1. Report Working Group 1: Objective Evaluation of the Assimilation System

Participants: R. Ménard (Chair), M. Fisher (rapporteur), Anthony Weaver, Loik Berr, Chiara Piccolo, William Grey, Andrew Lorenc, Gerhard Paul, Lars Isakseen

1.1. Evaluation of the quality of an assimilation system

The group felt that because there are different users of assimilation products it is not possible to identify the best evaluation criterion of an assimilation system. The group has identified several user groups namely; numerical weather prediction, instrument and retrieval validation, atmospheric and climate modellers, assimilation developers, and external users that drives other component of the earth system using meteorological analyses (such as ocean assimilation and chemical assimilation with a CTM). Each of these groups looks at the assimilation system from a different perspective and often arrives at different measure of evaluation of the assimilation system. That is to say that *the evaluation of an assimilation system is user dependent*. In the interest of time however, not all user groups had time to develop specific recommendations and concentrated on the NWP.

1.1.1. Numerical weather prediction

Although the verification of the forecast against analyses is an important tool for verification and that analyses are verified against observations, the group recommends an increase use of independent observations. In particular **the group recommends that forecast be verified against independent observations**.

The issue was then raised whether all observations should be assimilated or some kept for its use in verification. The main issues raised were how detrimental to the main system would it be to withdraw some observations for verification purposes and how independent anyway are those observations. The group came with the following recommendation: that **thinned satellite data (not used in the assimilation system) should be processed using a comprehensive or line-by-line radiation transfer model which would then form an independent data set.**

The use of ensemble forecast and analysis was discussed as a mean to evaluate the analysis and forecast errors. It was recognized that more work needed to be done such as validating the ensemble spread against observation-based estimates, compare the validation of case studies against statistical measures of analysis quality which can be achieved in a statistical sense if the ensemble size is increased. The group felt that the standard statistical measure was perhaps itself somewhat deficient and **recommended to investigate the use of alternative metrics, such as feature-following verification (i.e. spatial verification technique), probability distribution functions, etc …**

1.1.2. Instrument and retrieval validation

Whereas in data assimilation collocated observations are generally avoided, instrument developers actually need this kind of information for validation and error estimation. Inter-calibration (cross-validation) of satellite data, validation campaigns, is important for instrument validation. The group **suggest that more special assimilation experiments be conducted for the specific purpose of validating new instruments. Towards this goal providing the profile analysis error covariance would also be useful**.

1.2. Evaluation of the optimality of the assimilation system

Developing an assimilation system with consistent error statistics may not generate the best forecast. For instance the satellite observations of IASI and AIRS have inflated observation errors to compensate for unknown observation error correlation, and the inflation has shown to improve forecast. The problem is not so much how to account for such errors in the assimilation system than how to determine these error statistics. On the model side, the effect of model biases diminished by adjusting background error covariance to provide the best forecast. Since B and Q are too big to ever be known entirely, the problem is how to reduce the dimension of these matrices down to a few relevant unknown parameters based on physical insight. In light of these remarks, **it is recommended that comprehensive observation and background/model error covariance based on physical insight continue to be further improved.**

1.3. Observation network diagnostics

Although this topic is the main focus of another discussion group, some discussion about this issue was made. OSE's and adjoint methods provide different and somewhat complementary information. It is recommended that both methods be used to assess observation network. The adjoint sensitivity methods are useful but we must understand the limitations. **The group recommends that adjoint sensitivity be further developed and carefully interpreted such as with comparing with OSE', testing the sensitivity by perturbing observations, and investigate using a cost function with respect to observations rather than analyses**.

1.4. Diagnostic tools

It is believed that the development of the EnKF to be used as diagnostic tool is advantageous. It can help in better understanding the specification of the background and observation error covariance matrices given the fact the Kalman gain is provided. Also because the Kalman matrix is exactly computed would allow the exact computation of the observation influence and the related cross-validation score which can be used for quality control purposes.

2. Report Working Group 2:

Participants: R. Gelaro (chair) T. McNally (rapparteur), L. Xu, A. Rhodin, D. Daescu, A. Benedetti, E. Kalnay, F. Prates, P. Poli, W. Grey, M. Masutani

2.1. Sensitivity WRT DA input parameters

The group feels that in terms of the calculations that are made routinely to monitor observation forecast impact (FSO) and analysis sensitivity (AS), the implementation of additional diagnostics based on the sensitivity to input parameters (σ_o, σ_b) represents only a modest overhead (human and computer) and as such should be implemented routinely. In particular the sensitivity to the specification of B variances provides a level of detail that is not currently available with other methods.

