

# AN INTRODUCTION TO DATA ASSIMILATION

**Xiang-Yu Huang and Henrik Vedel**

Danish Meteorological Institute, Lyngbyvej 100, DK-2100 København Ø, Denmark.

Email: xyh@dmi.dk and hev@dmi.dk

## ABSTRACT

Data assimilation is discussed from an operational numerical weather prediction (NWP) point of view. Examples of the observations used operationally are given. Preprocessing and quality control of observations are then illustrated with a few examples. Different data assimilation methods are summarised, with a discussion of some of the basic assumptions adopted by the widely used statistical methods. The importance of good quantitative knowledge about the errors of both the NWP model fields and the observations is stressed. Finally it is discussed how the route from a novel type of observation being available to the observation having a proven, positive impact on the forecasts may well take many years, even decades, of hard work for each new data type, including not only the data assimilation system, but also extensive case studies and long parallel NWP model runs.

## 1. INTRODUCTION

Numerical weather prediction (NWP) is an initial value problem. Provided an estimate of the atmospheric state, in terms of the variables of the NWP model, the model simulates the atmospheric state at later times. It also calculates precipitation and other important properties used by weather forecasters in the production of the public weather forecasts.

Determination of the starting state for an NWP model is called data analysis or data assimilation. The resulting state we call the *analysis*, estimation of which is a non trivial problem for a number of reasons. On the other hand the quality of the analysis is crucial for the NWP forecast skill, and it is to a large degree due to improvements of the data assimilation systems and addition of satellite data that the forecast skill has improved significantly during recent years.

A central problem is that observationally determination of the initial state is an under-determined problem. The model variables outnumber the observations and further the observations do not sample in a homogeneous way the volume simulated by the NWP models. Auxiliary information is necessary. This could be in the form of assumptions (e.g. if using extra- and interpolation to populate data void areas of phase-space), but normally one uses instead the data from a short-term NWP forecast, the so called *background* or *first guess* field. This is illustrated in figure 1, in which the

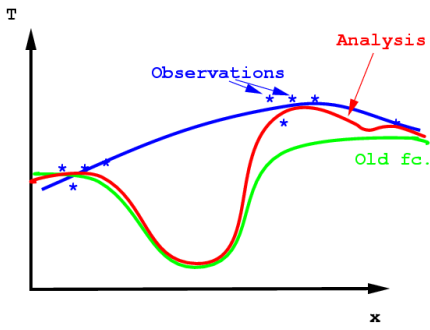


Figure 1. Gedanken data analysis

crosses can be thought to represent simultaneous (synoptic) temperature measurements along a line. The blue curve is the result one would obtain based on only the observations, the green curve is the prediction by a NWP model. Because of the skill of the model in advecting atmospheric information and in simulating the atmospheric dynamic evolution, the low temperatures expected in the data void region are far more likely than the prediction based on interpolation. As a consequence a proper analysis should be close to the forecast in that region, while it should be close to the observations near those, as indicated by the red curve.

A second problem is that many important observations do not correspond directly to the model variables, but rather to complicated combinations thereof. The ability of different data assimilation methods to cope with this varies significantly.

The basic idea in modern meteorological data assimilation is to combine many, widely different types of observations and the background field in the compilation of an estimate of the state of the atmosphere at a given time. In the case of statistical data assimilation the resulting field, the analysis, is the statistically most likely state of the atmosphere given the information at hand and provided the statistical assumptions adapted are valid. As long as the statistics give a proper description of the errors even poor data will lead to an improved analysis.

In the following we give a very brief introduction to data assimilation. Section 2 presents an overview of the observing systems presently contributing data to the data assimilation at DMI, section 3 discusses screening and preprocessing of observations, section 4 lists the main data assimilation methods, section 5 describes variational data assimilation, section 6 how to include new observations.

## 2. OBSERVATIONS

Figures 2 to 5 provide examples of the observations used operationally at DMI in the data analysis. TEMPs and AIREPs provide data about the column above the location, as do (A)TOVS data, while the other data are obtained at or near ground. The satellite data are down taken by DMI and are for one data assimilation time-slot. A longer period data assimilation window, such as used at ECMWF, would give a more even global coverage. Note the homogeneity of the distribution of the satellite data in contrast to all other data. Both this and the good time coverage by satellite data is a great asset. On the other hand the vertical resolution of the satellite data currently in use is not optimal. This is a prime reason GPS RO data are very interesting to meteorologists.

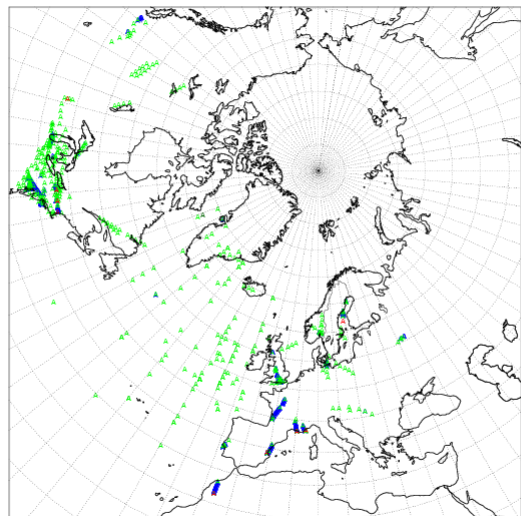


Figure 2. AIREP aircraft observations.

