generalized frames on the sphere, with application to the background error covariance modelling, *Proc. Seminar on Recent Developments in Numerical Methods for Atmospheric and Ocean Modelling, Reading, UK, ECMWF*, pp. 87–102.

Gauthier, P. and J.-N. Thépaut, 2001: Impact of the digital filters as a weak constraint in the preoperational 4D-Var assimilation system of Météo-France, *Mon. Weather Rev.*, **129**, 2089–2102.

IFS-Documentation, 2015: - CY41r1: Part II: Data assimilation, http://www.ecmwf.int/sites/default/files/elibrary/2015/9209-part-ii-data-assimilation.pdf, 103 pp.

Illingworth, A. et al., 2015: The earthcare satellite: The next step forward in global measurements of clouds, aerosols, precipitation, and radiation, *Bulletin of the American Meteorological Society*, **96**(8), 1311–1332.

Janisková, M. and P. Lopez, 2013: Linearized physics for data assimilation at ECMWF, in *S.K. Park and L. Xu (Eds), Data Assimilation for Atmospheric, Ocean and Hydrological Applications (Vol II), Springer-Verlag Berline Heidelberg*, pp. 251–286, doi:10.1007/978–3–642–35088–7–11.

Liebe, H. J., 1985: An updated model for milimeter wave propagation in moist air, *Radio Science*, **20(5)**, 1069–1089.

Liebe, H. J., P. Rosenkranz, and G. A. Hufford, 1992: Atmospheric 60-GHz oxygen spectrum - new laboratory measurements and line parameters, *Journal of Quantitative Spectroscopy and Radiative Transfer*, **48**, 629–643.

Marchand, R., G. G. Mace, T. Ackerman, and G. Stephens, 2008: Hydrometeor detection using cloudsatan earth-orbiting 94-ghz cloud radar, *Journal of Atmospheric and Oceanic Technology*, **25**(4), 519–533.

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

Stephens, G., D. Vane, R. Boain, G. Mace, K. Sassen, Z. Wang, A. Illingwort, E. O'Connor, W. Rossow, and S. Durden, 2002: The CloudSat mission and the A-train, *Bull. Am. Meteorol. Soc.*, **83(12)**, 1771–1790.

Veerse, F. and J. N. Thépaut, 1998: Multiple-truncation incremental approach for four-dimensional variational data assimilation., *Q. J. R. Meteorol. Soc.*, **124**, 1889–1908.

Winker, D., M. Vaughan, A. Omar, Y. Hu, K. Powell, Z. Liu, W. Hunt, and S. Young, 2009: Overview of the CALIPSO mission and CALIOP data processing algorithms, *J. Atmos. and Ocean. Tech.*, **26(7)**, 2310–2323.