Until these are tried it is difficult to envisage the outcome – it is also less obvious how we would act upon such information (e.g. to change σ_o, σ_b) if we have a strong requirement to respect balance and other constraints. However, in the GEMS type environment (where such constraints are less acute and the existing knowledge is less mature) the group sees a more ready application of these diagnostics.

Recommendation: ECMWF should investigate and adopt these diagnostic tools if appropriate

2.2. Error bars on FSO and AS diagnostics

These diagnostics are now widely used and there is increasing confidence in their application. There appears to be good consistency with established OSE based diagnostics and offers a much more potentially powerful dissection of the results (that would be computationally unfeasible with other methods).

Sources of uncertainty include:

- Errors in verifying analysis (errors comparable to 24hr forecast error ; should be less problematic in EnKF where verification not restricted to 24 hrs)
- accuracy / linearity of TL/AD model
- definition of the cost function / metric of success
- natural statistical uncertainty

The group feels that error bars can and should be placed upon these diagnostics. The simple statistical spread can be used (with significance testing), but this will not encapsulate / quantify some of the sources of uncertainty mentioned above (i.e. is not sufficient to quantify the total error in the diagnostic). There was a very clear feeling that this diagnostic should not be used in isolation, but as a complement to existing diagnostics (data statistics, OSE…).

Recommendation: ECMWF in collaboration with other centres should develop a method to place error estimates on these diagnostics

2.3. Metric for sensitivity analysis

The adjoint results measure the response of a single forecast metric to 'all' perturbations of the observing system, while OSEs measure the response of all metrics to a single perturbation of the observing system.

Therefore careful consideration should be given to the choice of forecast metric in the adjoint context (results to date focus mostly on a dry energy-based metric).

The choice is clearly application driven and if specific information on the impact of particular data (e.g. individual satellite channels) is required an appropriate metric that reflects what is actually measured should be chosen.

Similarly, if the impact of observations on a particular scale is required, a metric tailored to this application should be designed (e.g. short scales v longer scales).

One obvious candidate is a metric that reflects the errors forecasting of humidity and the sensitivity of humidity observations.

Recommendation: There is no single optimal forecast measure we can recommend. Therefore, some experimentation with a reasonable, limited set of forecast measures is encouraged.

2.4. OSEs

The OSEs are critical part of the measuring observation impact, but by their nature (granularity, cost, perturbing the system) cannot be comprehensive and should therefore be used in conjunction with other tools such as adjoint based methods.

Assuming that the centre is currently running as many OSEs as is possible given the available resources the group focussed on how these experiments are run and how the results are interpreted.

Given the cost it is obvious that as much coordination as possible over the periods tested is important (within the constraints of new data availability).

It is also important to appreciate that the construction of the OSE may influence the outcome of a data impact study.

"NOSAT baseline + new system" may be misleading and optimistic in an operational context due to the very different underlying background error structures (this would be less problematic in EnKF). These are very relevant and useful for re-analysis where significantly degraded observing systems are more likely.

Evaluation of the OSEs should routinely incorporate appropriate metrics (i.e. in addition to traditional 500z scores – weather parameters – extreme events). This is done by Operations and has relevance for users so it should be part of the development process. The sensitivity to the choice of verifying analysis when evaluating short range forecasts can be acute.

Recommendation 1: ECMWF should continue to exploit OSEs to the fullest extent possible and should be used in conjunction with other tools such as adjoint based methods.

Recommendation 2: Evaluation of the OSEs should routinely incorporate appropriate metrics (i.e. in addition to traditional 500z scores – weather parameters – extreme events).

2.5. OSSEs

The obvious advantage of OSSEs is that “truth” is known as opposed to other measures (OSEs and adjoint tools) where “truth” is uncertain. They are also the best available mechanism to test hypothetical scenarios such as new observing systems in a realistic environment.

It is important to verify that the OSSE has similar statistical characteristics as a real system with respect to existing observations (e.g. the forecast impact of all currently used systems in the OSSE is the same as that found in OSEs, also the tuning of observation errors to achieve similar innovation statistics studied).

The utility of an OSSE (measured by the reliability of the outcomes) is sensitive to what extent the use of a new (proposed) observing system stresses the skill of the current assimilation system (e.g. will the data assimilation system in 10 years time give the same result, possibly with a very different baseline observing system). It was noted that that while OSSEs can be optimistic about observation impact, advances in DA skill will usually mean we make better use of observations in the future and offset this to some extent.

