Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

Treatment of deterministic perturbations and stochastic processes within a quality control scheme

Birgit Eibl[1] and Reinhold Steinacker[1]

[1]University of Vienna, Department for Meteorology and Geophysics, Althanstraße 14, 1090 Vienna, Austria

Correspondence to: Birgit Eibl (birgit.eibl@univie.ac.at)

Abstract. Meteorological in situ observational data comes with a variety of errors and uncertainties. Any further usage of this data requires a sophisticated quality control to detect, quantify and possibly eliminate or at least to reduce errors and to increase the value of the information. It must be assumed, that each observational value Ψ_{obs} is contaminated by errors Ψ_{err} so that the true state Ψ_{true} is not known. Different kinds of errors can be identified. Each of them has different characteristics and therefore has to be detected through appropriate methods. For years, various methods as a self consistency test, clustering and nearest neighbour techniques have been implemented in the complex quality control scheme of the Vienna Enhanced Resolution Analysis (VERA). Thereby former elaborations adressed the elimination and treatment of gross errors. In successioon the present investigation adresses the determination of stochastic and deterministic perturbations. In a first step we implemented the method to split up the observational value to smooth out the stochastic errors to the best and retain deterministic perturbations thereafter. Through controlled experiments on two dimensions the performance and limitations of the complex quality control scheme has been investigated. The treatment of errors and signals on different scales and the limit of the usability of this property is the main focus of the presented investigation. We highly recommend to use the method for data quality control within a high resolution model analysing spatially distributed data in highly complex terrain.

1 Introduction

Meteorological observational in situ data comes with a variety of errors and uncertainties. Any further usage of this data requires a sophisticated quality control to detect, quantify and possibly eliminate or at least to reduce errors and to increase the value of the information. Different methodologies for detecting, handling and eliminating different kinds of data have been developed by for instance Gandin (1988) or Haiden et al. (2010). Phillips and Marks (1996), for example, suggest that for every model using spatial interpolation, should include an uncertainty map for the results, as every interpolation introduces an additional uncertainty to the original input values. Gandin (1988) suggests using a complex quality control able to treat all different kinds of errors individually as they occur. Working with or handling observational data within a data assimilation system of a weather analysis or prediction model requires individual quality control mechanisms according to different priorities.

The Vienna Enhanced Resolution Analysis (VERA) (Mayer and Steinacker, 2012) used in this work, is independet from any prognostic model or model first guess field and focuses on the statistical behaviour and the spatial and temporal consistency of observational data to detect and correct errors before the data is brought to a regular grid. Several models on the meso scale

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

work with data assimilation methods, where the detection of errors depends on the difference between observed and model data. One such model is, for example, the Integrated Nowcasting through Comprehensive Analysis (INCA) system (Haiden et al., 2011). High resolution prognostic models like those developed by the Consortium for Small Scale Modelling (COSMO) (Schraff and Hess, 2012) or the Weather Research and Forecasting Model (WRF) (Wang et al., 2008) use different assimilation schemes and first guess fields from global or lower resolution models to generate the initial condition. For the purpose of the presented investigation it is highly recommended to exclude errors that are introduced by model first guess fields.

The INCA system (Haiden et al., 2011) relies on the Numerical Weather Prediction (NWP) model output and uses operational in situ data and additionally high resolution remote sensing data, to interpolate between observations. Analyses of INCA are not used as an initial condition for NWP model integration (Haiden et al., 2010) but start instead with first guess fields coming from the ALADIN short-range forecast. This first guess is adjusted by the difference between observation and forecast at the observational locations. The data quality control of the high resolution COSMO-model is performed by the observation nudging technique for data assimilation before producing high resolution analyses. The initial conditions come from various coarse-grid driving models (GME, EZMWF, COSMO-Model) or from the continuous data assimilation stream (Schraff and Hess, 2012). As Haiden et al. (2011) pointed out, VERA is an analysis scheme which is independent from prognostic models, whereas INCA relies on NWP model output. The downscaling in VERA is carried out with the aid of high resolution patterns, generated by topography and land surface characteristic, INCA basically uses remotely sensed data and topographic information. Both systems aim to create analysis fields as close as possible to observational data (Haiden et al., 2011), but VERA is superior for model verification, because the model based background field of INCA does not allow an independent comparison between analysis and model. (Troupin et al., 2012) proposed a similar interpolation method like VERA, based on the minimization of a cost function and a finite element solver. As Phillips (1986) pointed out, the meteorological noise in the initial data can be reduced by adjusting amplitudes and phases of gravity modes, in this case, to values that are forced by non-linear interaction between Rossby modes. This was an early attempt to suppress meteorological noise in fields, generated by observational data.

For a better understanding of the methodology in section 2, a brief description of the error detection and eliminating procedure is being given. When it comes to gross error detection, and the overall complex quality control system, a detailed description can be found in (Mayer et al., 2012). This paper focusses on the further development of the complex quality control system within VERA and the possibility of the added value that can be gained when used in complex terrain. Different types of errors and noise are being separated and filtered to obtain a pure signal in meteorological fields. The described method within the VERA quality control scheme wants to preserve not only the synoptic scale, but also small scale orographically induced signals in the data. Therefore the accuracy of high resolution prognostic models could be enhanced if the quality controlled data is used within or as a part of a data assimilation process. We recommend to use the proposed scheme for the highly complex terrain and a high resolution model. But on the global scale and on flat terrain the 4D-VAR data assimilation scheme will be the better choice, as the model first guess field deliver a robust basis for data analysis on flat terrain. A positive impact on the

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

WRF performance in the Alpine region, when using VERA quality controlled data, has been observed recently (pers. comm. Mayer, 2016).

The main goal of this paper is to investigate the performance, uncertainty and limitation of the proposed complex quality control by carrying out controlled experiments on two dimensions over complex terrain. An expansion to three dimensions as is common for regional models like INCA, WRF or COSMO could easily be carried out. The more the (wanted) signal is preserved and the more (unwanted) noise is filtered out from the data, the better the performance of the quality control scheme is. Section 2 explains the methodology, used data and performed controlled case studies are presented in Section 3, followed by the presentation and discussion of the results in Section 4. Conclusions and outlook finalize the paper in Section 5.