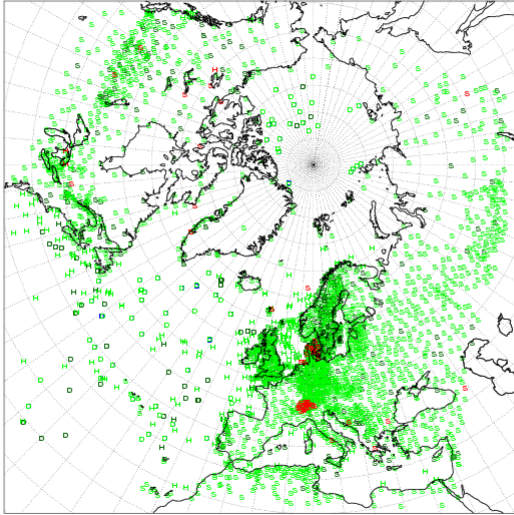


Figure 3. SYNOP and SHIP surface observations

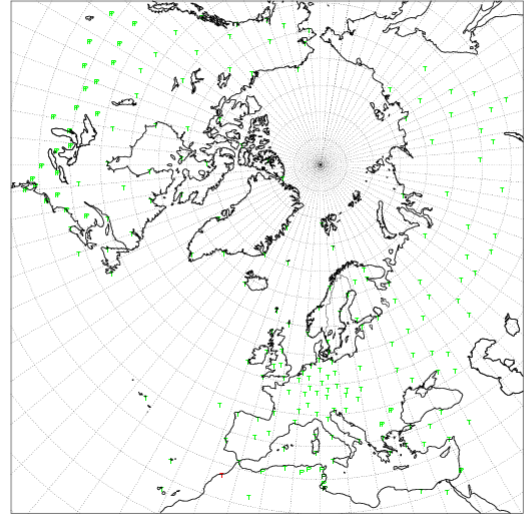


Figure 4. TEMP and PILOT balloon observations

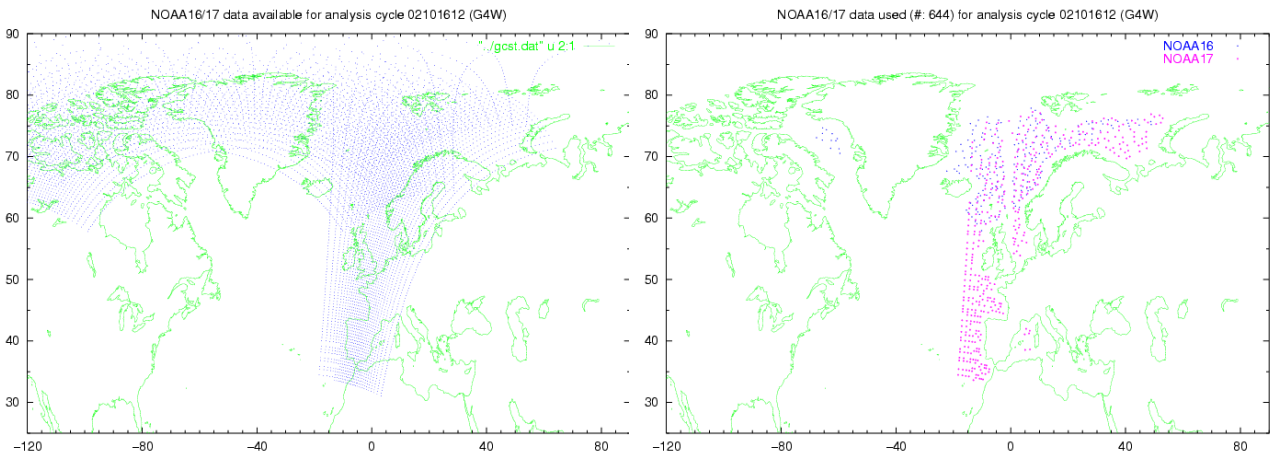


Figure 5. Satellite (A)TOVS data available for assimilation before (left) and after thinning and screening of observations over land and in cloudy areas.

The above describes observations used in data assimilation, but one should not forget that a wealth of observations are made to assist the forecasters in other ways, often not intended for assimilation at all.

### 3. QUALITY CONTROL OF OBSERVATIONS

All observing systems have their limitations, problems, and failures, resulting in the reported measurements being sometimes incorrect. Such data must be identified and rejected by the data assimilation system in order to avoid corruption of the analysis. Due to the amount of data handled this is done by automatic routines, both in the form of preprocessing and during the data assimilation stage.