Recommendation: ECMWF should continue to contribute to the existing international OSSE effort.

2.6. Re-analysis as a diagnostic tool

RA has demonstrated the need for a robust bias correction system and tested the long-tern stability of these. It also documents improvements of the DA system over time subject to a changing observing system. The longer-term evaluation of new diagnostic tools will be possible with RA to gain experience (e.g. understanding the uncertainties and statistical significance) that would otherwise be slowly gathered in the operational context.

Recommendation: ECMWF should continue its RA efforts and extend to validating DA diagnostics

3. Working Group 3: Diagnosing model and forecast error

Participants: Ricardo Todling (Chair), Niels Bormann, Carla Cardinali, Dick Dee (rapporteur), Gerald Desroziers, Olivier Talagrand, Yannick Tremolet, Nedjeljka Zagar

3.1. General points made during the working group discussions

- No diagnostics can provide clear-cut answers without the introduction of additional or external hypotheses on the errors affecting the data (including the model-generated background fields as well as observations)
- It is important to understand the specific limitations of individual diagnostics
- Therefore there is need for a diversity of diagnostics and for the accumulation of many pieces of evidence
- We need to identify the key pieces of information and make them easily accessible, and also educate the users regarding the use of this information
- We recognize that the complexity and diversity of the available tools requires good communication among different efforts
- It is important not to lose sight of the connection with the underlying physical aspects of the atmosphere or ocean in diagnostic work

3.2. The role of model bias and how to diagnose it

- The presence of model bias affects the outcome of most (if not all) performance measures
- Resources and methods for diagnosing model biases can include:
    - time series of analysis increments, including modal decomposition
    - short- to medium-range forecasts vs analysis (as used in ECMWF's Predictability and Diagnostics Section)
    - short- to medium-range forecast vs observations: extended departures
    - reanalysis data
    - carefully designed Observation Simulation Experiments (OSEs)
    - information obtained with variational bias correction (VarBC)
    - comparison with independent data
    - examination of dynamical consistency: balance, energetics, modal analysis
- It is recognized that biases can also be introduced into the system due to improperly specified analysis ingredients (e.g., misspecification of background errors; observation operators)
- Weak-constraint 4D-Var will provide estimates of model biases that should be compared with estimates obtained from independent diagnostics
- There is a need for diagnostic techniques that do not depend on the assumption that the system is unbiased

3.3. Diagnostics for (non-systematic) model errors

- We still do not know much about the forecast model's signal-to-noise ratio
- Possible techniques for estimating model error statistics include:

  - Use of Lorenz 1982 curves to estimate model error growth
  - Varying the window length in (strong-constraint) 4D-Var
  - Studying the evolution of analysis departures within a 4D-Var window
  - Application of Desrozier's diagnostics with an ensemble of analyses
- These estimates should play a role in the development of weak-constraint 4D-Var

3.4. Diagnostics for background error covariance models

- These will be especially important when flow-dependent covariance models are introduced
- Currently, there are no obvious ways to visualize the background error covariance B as used in the ECMWF system. Possible techniques include:
  - Applying B to columns of the identity matrix at a few selected grid points
  - Performing single-observation analyses
  - Comparing maps of observation influence diagnostics for satellite data with maps of specified background and observation error variances
  - Compute and visualizing maps of prescribed background error variances for model variables and for observables
- Routine use should be made of Desrozier's diagnostics, also for diagnosing vertical correlations
- In case of flow-dependent variance specifications:
  - Consistency of specified background error variances with Desrozier's estimates should be checked on a map
  - Sensitivity of the cost J to the specified background error variances could be computed (as discussed by Daescu in his presentation)

3.5. Observation impact assessment in the presence of model bias

- Model biases can make it appear that observations have negative impact, when in fact they are partially correcting the model biases
- One should avoid using the analysis as verification in such cases, but rather verify against observations (e.g. Todling's poster)
- It is important to keep in mind that observation impact is a statement about the data assimilation system – not necessarily about the quality of the observation

3.6. Recommendations

1) A variety of diagnostics should be looked at for reliable performance assessment and problem identification
2) The technical work to allow the use of forecast departures for all observations should be completed and these should be used routinely to assess forecast skill
3) Desrozier's diagnostics should be routinely monitored in operations along with observation departures
4) Observation impact assessment tools should be extended to verify against observations and to consider a variety of forecast aspects (e.g. Todling's poster)
5) A number of innovative diagnostic tools and applications should be developed, such as the use of Desrozier's diagnostics for estimating vertical error correlations (see items listed on previous slides).