2 Methodology

Before irregularly distributed data are interpolated to a regular grid, complex quality control should be performed, to eliminate or correct errors (Gandin, 1988). According to the methodology of Steinacker et al. (2000), Sperka and Steinacker (2011) and Mayer et al. (2012), it must be assumed that each observational value Ψ_{obs} is contaminated by errors Ψ_{err} so that the true state Ψ_{true} is not known.

$$\Psi_{obs} = \Psi_{true} + \Psi_{err} \tag{1}$$

As we normally only have observations available at discrete intervals, at stations at specific distances from each other, we can only derive those scales of the true field, which are much larger than the average station distance. We call this resolvable, generally smooth part of the field "synoptic" Ψ_{syn} and denote the unresolvable rest by the term "sub scale" Ψ_{sub}. Concerning sub scale patterns, a downscaling, which is performed in the VERA-system by the so called fingerprint technique, can be carried out, if access to additional information is available. Fingerprints Ψ_{fp} are high resolution - with regard to the station distance - fields, for example, from remote sensing platforms like radar for precipitation, satellite infrared radiometric information for temperatures, high resolution topographic or land type information for parameters, which are correlated to elevation or other topographic and land type features, etc. The strength c_{fp} of the fingerprint pattern has to be calibrated (weighted) by observations through statistical regression. The stronger the fingerprint pattern is present in the observational data, the higher the weighting factor c_{fp} is. Several different fingerprints may be offered to the system. Fingerprints have some similarities to EOFs, but are physically, rather than statistically, determined .

Sub scale signals can also be investigated by a multivariate approach. If Ψ for example is the precipitation rate, cases are commonly found, in which just one station reports precipitation in a larger area without any precipitation. Without any additional information it is impossible to decide, whtether the precipitation report is erroneous or if a local shower really has occurred just at the one station. If we consider the fact, that during a typical shower the temperature drops, humidity rises, wind speed increases, wind direction changes, pressure rises, etc., we can get a more robust estimate of whether the value represents a signal or just a random error, when we also consider the spatial structure of the other mentioned parameters. The difference

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions

to the fingerprint technique is that despite we can distinguish between signal and error or noise by the multivariate approach we cannot derive the scale of the phenomenon or the sub scale spatial pattern. Sub scale signals, uncovered by a multivariate approach are denoted by Ψ_{subsig}. The residual of Ψ_{sub}, which is neither detectable by the fingerprint nor by the multivariate approach is denoted by meteorological noise Ψ_{mn}. The error part of the observations Ψ_{obs}, which may be caused by a sensor calibration error, wrong reading, error introduced during transmission, coding or decoding, etc., can be split up into random errors Ψ_{re}, systematic errors (bias) Ψ_{se} and gross errors Ψ_{ge}. Hence it is possible to split up each observational value into a number of parts:

$$\Psi_{obs} = \Psi_{syn} + \sum_i \left(c_{fp}\Psi_{fp}\right)_i + \Psi_{subsig} + \Psi_{mn} + \Psi_{re} + \Psi_{se} + \Psi_{ge} \tag{2}$$

It should be noted here, that the scale separation between the large (synoptic) scale and the subscale depends on the station density. If the mean station distance is in the order of 100 km, the large scale basically covers extra tropical cyclones and anticyclones. If an observational micro-net with a 1 km station distance is available, even convective systems or urban heat islands may become "synoptic" features. Furthermore it is impossible to separate the meteorological noise and random errors which we therefore combine to $\Psi_{mn} + \Psi_{re} = \Psi_{noise}$. Hence an observational value can be split up into 6 separable parts:

$$\Psi_{obs} = \Psi_{syn} + \sum_i \left(c_{fp}\Psi_{fp}\right)_i + \Psi_{subsig} + \Psi_{noise} + \Psi_{se} + \Psi_{ge} \tag{3}$$

Normally, with the exception of Ψ_{ge}, the amplitude of Ψ_{syn} is larger than the amplitude of the other components of Ψ_{obs}. After the removal of gross errors, a low pass spectral, Gaussian, Laplacian or other adequate spatial filter will therefore create a field which is close to the synoptic component. The problem thereby is that such a filter will not only dampen random errors but also the meteorologically relevant smaller scale patterns. What we usually want are both the "clean" synoptic and the sub scale patterns as well. The difference between the observed value at a station and the filtered value $\Psi_{obs} - \Psi_{syn}$ ("deviation") represents the basis for the error detection and qualification scheme of VERA. The whole procedure to separate the terms of 3 has to be carried out iteratively:

- *Iteration I (gross error detection)*: Many gross errors can be detected, when the deviation exceeds certain physical or statistical limits. VERA uses the following criterion: If the deviation at one station exceeds physical limits or (for normally distributed variables) the three-fold long term interquartile range of the same station, the observation is treated as a gross error. To avoid the impact of gross errors on the spatial analysis in the next iteration, observations characterized as gross errors are omitted in the further analysis.

$$\Psi_{obs} = \Psi_{syn} + \sum_i \left(c_{fp}\Psi_{fp}\right)_i + \Psi_{subsig} + \Psi_{noise} + \Psi_{se} \tag{4}$$

- *Iteration II (systematic error /bias correction)*: If the temporal mean value of the deviations over a long time (e.g. a month) at a station is different from zero, such a mean deviation is characterized as a bias. In the next iteration the data set of observations is corrected with regard to the detected biases.

$$\Psi_{obs} - \Psi_{se} = \Psi_{syn} + \sum_i \left(c_{fp}\Psi_{fp}\right)_i + \Psi_{subsig} + \Psi_{noise} \tag{5}$$

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

– *Iteration III (finger print elimination)*: To be able to detect deterministic small scale patterns in the field, we need suitable fingerprints as mentioned above. We can offer the analysis system several possible fingerprints, for which the weights are determined by regressions. If a pattern is recognized in the data, the weight will be positive, if it is not recognized, the weight will be zero. A negative weight means that the inverse of a given pattern has been recognized. In addition subtracting the deterministic small scale components in the form of weighted fingerprints from the observed value equation (5) yields

$$\Psi_{obs} - \Psi_{se} - \sum_i (c_{fp}\Psi_{fp})_i = \Psi_{syn} + \Psi_{subsig} + \Psi_{noise} \tag{6}$$