The systems used to quality control the observational data include

- *Bad reporting practise check.*
- *Blacklist check.* For stations which are found to 'always' report erroneous data.
- *Gross check.* Against some limits, e.g. from climatology.
- *Background check.* Based on the deviation between the observation and the expectation based on a short term forecast.
- *Buddy check.* Checking against nearby observations.
- *Redundancy check.*

In addition observations are selected or rejected according to the time they were obtained. For a given type of observation from a given site the observations closest to the centre of the data assimilation time window is chosen. For 3DVar and most other assimilation systems (see later for definitions) this window consists of a single rather large time slot, e.g. 3 or 6 hours, while in 4DVar the window is broken down into smaller sub-windows, being typically 1 hour wide each, while the full window may be larger than in 3DVar. As a result more observations can be utilised in 4DVar.

As an example table 1 shows the number of active respectively rejected observations in two runs for the early part of December 1999, one with 3DVar and one with 4DVar. (Satellite data were not included in these runs. The full assimilation windows had equal length.)

	3DVAR		4DVAR	
	Active	Rejected	Active	Rejected
LAND SYNOP	2291	3828	7159	1072
AIREP	5012	3338	5125	3250
BUOY	136	616	476	334
TEMP	115	39	116	38
PILOT	20	67	42	49
SHIP	306	386	749	97

*Table 1. Observation data usage (average from 36 cycles during 1999.12.01 - 09).*

The highest level quality control takes place during the assimilation cycle itself, where the statistical description of the errors of both the forecast field and the observations enables a graduation of the observations between highly likely and probably not true. This is exemplified in table 2. Notice also how the radiosondes (TEMP) provide a significant amount of the total data, despite the number of radiosondes launched being relatively small. The radiosondes are the only ones providing humidity information. From the SYNOP stations only pressure is used by the current system. Quality indicator 1 means the observation is probably correct degrading to factor 4, which means probably incorrect.

From the tables it can be seen we are far from using all the observations currently available. This depends not only on the data quality, but also on the observing system, on the NWP model resolution, and on the the data assimilation system and method. We would like to utilise more of the existing

VARQC flags:		3DVAR				4DVAR			
		1	2	3	4	1	2	3	4
SYNOP	height	2121	15	7	26	6726	59	30	74
BUOY	height	69	0	0	0	300	0	0	1
SHIP	height	282	1	0	4	695	1	0	5
AIREP	temperature	4958	3	2	1	5063	6	2	2
	U-wind	4857	16	10	29	4899	50	30	45
	V-wind	4857	16	10	29	4899	50	30	45
TEMP	temperature	4074	11	6	10	4073	20	10	12
	U-wind	3313	8	6	35	3291	26	17	55
	V-wind	3313	8	6	35	3291	26	17	55
	Humidity	3447	1	0	2	3458	3	2	3
PILOT	U-wind	323	0	0	1	1193	7	4	12
	V-wind	323	0	0	1	1193	7	4	12

Table 2. Variational quality control flags (averaged from 40 cycles during 1999.12.01 - 12.10).

data, but at the same time we have to prepare for assimilation of novel data expected to be important to NWP model skill, like GPS data.

#### 4. DATA ASSIMILATION METHODS

A large number of data assimilation methods exists for meteorology. Which method is adopted depends mainly on the NWP model in question (type, area) and the available resources for data assimilation (cpu and manpower). It is a significant fraction of the total NWP computing time which is spent on data assimilation, wherefore time constraints play a dominating role when selecting data assimilation method. Mainly for this reason adaptive statistical methods are not yet used operationally.

The methods can be divided into classes:

- Empirical methods
  - Successive Correction Method (SCM)
  - Nudging
  - Physical Initialisation (PI), Latent Heat Nudging (LHN)
- Constant statistical methods
  - Optimal interpolation (OI)
  - 3-dimensional variational data assimilation (3DVar)
  - 4-dimensional variational data assimilation (4DVar)
- Adaptive statistical methods
  - Extended Kalman filter (EKF)
  - Ensemble Kalman filter (EnFK)

## 5. 3DVAR AND 4DVAR

3DVar is the data assimilation method currently used at DMI and most other met. centres. We are moving toward 4DVar, which is presently used at only a few, very large met. centres.

In variational data analysis the analysis is found by minimisation of a cost function,

$$J_{3DVar} = \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}_b) + \frac{1}{2}(H(\mathbf{x}) - \mathbf{y})^T \mathbf{R}^{-1}(H(\mathbf{x}) - \mathbf{y}) \quad (1)$$

$$J_{4DVar} = \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) + \frac{1}{2}(H(M_i(\mathbf{x}_0)) - \mathbf{y}_i)^T \mathbf{R}^{-1}(H(M_i(\mathbf{x}_0)) - \mathbf{y}_i) \quad (2)$$

with respect to the state vector  $\mathbf{x}$ . Here  $\mathbf{y}$  is a vector containing all observations, while  $H$  is a so-called *observation operator*, which maps from model space to observation space (examples will be given later).  $\mathbf{x}_b$  is the background (a recent NWP forecast, valid at the time of the data analysis).  $\mathbf{B}$  is the error covariance matrix of the background field,  $\mathbf{R}$  is the error covariance matrix of the observations. The matrices are of order  $N_{NWP\text{-variables}}^2$  and  $N_{obs}^2$ , approximately  $10^{14}$  and  $10^8$  for a typical NWP run at DMI.