– *Iteration IV (multivariate small scale signal elimination)*: If single subscale signals, found by a multivariate approach in a scale are kept, the corresponding deviations from the left side of equation (6) can be subtracted to obtain:

$$\Psi_{obs} - \Psi_{se} - \sum_i (c_{fp}\Psi_{fp})_i - \Psi_{subsig} = \Psi_{syn} + \Psi_{noise} \tag{7}$$

Alternatively if it is desired that these sub scale signals are filtered, Ψ_{subsig} can be left on the right hand side as part of Ψ_{noise}:

$$\Psi_{obs} - \Psi_{se} - \sum_i (c_{fp}\Psi_{fp})_i = \Psi_{syn} + \Psi_{noise} \tag{8}$$

– *Iteration V (random error elimination)*: Now the noise can be eliminated from the field by applying a suitable filter. VERA takes an overlapping spatial Laplace filter (Mayer et al., 2012) to quantify the deviations, which are interpreted as random errors. By subtracting the latter from the left hand side of equation (7) or equation (8) the "clean" deterministic large scale (synoptic) part of the observation can finally be obtained.

$$\Psi_{syn} = \Psi_{obs} - \Psi_{se} - \sum_i (c_{fp}\Psi_{fp})_i - \Psi_{subsig} - \Psi_{noise} \tag{9}$$

or

$$\Psi_{syn} = \Psi_{obs} - \Psi_{se} - \sum_i (c_{fp}\Psi_{fp})_i - \Psi_{noise} \tag{10}$$

The field of the quality checked and corrected "clean" synoptic and the deterministic subscale patterns can be recombined in the corresponding parts:

$$\Psi_{syn} + \sum_i (c_{fp}\Psi_{fp})_i + \Psi_{subsig} = \Psi_{obs} - \Psi_{se} - \Psi_{noise} \tag{11}$$

or

$$\Psi_{syn} + \sum_i (c_{fp}\Psi_{fp})_i = \Psi_{obs} - \Psi_{se} - \Psi_{noise} \tag{12}$$

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions

For a simple one dimensional example and for a data set without gross errors and biases the result of the filter process is shown in Fig. 1. As one can easily recognize, the filter response strongly depends on the scale and the amplitude of the synoptic pattern, and the amplitude of the noise (signal to noise range) with regard to the station distance. The VERA scheme published by (Steinacker et al., 2011) and (Mayer et al., 2012) executes the whole quality control package before calculating the spatial analysis fields. The presented quality control scheme within the analysis process is shown in (Fig.2) and allows small scale deterministic signals in meteorological fields to be conserved.

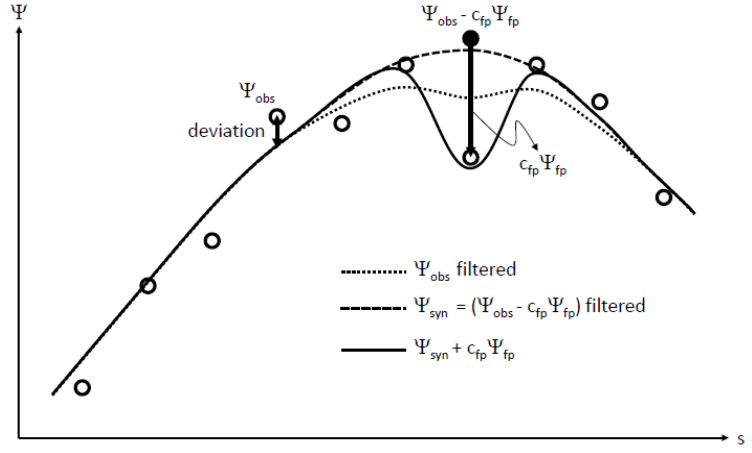

Figure 1. One dimensional example (observational along a space coordinate s) of the effect of filtering observational data with and without consideration of small scale patterns (fingerprints). When observational data are filtered directly (dotted curve), much of the deterministic small scale pattern is lost. When filtering the observed data without the fingerprint pattern (dashed curve), we damp only the noise. The sum of the filtered synoptic part and the fingerprint part (continuous curve) results in a pattern, where small scale deterministic features stay unfiltered despite the efficient noise filtering.

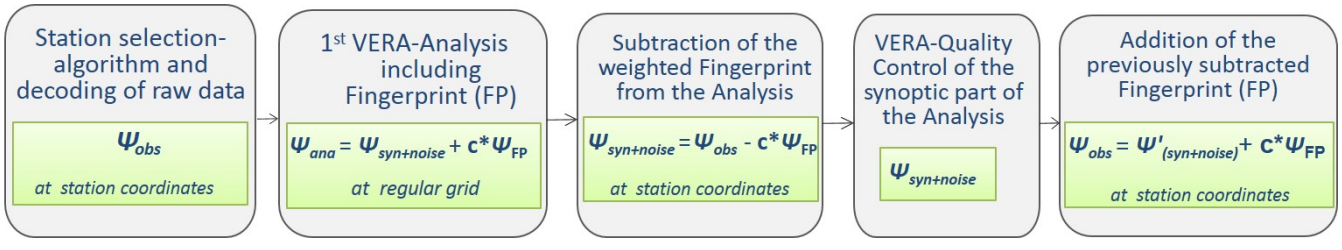

Figure 2. Process of the quality control scheme. Ψ_{obs} is the initial data at irregularly distributed observational station coordinates. Ψ_{ana} is the analysed value, where possible deterministic, physically explicable patterns (Ψ_{FP}) are extracted and weighted with the calculated factor c. $\Psi_{syn+noise}$ (large scale signal and meteorological noise) is the part of the analysed initial data that is unexplained by deterministic, physically explicable patterns. $\Psi'_{syn+noise}$ is the quality controlled part of the initial data.

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

3 Data

The performance of the presented quality control scheme cannot seriously be verified when solely error afflicted operational in situ data sets are used. For verification purposes the generation of data is proposed. The presented data processing makes it possibele to calculate the exact signal to noise ratio and therefore the exact mean and standard deviation of the desired atmospheric information and the noisy part of data. If not generated, the statistical terms of the components described in equation (2) are not known a priori. To prove the technical accuracy of the method and outline a sharp control it is indispensable to generate the different components of an observational value seperately and then analyse them. Therefore control experiments have been performed, where the set of non-dimensional components in equation (2) were generated. Data sets without any gross errors and biases were assumed, because the gross error detection and bias correction procedure is described in detail and extensively tested in (Mayer et al., 2012). For simplicity reasons we just take one fingerprint pattern (Ψ_{FP}). Anexemplary presentation is shown in (Fig.3) . Furthermore subscale signals were not separated from random errors and hence it is possible to stick with the formulations of equations 8, 10 and 12. Then equation (2) reduces to

$$\Psi_{obs} = \Psi_{syn} + c_{fp}\Psi_{fp} + \Psi_{noise} \tag{13}$$

The synoptic part of the field is analytically generated by a two dimensional, smooth, chess pattern wave system

$$\Psi_{syn} = A * (\sin(\mu_x x + \mu_y y)) + A * (\sin(\mu_x x - \mu_y y)) \tag{14}$$

The amplitude A of the wave pattern is set arbitrarily to 1 and the wave numbers μ_x and μ_y vary for the different experimental settings between 0.005 km^{-1} for large scale waves and 0.04 km^{-1} for meso-β scale waves, which corresponds to wave lengths λ_x and λ_y of approximately 1250 km and 150 km respectively. For the fingerprint pattern the thermal fingerprint (Steinacker et al., 2006) and (Bica et al., 2006) has been chosen, which indicates the different heating/cooling pattern induced by lowlands, mountains and water bodies (Fig. 3). In the setting for the discussed examination, the dimensionless values of Ψ_{fp} vary between 0 and 1. The weight c_{fp} of the thermal fingerprint varies for the experimental settings between 1 and 5. The noise part of the field has been produced by a random generator leading to spatially uncorrelated Gaussian distributed numbers with a mean of 0 and a standard deviation between 0.2 and 2 and represents the roughest part of the field. Due to the variable settings of the wave length of the synoptic part, the amplitude of the fingerprint part and the amplitude of the noise part with regard to the amplitude of the synoptic part (signal to noise ratio) we can investigate, how well and effective the suggested quality control procedure can filter and eliminate the noise and retain the synoptic and fingerprint parts of the field and if or under what conditions there are limits of its applicability.

3.1 Test Domain

In Fig.4 the location of 1311 observational stations within the European domain is shown.
The test domain encompasses a large part of Europe and North Africa and is shown in Fig.4. The station location has been taken from the an available set of surface weather stations on a particular day. The density of observation sites is high in Central

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

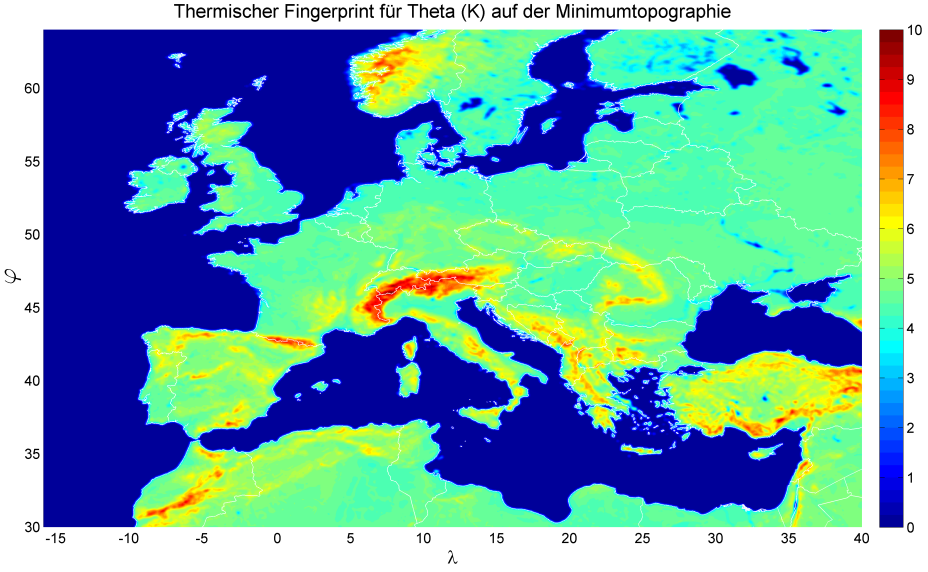

Figure 3. Thermal fingerprint of Europe as used in the VERA downscaling procedure. Contour lines are dimensionless and range between 0 and 10 in the operational setting.

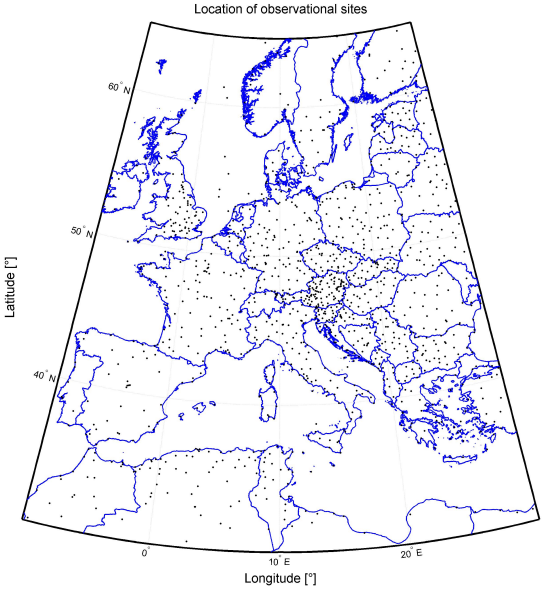

Figure 4. Distribution of 1311 observational stations within the European section.

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

Europe, whereas in Scandinavia, on the Iberian peninsula and especially over the oceanic areas it is much lower. The mean distance between two adjacent stations in the whole domain is close to 90 km. In Central Europe it is around 30 km and in the data sparse maritime areas several hundred km.