In 3DVar one operates with a single data assimilation time window centred on the time of the analysis, e.g. a 3 or 6 hour window for a NWP system making a new analysis for every 6 hours, or a 3 hour window for a 3 hour cycling system. For each observational site the system selects the datum closest in time to the analysis time, if there are other observations from the site they are rejected.

In 4DVar the observations are binned into the time-slots  $i$ , and  $M_i$  is the NWP model operator which turns state vector  $\mathbf{x}_0$  into its forecast value at time  $i$ . Typical time-slots are 1 hour wide. This enables one to use observations from the same site obtained at different times, and it minimises the offset in time between the time of the observations and the valid time for the forecast fields against which the observations are compared. The drawback is that 4DVar requires far more computer-time.

Normally  $\mathbf{R}$  is assumed diagonal, meaning the errors of the observations are not correlated. However,  $\mathbf{B}$  contain non diagonal elements, which are very important to the data analysis, as they determine how the corrections of the background state due to the observations are spread spatially in the region surrounding the observation, both vertically and horizontally. Further the correlations lead to a spreading of observational information concerning one variable to other variables not directly related to the observation. For example an observation of temperature will lead to corrections not only of the temperature field of the background state, but also of pressure, wind, etc. In practise the minimum is found iteratively. This requires an expression for  $\nabla_{\mathbf{x}} J$ . In deriving that it is assumed that  $H$  and  $M_i$  can be described by a Taylor expansion to first order around  $\mathbf{x}_b$  for relevant values of  $\mathbf{x}$ , so that  $H(\mathbf{x}) = H(\mathbf{x}_b) + \mathbf{H}(\mathbf{x}_b)\delta\mathbf{x}$ , where  $\mathbf{H}$  is the so-called *tangent linear observation operator*, and  $\delta\mathbf{x} \equiv \mathbf{x} - \mathbf{x}_b$  is the so-called *analysis increment*. In 4DVar it is further assumed that within the data assimilation time slot the model can be linearised for the time evolution of the increments, such that  $M_k(\mathbf{x}_0) - M_k(\mathbf{x}_b) \approx (\prod_{l=0}^k \mathbf{M}_l) \delta\mathbf{x}_0$ .

### 5.1 Observation operators

The observation operator  $H$  maps from model state  $\mathbf{x}$  to observation space.

For an observation corresponding to a variable of the model,  $H$  will be an operator performing just an interpolation in the model field to the location of the observation. (In addition it will perform a

number of other transformations, but they are related to the particular properties of the NWP model and the data assimilation system, and are disregarded here.)

For zenith total delays (ground based GPS measure)  $H$  includes

$$ZTD = p_a f(\theta, h) + \frac{1}{g(\theta)} \sum_{i=1}^N q_i (p_{i+1/2} - p_{i-1/2}) \quad (3)$$

where  $p_a$  is the pressure at the GPS antenna,  $p_{i\pm 1/2}$  are pressures at the model half levels,  $q$  is specific humidity,  $g$  is the gravitational acceleration,  $\theta$  is the latitude and  $f$  is a function depending on the geographical location of the site.

For refractivity the observation operator includes the part

$$N = \frac{p}{T} \frac{1}{1 + q(1/\epsilon - 1)} \left( k_1 + \frac{q}{\epsilon} (k_2 - k_1/\epsilon + \frac{k_3}{T}) \right), \quad (4)$$

where  $p$ ,  $T$ , and  $q$  are pressure, temperature and humidity at the location in question and the remainder are constants.

For radiance data, (A)TOVS, the observation operator includes a complicated integral depending on the radiative transfer up through the atmosphere.

The observation operator depends on the level of preprocessing. For example the refractivity profile could be converted into a temperature versus pressure profile, which could be assimilated instead. It is widely found that it is beneficial to assimilate the observations in as raw a form as possible, but for many types of satellite data some preprocessing is needed in order to speed up the data assimilation. Operational NWP models run under strict time constraints, which calls for compromises when making the data assimilation system.

## 5.2 Determination of error statistics

If  $\mathbf{B}$  and  $\mathbf{R}$  are correct the outcome of the data analysis corresponds to the maximum likelihood estimate of the atmospheric state, given the available information. Obviously precise determination of  $\mathbf{B}$  and  $\mathbf{R}$  is mandatory for a high level data assimilation system.

In principle the covariance matrices should be determined by comparing respectively the model fields and the observations to the *truth*, the real atmospheric state. Being out of reach a number of methods are adopted for estimation of the two important matrices by other means.

Dealing with this it is assumed that the observations and the background are unbiased. If known biases exist, they are removed. Further it is assumed that errors of the observations and the background are not correlated.