3.2 Case Studies

For the evaluation of the performance of the filtering of the noisy part of the data, various case studies with different settings of parameters were performed. The settings of these case studies are listed in (Tab.1) and the associated statistics in (Tab.2). The designation of the case studies consists of the three parts that build the generated data value, characterized by different capital letters W, N and FP. W stands for the wavenumber, N for the noisy part and FP for the "fingerprint". The numbers directly following the capital letters indicate the weight (for FP) or the standard deviation (for the noise) or the applied wavenumer (for W). Within the quality control scheme the Bias correction and gross error correction was switched off. These parts have been extensively tested in previous elaborations (Steinacker et al., 2011) and (Mayer et al., 2012).

Table 1. *Conditions and characteristics of initial data components (Ψ_{syn}, Ψ_{noise}, Ψ_{fp}) for various case studies. The designation of the case studies consists of the capital letters W, N, FP; following numbers indicate either the applied weight, standard deviation or wavenumber. μ_x, μ_y = wavenumber, STD=standard deviation of randomly distributed data (mean=0), c_{fp}=weighting factor.*

Case study	Ψ_{syn} μ_x, μ_y	Ψ_{noise} STD	$c_{fp}\Psi_{fp}$ c_{fp}
W001N02FP1	0.0015 km^{-1}	0.2	1
W001N1FP1	0.0015 km^{-1}	1	1
W001N5FP1	0.0015 km^{-1}	5	1
W001N5FP5	0.0015 km^{-1}	5	5
W001N1FP5	0.0015 km^{-1}	1	5
W005N1FP1	0.005 km^{-1}	1	1
W005N02FP1	0.005 km^{-1}	0.2	1
W005N5FP5	0.005 km^{-1}	5	5

3.3 Statistics

For a robust interpretation and evaluation of the filter and its performance and limits, statistical analyses were performed. Formulas from (Wilks, 2006).

- Noise ratio (NR)

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

- Regression
- Correlation Coefficient (CC)
- Spectral analysis

As there are several signals involved in a single initial data set, the calculation of a noise ratio (NR) is crucial. $NR = \frac{STD_{\Psi_{noise}}}{STD_{\Psi_{(syn+noise)}}}$ where $STD_{\Psi_{(syn+noise)}}$ is the standard deviation of the input signal before the application of the quality control and $STD_{\Psi_{noise}}$ the standard deviation of the noisy part of the initial signal. For calculating the ratio with quality controlled data, the $STD_{\Psi'_{(syn+noise)}}$ which is the standard deviation of the output signal after the initial data was quality controlled can be applied in the formula. Therefore the NR could be described as the power of the noise devided by the power of the signal Kieser et al. (2005).

The correlation coefficient (CC) indicates, how well two series of data fit together. The squared CC gives the fraction of the variance, which is statistically explained by the regression $CC^2 = \frac{\sum_{j=1}^{n}[y_j-\bar{y}][\hat{y}(x_j)-\bar{y}]}{\sqrt{\sum_{j=1}^{n}[y_j-\bar{y}]^2 \sum_{j=1}^{n}[\hat{y}(x_j)-\bar{y}]^2}}$ where y_j are the observed values, $\bar{y}$ their mean value and $\hat{y}(x_j)$ the predicted values by the regression (Wilks, 2006). The correlation coefficient between the initial data $\Psi_{syn+noise}$ and the quality controlled data $\Psi'_{(syn+noise)}$ is shown in Tab.3 in column CC. The correlation between the $\Psi_{syn+noise}$ and the Ψ_{syn} part within the same case study and $\Psi'_{(syn+noise)}$ with Ψ_{syn} of the same case study is depicted in Tab.2 (column $C1$) respectively in Tab.3 in column $C2$.

For the spectral analysis a fast Fourier transformation (fft) was performed. The purpose is to visualize the different wavelengths and energy spectra of the initial and quality controlled signal. In section 4 the performance is discussed and the spectra depicted.

Table 2. *Statistical information for parts of the initial data (Ψ_{syn} and Ψ_{noise}) used in case studies before the quality control was applied. NR is the noise ratio between $\Psi_{syn+noise}$ and Ψ_{noise}. $C1$ is the correlation coefficient between $\Psi_{syn+noise}$ and Ψ_{syn}.*

Statistics	MEAN	STD	MEAN	STD	MEAN	STD	NR	C1
Case study	Ψ_{syn}	Ψ_{syn}	Ψ_{noise}	Ψ_{noise}	$\Psi_{syn+noise}$	$\Psi_{syn+noise}$		
W001N1FP1	0.02	0.73	0.03	1.01	0.33	1.24	0.79	0.60
W001N02FP1	0.02	0.73	0.01	0.22	0.40	0.76	0.26	0.96
W001N5FP1	0.01	0.73	0.04	4.77	-0.13	4.77	0.99	0.17
W001N1FP5	0.01	0.73	0.03	0.98	0.57	1.30	0.75	0.63
W001N5FP5	0.01	0.73	0.01	4.72	-1.76	4.95	0.95	0.25
W005N1FP1	0.00	1.01	0.03	0.98	0.32	1.38	0.71	0.70
W005N02FP1	0.00	1.01	0.01	0.19	0.24	2.16	0.09	0.43
W005N5FP5	0.01	1.00	0.04	4.72	-1.72	4.90	0.96	0.18

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

For the statistical evaluation the noise ratio (NR), the standard deviation (STD) and the correlation coefficients (CC, $C1$ and $C2$) were calculated for the original (Tab.2) and quality the controlled data (Tab.3).

4 Results

4.1 Performance

Table 3. *Performance of the quality control system. Correlation coefficient (CC), noise ratio (NR), MEAN and standard deviation (STD) of* $\Psi'_{(syn+noise)}$ *data after the application of the quality control is listed.*

Performance case studies	MEAN $\Psi'_{(syn+noise)}$	STD $\Psi'_{(syn+noise)}$	CC	NR	C2
W001N1FP1	0.31	0.91	0.64	1.08	0.83
W001N02FP1	0.39	0.74	0.97	0.27	0.98
W001N5FP1	-0.15	2.57	0.36	1.83	0.35
W001N1FP5	0.55	0.99	0.68	0.98	0.85
W001N5FP5	-1.78	2.83	0.42	1.67	0.46
W005N1FP1	0.31	1.04	0.72	0.94	0.86
W005N02FP1	0.22	1.31	0.50	0.15	0.65
W005N5FP5	-1.73	2.73	0.39	1.73	0.31

Comparing the performed statistics before (Tab. 2) and after (Tab. 3) the application of the quality control on $\Psi_{(syn+noise)}$, a significant improvement is apparent from the lower STD of quality controlled data shown in Tab. 3. To get an idea of how the quality control is effecting the different signals originating from different scales a Fast Fourier Transformation (fft) was performed. For this purpose the initial data $\Psi_{(syn+noise)}$ and the quality controlled data $\Psi'_{(syn+noise)}$ were detrended and a window function was applied. For the spectral analysis only data after the subtraction of the $c_{fp}\Psi_{fp}$ part was used and is presented in the log-log graphs in Fig. 5. Since the observational data and therefore the quality controlled data is a mixture of different signals characterized by different wavelengths, a fft provides an insightful analysis. After the quality control the signals are no longer properly separable, but the fft gives an idea of the effect the quality control has on the initial data. The graphs in Fig. 5 show the spectrum of wavelengths from longer wavelengths on the left to shorter wavelengths and their dissipation at the right end of the scale. With high energetic large scale vortices on the left end of the scale and the small eddies, noise and dissipation at the right end. With the preservation of large vortices and the reduction of smaller scale eddies one can say that the performance of the quality control scheme is as anticipated (Stull, 2009).

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

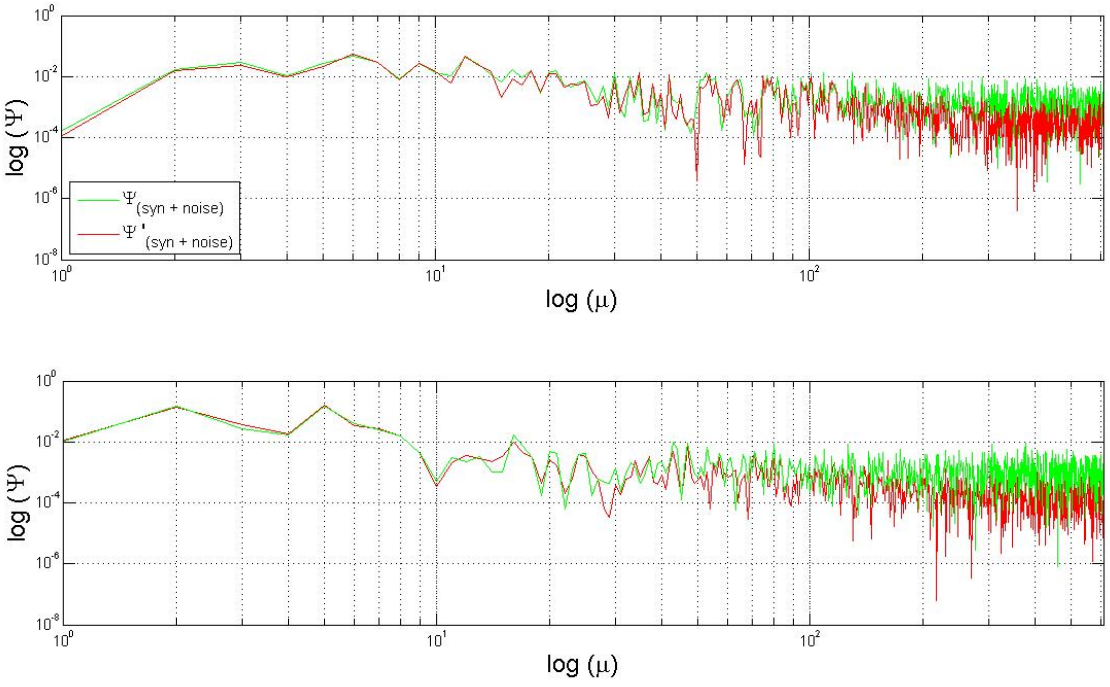

Figure 5. *Spectral analysis performed with a fast Fourier transformation (fft) with the initial data $\Psi_{(syn+noise)}$ (green line) before the quality control. Red line represents the data $\Psi'_{(syn+noise)}$ after the application of the quality control. The figure at the top shows the case study W005N1FP1 whereas at the bottom case study W001N1FP5 is depicted. In both case studies the noise input is exactly the same whereas the Ψ_{syn} part is of shorter wavelength in the case study on the left.*

4.2 Limits of the filter

For different simulated atmospheric conditions the expected performance of the filter shows its limits. In table 1 the different conditions of the performed case studies are listed. In case study W001N1FP1 with a long wavelength in the Ψ_{syn} part of the signal and the standard deviation of the Ψ_{noise} around 1, the NR is significantly higher after the approach of the quality control scheme. Whereas the NR has barely improved in case study W001N02FP1, with the same data for Ψ_{syn} but a standard deviation for the meteorological noise Ψ_{noise} of approximately 0.2. For a Ψ_{noise} with $STD = 5$ the NR shows significantly different ratios in all cases. In Fig. 6 the values for different parts of initial data is plotted in order of the magnitude of Ψ_{syn} data. In the formula for the Ψ_{syn} signal (Eq.14) A is set to 1 for all case studies. Therefore the maximum amplitude should be located around $+2$ respectively -2, depending on the added noisy part Ψ_{noise}. Obviously visible is the damping of the noisy part of the initial data (green) due to the application of the filter (quality control). The fluctuations of the quality controlled data (red) are of smaller amplitude than before the filter treatment. Another impact of the filter treatment not shown here is an additional damping of the Ψ_{syn} which is often not requested and only appearing if both parts of the initial data are within a relative similar range of wavenumbers. This happens to a greater extent the smaller the difference between the wavelength of

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions

Ψ_{syn} and Ψ_{noise} gets. The latter impact of the filter is not likely to appear in real meteorological conditions where the synoptic scale signal and the meteorological noise is explicitly differentiable.