Estimation of  $\mathbf{R}$  :

Normally a data provider has some sort of error estimate for the data. This can be made, for example, from theoretical arguments or from comparisons with auxiliary measurement equipment. However, what  $\mathbf{R}$  represents is the errors of the observing system seen from an NWP model point of view. Often the properties measured by an observing system varies within one NWP model grid-box, and

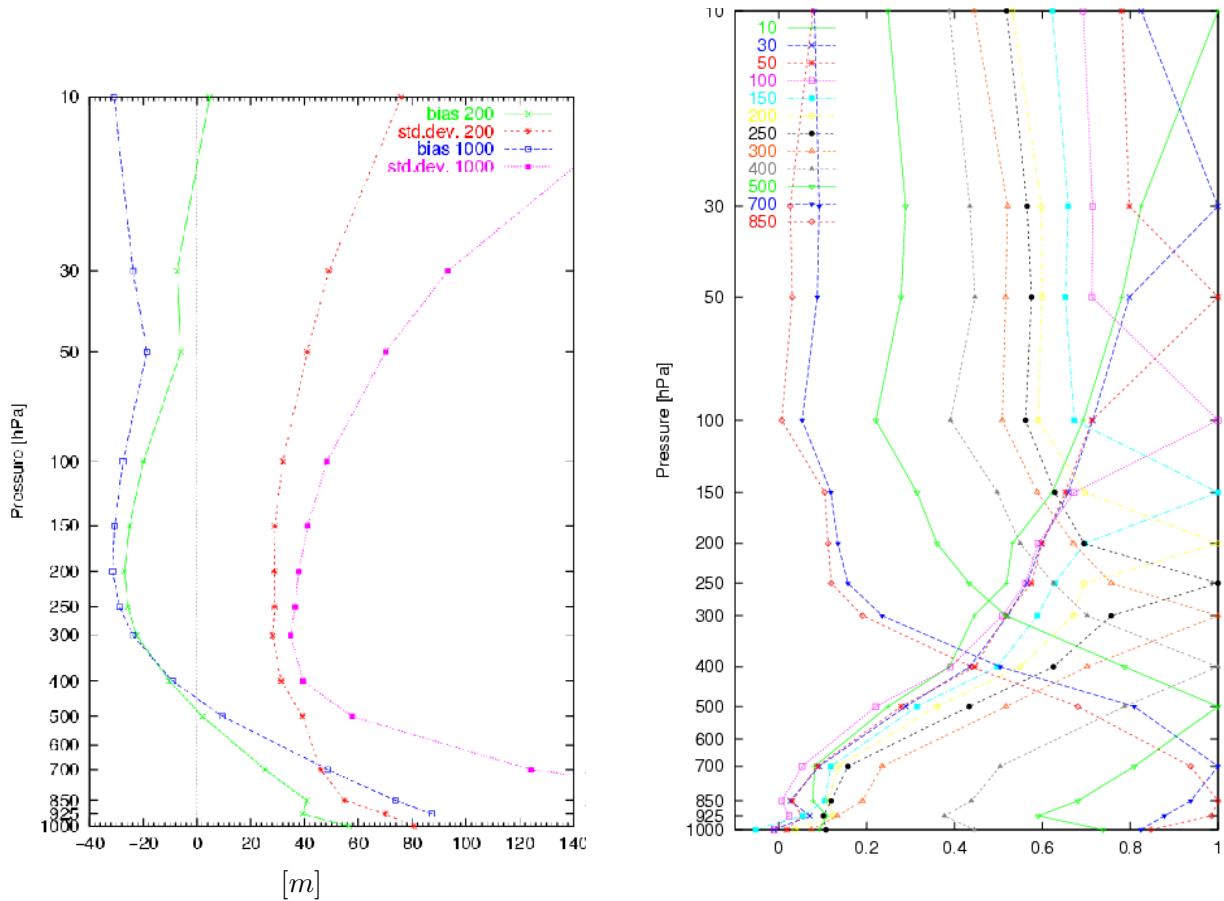


Figure 6. Bias and standard deviation between GPS/MET RO geopotential height data and ECMWF analyses for two values of accepted maximum deviation (left). Correlations of errors for GPS/MET RO geopotential height data.

$\mathbf{R}$  should include these variations, named 'errors of representativeness'. For some observations they can be dominant. When they are significant,  $\mathbf{R}$  will depend on the model resolution.

One way in which to get a rough estimate of  $\mathbf{R}$  for a novel observation type is to determine the statistics of the offsets between the observations and the NWP model analyses. This gives an upper limit to the observational errors, as the errors of the analysis are included. An example is seen in figure 6 in which GPS/MET RO data in the form of geopotential heights were compared to ECMWF analyses prior to an impact experiment with inclusion of GPS/MET data. The strong dependence of the bias on the maximum allowed offset is due to a significant number of poor observations in the GPS/MET data set.

A much better option is to use the so-called Hollingworth Lönnerberg method ( see Hollingworth and Lönnerberg (1986) and Lönnerberg and Hollingworth (1986)), which enables determination of both observation error, background error, and background error correlations, under the assumption that the observation errors are not correlated.

But this may not always be practically possible. In this regard a particular problem related to GPS observations is the existence of spatial error correlations of the observations. For example errors in the orbits of the GPS satellites will introduce correlations in the errors of the RO data.