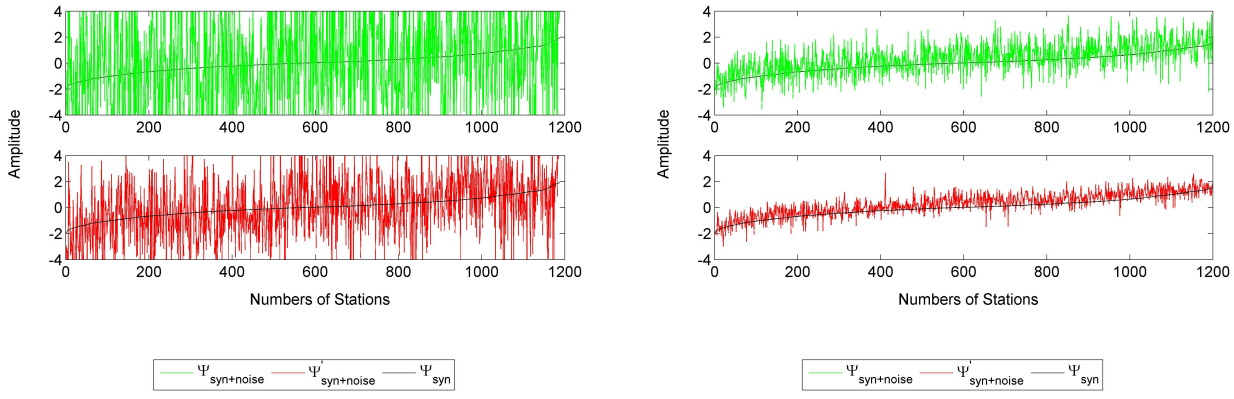

Figure 6. *Distribution of Ψ_{syn} (black solid line) at 1250 observational station coordinates ordered by the magnitude of Ψ_{syn}. In green the initial data $\Psi_{(syn+noise)}$ before the filter performance test, red the data $\Psi\prime_{(syn+noise)}$ after the application of the quality control. The wavelength for Ψ_{syn} is 3600 km, the standard deviation for the Ψ_{noise} part is two in the figure on the right and 1 in the left chart. Note the different scaling of both charts.*

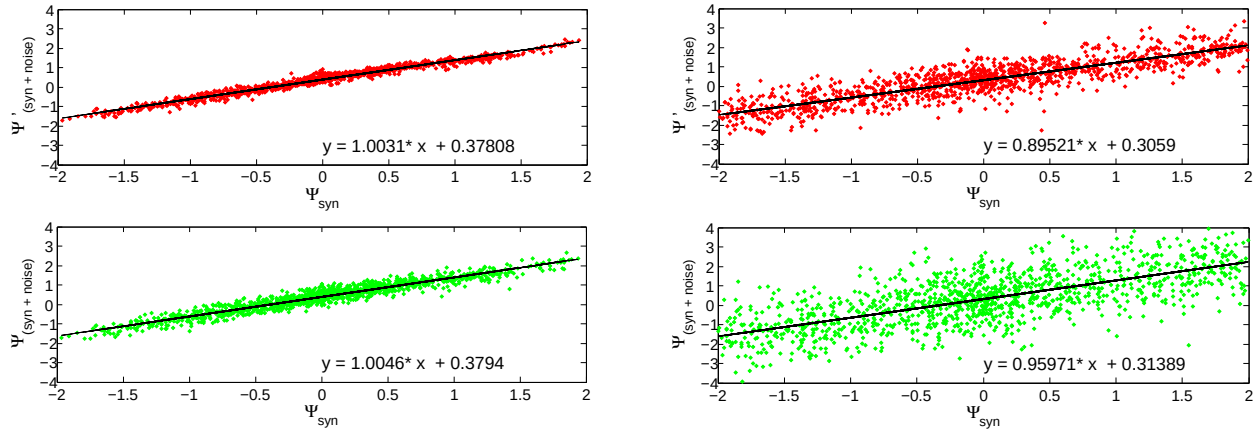

Figure 7. *Scatter plot showing the correlation of Ψ_{syn} with the initial data $\Psi_{(syn+noise)}$ (green diamonds) before the filter performance test, red diamonds represent the data $\Psi'_{(syn+noise)}$ after the application of the quality control. The figure on the right shows a correlation coefficient (CC) of 0.6 for the initial data 0.8 for quality controlled data (red). CC for the initial data in the left chart is 0.95 and after the performed quality control 0.98.*

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

The two case studies shown in Fig. 7 show significant improvement with respect to the reduction of deviation and data variability. In the chart on the left with the initial data composition of a wave number μ = 0.0015 in the Ψ_{syn} part and a STD = 0.2 for Ψ_{noise}, the $C1$ could be enhanced from 0.95 to $C2$ with 0.98 for the quality controlled data. The case study on the left with μ = 0.005 in Ψ_{syn} and STD = 1 for Ψ_{noise} had a $C1$ of 0.6 for the initial data which increased to a value of 0.8 for $C2$, the quality controlled data. As depicted in Tab.2 and Tab.3 correlations between the Ψ_{syn} and the $\Psi_{(syn+noise)}$ respectively $\Psi'_{(syn+noise)}$ could be enhanced significantly, which was somehow the aim of changing the routine of the quality control scheme.

5 Conclusions

A sophisticated data quality control forms the basis for a comprehensive analysis and subsequent use of measured data for data assimilation and forecasting purposes. The presented step within a continuously and long-lasting development process of a complex quality control system describes only a small part of the comprehensive and extensive field dealing with broad variety of errors, their detection and correction. The overall target of different quality control systems is to preserve and represent the current state of the atmosphere which is the closest to the truth someone can get.

Overall the performance of the quality control scheme is able to reduce the noisy part of an initial data set even if the variation is small. The more the wavenumber of the Ψ_{syn} part distinguishes from the Ψ_{noise} part of initial data fields, the more significant the filtering of the erroneous part of data will be. If the noisy, erroneous data and the "fingerprint" pattern are of the same scale, the subtraction of the "fingerprint" Ψ_{fp} from the observational value Ψ_{obs} would not be satisfying, as the subtraction would be vague and not sharp enough for preserving phenomena. Subsequently this quality control scheme would not yield best performances within the latter conditions. Considering real conditions within a complex terrain, a so called synoptic signal and the terrain induced modification will be of different scales and therefore the quality control system is able to manage the separation of the different signals. Even the meteorological noise is generally appearing on a different scale than the terrain induced signal.