Estimation of  $\mathbf{B}$  :

- 1 The Hollingsworth Lönnerberg method (discussed above).
- 2 NMC method. Assumes that background error covariances are proportional to correlations of forecast differences between forecasts of different age, e.g. 24 and 48 hours, but same valid time.
- 3 Analysis ensemble method. The offsets between short term forecasts made from different perturbed analyses are used to estimate  $\mathbf{B}$ . Various methods exist for perturbing the analyses. Sometimes also varying the model 'physics' (met. language for the non-resolved, parametrised part of the NWP model) is included in the variations.

In real applications, because we need the inverse of this enormous matrix, a selection of the analysis variables is made to remove some dependencies among the model variables and often a transformation is made to spectral space (like when using a spectral NWP model) in which  $\mathbf{B}$  is assumed diagonal. Under some additional assumptions and constraints the corresponding transformed  $\mathbf{B}$  is determined. In reality all NWP model variables have error correlations locally, leading to a multivariate data analysis. However, humidity is currently constrained to be separate, because both observations and models contain relatively large humidity errors, and it is important to maintain the proper balances between the pressure gradients, the wind field and the temperature field, as otherwise perturbations corresponding to fast developing waves may be introduced in the model state.

Figure 7 shows as an example the assimilation of GPS/MET data from 5 occultation measurements, in the form of the difference between the analysis field with/without the GPS/MET geopotential height data included in the data analysis. On the left one notices the spreading of the observational impact to the region surrounding the observations. The figure at right illustrates the multivariate nature of the assimilation, as data assimilation has also changed the wind field.

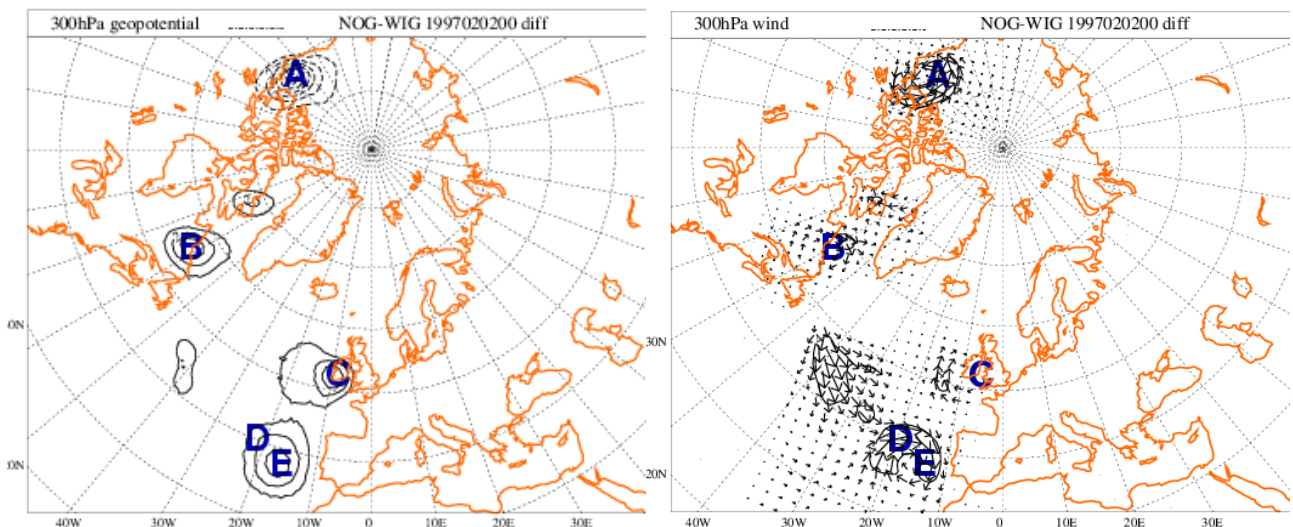


Figure 7. Impact of GPS RO observation on analysis. Left: Difference in 300 hPa geopotential height with/without GPS data. Right: Difference in wind at 300 hPa level with/without GPS data. Data from GPS/MET. Letters indicate location of GPS RO measurement.

In 3DVar and 4DVar  $\mathbf{B}$  is constant. In reality the errors of the background field depend partly on the atmospheric situation, and ideally  $\mathbf{B}$  should reflect that. A possible solution is to use Kalman filter methods, which attempt to predict the evolution of  $\mathbf{B}$ . However, currently they are prohibited by the large cpu requirements. As a cheap remedy some use different  $\mathbf{B}$ 's, depending on, e.g., the season.

As more observations become included in the NWP system and the NWP model itself is improved,  $\mathbf{B}$  will change gradually. Similarly the observing system normally improves with time. Consequently the two error covariance matrices must be re-estimated now and then.

In the normal formulation of variational data assimilation it is assumed that the error distributions are Gaussian. This is not always the case, in particular when physical on/off processes are involved. One such process is the condensation of water vapour which speeds up dramatically when the water vapour saturation pressure is reached. Sometimes such problems can be alleviated by a change of variable, but not always.

## 6. INCLUDING A NEW TYPE OF OBSERVATIONAL DATA

The inclusion of a novel data type goes through a long sequence of steps, from the research level to the operational level. These steps are

- Development of observation operators and simple checks. Determination of error statistics.
- Implementation of the above.
- Analysis increments are studied to verify proper functionality of the routines handling the new data.
- Case studies are performed to assess the impact of the new data in selected cases. These can include important weather events often used for testing at a given met. centre, or cases in which the novel data type is expected to be most useful.
- Extensive parallel experiments (e.g. one month for each season), in which the forecast skill with and without the new data are compared. The comparison is done both for *Standard scores* bias, rms, correlations, etc. of the forecast variables and closely related measures, as well as for *Special properties*, such as precipitation, surface fluxes, etc.
- Special aspects: noise, spin-up effects, etc.
- Pre-operational tests (additional parallel runs with/without the data included).
- Monitoring operational use to provide feedback to further research, and to tune use of data.

In practise it can take years and decades for an observing system to reach operational status. On the other side an observing system which is used for operational NWP may become redundant due to new observing systems, advances in assimilation techniques, and improvements in other components of the forecast system. Continuous monitoring and further tuning are necessary to optimise the use of data from an observing system in use.

Verification can be done in numerous ways, but the most trustworthy is statistical verification of forecasts against observations and against analyses. Figure 8 gives an example of observational verification with and without satellite radiance data included in the data assimilation. In each figure the pair of

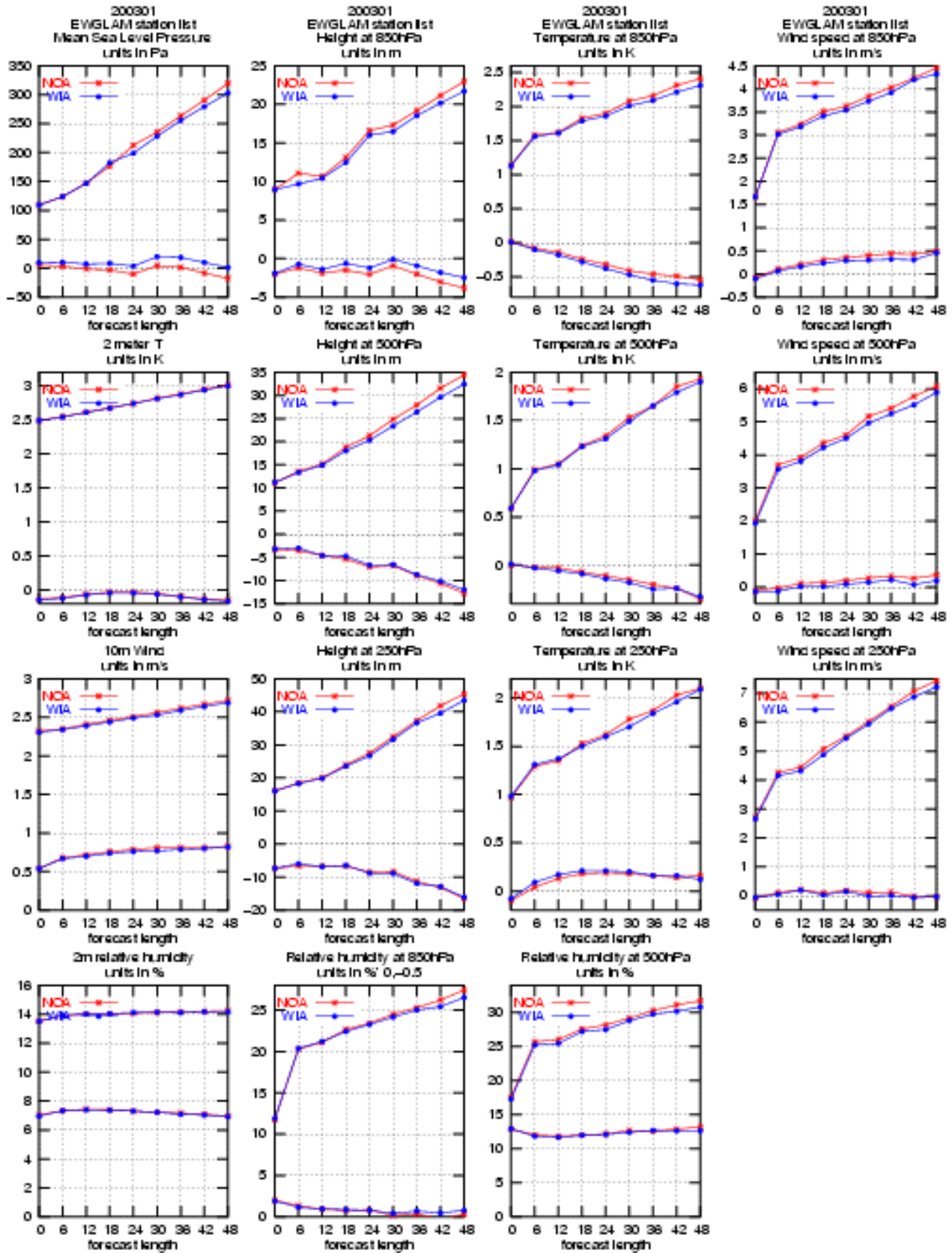


Figure 8. Example of statistical verification of forecasts against observations. In each panel, the upper two curves are for rms errors and the lower two curves are for bias. The forecasts are based on analyses made with/without (A)TOVS data (labelled: WIA/NOA). Plot made by Bjarne Amstrup, DMI.

lower curves show biases, the pair of upper curves the rms. The closer the curves to the x-axis the better. A clear positive impact of the (A)TOVS data is seen.