Since the present composition is based on generated data a comprehensive evaluation using observational data would be the obvious next step. Further a detailed performance analysis within different environments in complex terrain will be carried out. The main focus will lie on the applicability of the presented complex data quality control system to an area with dense observational data availability on the one hand and on the other hand to determine the opposite limit for useful analysis in data sparse areas. As in the present paper further investigations and analysis will be executed in highly complex terrain environments. The usability of open access observational data from partly private weather stations should be addressed by a data quality control scheme. The analysis of different parameters requires the development of different "fingerprints" and/or the usage of their combination to identify various meteorological phenomena. For this purpose an area in the Tropics with highly irregularly distributed in-situ observations within a diurnal climate is selected to evaluate the possibility of the presented methodology in the given environment. Additionally the benefit for an analysis by adding small areas where data is collected within a denser observation network should be determined.

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Now that the limits of the filter (high resolution analysis, significant differnece between the signals) are known, real data can be analysed with this method. The difference here is, that the exact noise ratio of real data is not known, but it is reasonable to assume that it is higher at situations with significant and strong synoptic gradients and therefore coherent atmospheric conditions in contrast to situtations where the gradient is weaker and therefore the signal to noise ratio is very low. A compariston with real observational data is the reasonable next step. For best possible outcome, the same location as shown in Fig. (4.) will be used. The temperature and pressure data will be analysed and the selected case studies should fit the framework of the generated data. Coherent atmospheric conditions like gradient intensive synoptic patterns will be selected. Further evaluation will examine the performance of the presented method on different dense observational networks. It is expected that a denser network does not bring significant information to the performed analysis but the investigation will point out future perspectives.

References

Bica, B., Knabl, T., Steinacker, R., Ratheiser, M., Dorninger, M., Lotteraner, C., Schneider, S., Chimani, B., Gepp, W., and Tschannett, S.: Thermally and Dynamically Induced Pressure Features over Complex Terrain from High-Resolution Analyses, Journal of Applied Meteorology and Climatology, 46, 50–65, 2006.

Gandin, L. S.: Complex Quality Control of Meteorological Observations, Monthly Weather Review, 116, 1137–1156, doi:10.1175/1520-0493(1988)116<1137:CQCOMO>2.0.CO;2, 1988.

Haiden, T., Kann, A., Pistotnik, G., Stadlbacher, K., and Wittmann, C.: Integrated Nowcasting through Comprehensive Analysis (INCA) System description, 2010.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region, Weather and Forecasting, 26, 166–183, doi:10.1175/2010WAF2222451.1, 2011.

Kieser, R., Reynisson, P., and Mulligan, T. J.: Definition of signal-to-noise ratio and its critical role in split-beam measurements, ICES Journal of Marine Science, 62, 123–130, doi:10.1016/j.icejms.2004.09.006, 2005.

Mayer, D. and Steinacker, R.: VERA – Ein objektives Verfahren zur Analyse von meteorologischen Messwerten VERA – an objective method to analyse meteorological observations, Mitteilungsblatt des Hydrografischen Dienstes, 88, 9–34, 2012.

Mayer, D., Steiner, A., and Steinacker, R.: Innovations and applications of the VERA quality control, Geoscientific Instrumentation Methods and Data Systems, 1, 135–149, doi:10.5194/gi-1-135-2012, 2012.

Phillips, D. L. and Marks, D. G.: Spatial uncertainty analysis: Propagation of interpolation errors in spatially distributed models, Ecological Modelling, 91, 213–229, doi:10.1016/0304-3800(95)00191-3, 1996.

Phillips, N.: The spatial statistics of random geostrophic modes and first-guess errors, Tellus A, pp. 314–332, doi:10.3402/tellusa.v38i4.11721, 1986.

Schraff, C. and Hess, R.: A Description of the Nonhydrostatic Regional COSMO-Model Part III : Data Assimilation, Www.Cosmo-Model.Org, p. 93, 2012.

Sperka, S. and Steinacker, R.: A quality-control and bias-correction method developed for irregularly spaced time series of observational pressure data, Journal of Atmospheric and Oceanic Technology, 28, 1317–1323, doi:10.1175/JTECH-D-10-05046.1, 2011.

Steinacker, R., Häberli, C., and Pöttschacher, W.: A Transparent Method for the Analysis and Quality Evaluation of Irregularly Distributed and Noisy Observational Data, Monthly Weather Review, 128, 2303–2316, 2000.

Steinacker, R., Ratheiser, M., Bica, B., Chimani, B., Dorninger, M., Gepp, W., Lotteraner, C., Schneider, S., and Tschannett, S.: A Mesoscale Data Analysis and Downscaling Method over Complex Terrain, Monthly Weather Review, 134, 2758–2771, 2006.

Steinacker, R., Mayer, D., and Steiner, A.: Data Quality Control Based on Self-Consistency, Monthly Weather Review, 139, 3974–3991, doi:10.1175/MWR-D-10-05024.1, 2011.

Stull, R. B.: An Introduction to Boundary Layer Meteorology, Springer, 13 edn., 2009.

Troupin, C., Barth, A., Sirjacobs, D., Ouberdous, M., Brankart, J. M., Brasseur, P., Rixen, M., Alvera-Azcárate, A., Belounis, M., Capet, A., Lenartz, F., Toussaint, M. E., and Beckers, J. M.: Generation of analysis and consistent error fields using the Data Interpolating Variational Analysis (DIVA), Ocean Modelling, 52-53, 90–101, doi:10.1016/j.ocemod.2012.05.002, http://dx.doi.org/10.1016/j.ocemod.2012.05.002, 2012.

Geosci. Instrum. Method. Data Syst. Discuss., https://doi.org/10.5194/gi-2017-42
Manuscript under review for journal Geosci. Instrum. Method. Data Syst.
Discussion started: 21 December 2017

Geoscientific Instrumentation Methods and Data Systems Discussions
Open Access
EGU

Wang, X., Barker, D. M., Snyder, C., and Hamill, T. M.: A Hybrid ETKF–3DVAR Data Assimilation Scheme for the WRF Model. Part II: Real Observation Experiments, Monthly Weather Review, 136, 5132–5147, doi:10.1175/2008MWR2445.1, 2008.

Wilks, D.: Statistical Methods in the Atmospheric Sciences, Elsevier Academic Press, 2 edn., 2006.