More often the impact will be more marginal or even neutral for a certain type of new observations, but they may improve the forecasts locally or under specific, important weather conditions, in which case the observing system is still of interest to the meteorological centres. Figure 9 shows an example for a case study done by us in which forecasts made from analyses with and without ground based GPS zenith total delays were compared. The finding in the study was that ground based GPS improves the forecasts of significant precipitation, but the statistical verification showed an overall neutral impact.

It is getting more difficult for a new observing system to have a positive impact, as the NWP models become better and the existing observing systems improve.

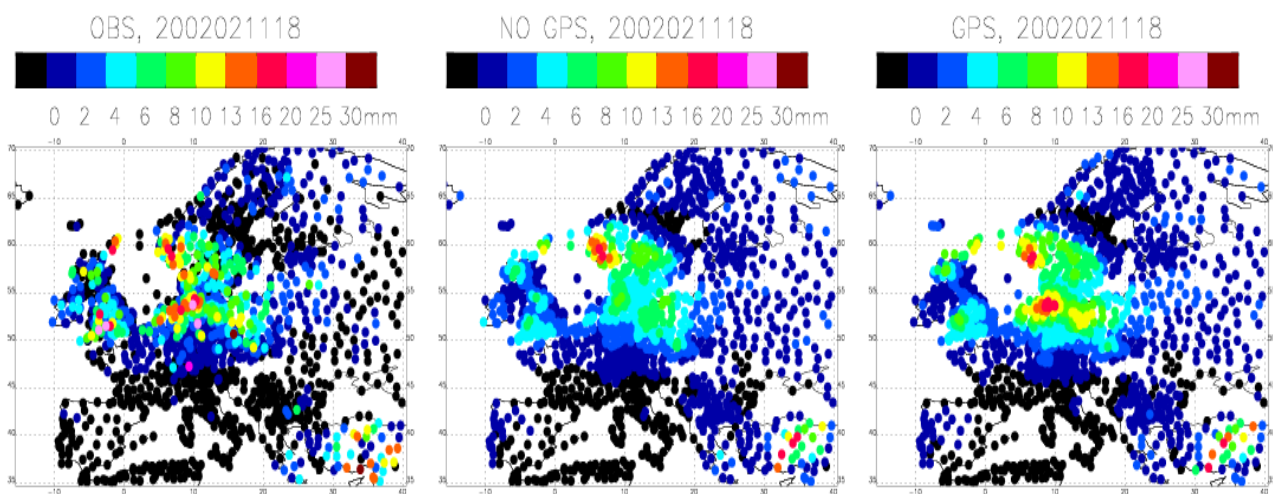


Figure 9. Comparison of observed and predicted 12 hour precipitation with/without ground based GPS zenith total delay observations in the data assimilation.

When it comes to GPS RO data our strong interest in the data is due to the fine global coverage in connection with a good vertical resolution and insensitivity to clouds. The two latter facts contrast most other satellite data, and use of GPS RO is expected to be not only beneficial in itself, but also improve our use of for example radiance data, which have very fine horizontal resolution and play a major role in today's data analyses. Positive impact from real GPS RO data collected from a single LEO has already been found on one of the most advanced data assimilation systems.

## 7. FURTHER READING

Among recommended literature for people interested in data assimilation are Daley (1991), Lorenc (1986), Kalnay (2003), and references therein. More details about the data assimilation system used at DMI can be found in Gustafsson et al. (2001) and Lindskog et al. (2001).

## References

Daley, R., *Atmospheric data analysis*, Cambridge University Press, 1991.

- Gustafsson, N., Berre, L., Hornquist, S., Huang, X.-Y., Lindskog, M., Navascues, B., Mogensen, K. S., and Thorsteinsson, S., Three-dimensional variational data assimilation for a limited area model. Part I: General formulation and the background error constraint., *Tellus*, 53, 425, 2001.
- Hollingworth, A. and Lönnberg, P., The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field., *Tellus*, 38A, 111, 1986.
- Kalnay, E., *Atmospheric modeling, data assimilation and predictability*, Cambridge University Press, 2003.
- Lindskog, M., Gustafsson, N., Navascues, B., Mogensen, K. S., Huang, X.-Y., Yang, X., Andrae, U., Berre, L., Thorsteinsson, S., and Rantakokko, J., Three-dimensional variational data assimilation for a limited area model. Part II: Observation handling and assimilation experiments., *Tellus*, 53A, 447, 2001.
- Lönnberg, P. and Hollingworth, A., The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: The covariance of height and wind errors., *Tellus*, 38A, 137, 1986.
- Lorenz, A., Analysis methods for numerical weather prediction., *Q. J. R. Meteor. Soc.*, 112, 1177, 1986.