AUTOMATED QUALITY EVALUATION FOR A MORE EFFECTIVE DATA PEER REVIEW

A Düsterhus[1,2]* and A Hense[2]

*[*1] National Oceanography Centre, Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, United Kingdom*
Email: andhus@noc.ac.uk
[2] Meteorological Institute, University of Bonn, Auf dem Hügel 20, 53121 Bonn, Germany
Email: ahense@uni-bonn.de

ABSTRACT

A peer review scheme comparable to that used in traditional scientific journals is a major element missing in bringing publications of raw data up to standards equivalent to those of traditional publications. This paper introduces a quality evaluation process designed to analyse the technical quality as well as the content of a dataset. This process is based on quality tests, the results of which are evaluated with the help of the knowledge of an expert. As a result, the quality is estimated by a single value only. Further, the paper includes an application and a critical discussion on the potential for success, the possible introduction of the process into data centres, and practical implications of the scheme.

Keywords: Data peer review, Data publication, Quality evaluation, Statistical quality assurance, Meteorological data

1 INTRODUCTION

Today publications reporting scientific work constitute a major contribution to the quality assessment of the involved research parties: individuals, projects, and organisations. This is definitely true for the publication of articles in established scientific journals, but other types of publications are also gaining importance. One example of this is the publication of data. Data are the basis of science and research and are therefore a special case. Their publication in trusted repositories is already an opportunity for scientists to enhance the credibility of their work (Costello, 2009; Piwowar & Vision, 2013; Henneken & Accomazzi, 2011). Additionally, data publication allows scientists to fulfil the requirements of funding agencies to make their data available (National Science Foundation, 2013; Research Councils UK, 2013; Deutsche Forschungsgemeinschaft, 2013). Integrating this type of publication into the regular scientific process can lead to more transparency and a simpler reproduction of scientific work (Klump et al., 2006). Nevertheless, up to now publication of raw data has not had the same standards of recognition as journal publications because it lacks a peer review process comparable to that of traditional paper publications. Such a process would make raw data publications comparable to journal and data papers and could lead to the establishment of a third string of scientific publications. (Quadt et al., 2012).

A need for a comprehensive peer review of data publications was stated by Parsons, Duerr, and Minister (2010) who urged the development of such schemes and their use at data centres. The topic was also addressed by the InterAcademy Council, which reviewed the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) and asked for guidelines on how non-peer-reviewed literature and observations could be used for the IPCC report (InterAcademy Council, 2010). The first steps for a formal data peer review were given by Lawrence et al. (2011). The authors described the possible requirements for different kinds of data publications and the different roles that have to be filled within these schemes. Nevertheless, methods to enable reviewers to perform effective and transparent reviews, which would approach those expected for paper publications, are not yet available.

This paper focuses on such methods and gives a roadmap for how datasets can be evaluated more effectively. We present the development of a statistical scheme to evaluate the quality of the content of datasets. This scheme is demonstrated on data content that can be described by real-valued numbers. Also there is a brief discussion about which requirements must be met to extend the scheme to other types of datasets. The quality evaluation scheme presented here has been designed to combine the knowledge of an expert with the results of

quality checks to assign a numerical quality estimate to a dataset. A description of quality evaluation is given in Section 2. An evaluation of the application of this method to a dataset from a meteorological weather station is presented in Section 3. Section 4 has a general introduction on how such schemes can be implemented in the data publication process at data centres and explains what a general peer review of a dataset looks like. This is illustrated by the procedures and workflows developed by Quadt et al. (2012). Section 6 discusses the scheme and the application in detail with a special focus on the problems of assigning a simple quality estimation to a dataset. The paper concludes in Section 7 with some remarks on how this scheme could also be used for unstructured datasets.

2 QUALITY EVALUATION

The quality evaluation scheme at hand is designed for observational datasets where the data content is a set of real-valued numbers, such as the time series of a physical measurement, field observations at different locations in space, or combinations of both. Likewise the scheme can be used for the output of numerical simulations or statistical analyses. To estimate the quality of a dataset, it is important to define what the quality of a dataset means. In the case of this scheme, quality is seen as a statement of how well the observations O represent the truth T. Due to the limitations of measurement instruments, observations are only able to approximate the truth (Gandin, 1988). Taking this into account, a measurement operator M_O is introduced that transforms the truth to the observational space. The quality of a dataset is given by the probability of a quantity Q, given the observations and the modified truth: $p(Q|O, M_O(T))$. Q stands for a "good quality dataset", which means in this case that it fulfils the expectations of being a reasonable representation of the truth T. In a very simplistic approach, one might say that Q represents the number of missing values in the observations.

The quality is estimated by quality checks and tests. In the following, it is assumed that a quality assurance test can be uniquely defined by a parameter set θ. Using this, the probability defining the quality is derived as follows:

$$p(Q|O, M_O(T)) = \int_{\theta} p(Q|\theta', O, M_O(T))\, p(\theta'|O, M_O(T))\, d\theta'. \quad (1)$$

Eq. (1) is derived from elementary probability theory involving the definition of joint probability $p(Q|O,M_O(T))$, which can be split with the help of the conditional probability $p(Q|\theta,O,M_O(T))$, the marginal probability $p(\theta|O,M_O(T))$, and the marginalisation (integration) over all possible parameter values θ.

In order to include the prior knowledge of an expert, for example a reviewer, the second term on the right hand side is expanded by the same method using variable I, which characterises the expert knowledge:

$$p(\theta|O, M_O(T)) = \int_{I} p(\theta|I', O, M_O(T))\, p(I'|O, M_O(T))\, dI'. \quad (2)$$

The number of tests as well as the number of experts whose knowledge is used has discrete values. Then the combination of Eq. (1) and (2) is written as discrete sums over probabilities:

$$p(Q|O, M_O(T)) = \sum_i p(Q|\theta_i, O, M_O(T)) \sum_j p(\theta_i|I_j, O, M_O(T))\, p(I_j|O, M_O(T)). \quad (3)$$

The right hand side of Eq. (3) starts with a term that estimates the probability of Q by using the test defining sets of the parameter. This measures the result of a quality assurance check measured in terms of probability. The second term uses the expert knowledge to assess the probability of the quality assurance test, described by the parameter set. It can be interpreted as a weighting of the test by the expert. The last term represents the prior knowledge of the expert.

With these three terms, which are fully defined by the knowledge of the expert and the chosen tests, based on the observations and the assumed truth, the quality estimate is assessed. A problem to be solved is an effective way of making a practical implementation of this scheme. It is of course possible for an expert to give his interpretation and the prior separately for every test performed. Nevertheless, from a practical point of view, this approach is not an effective solution. Therefore, the statistical quality evaluation scheme is implemented as shown in Figure 1.

An expert plays the key role in this scheme, by defining the parameters, priors, and weightings of each test. Because the priors define the tests as a whole, these decisions determine the kinds of tests as well. The prior translates each test result to a percentage of quality. This is necessary because not every test has a linear dependence between its result and the quality. To prevent the expert from having to define the priors for each result separately after the performance of the tests, s/he is instead asked to define a general prior beforehand. This general prior, called an adjusted prior, should deliver the corresponding quality statement for every possible outcome of a test. By combining this statement with the weighting and afterwards with the test results, the quality estimate is gained for one set of parameters for one test. When the weighting is used appropriately (i.e., normalised), it is of course possible to use several tests and parameter sets to estimate the quality of a dataset.

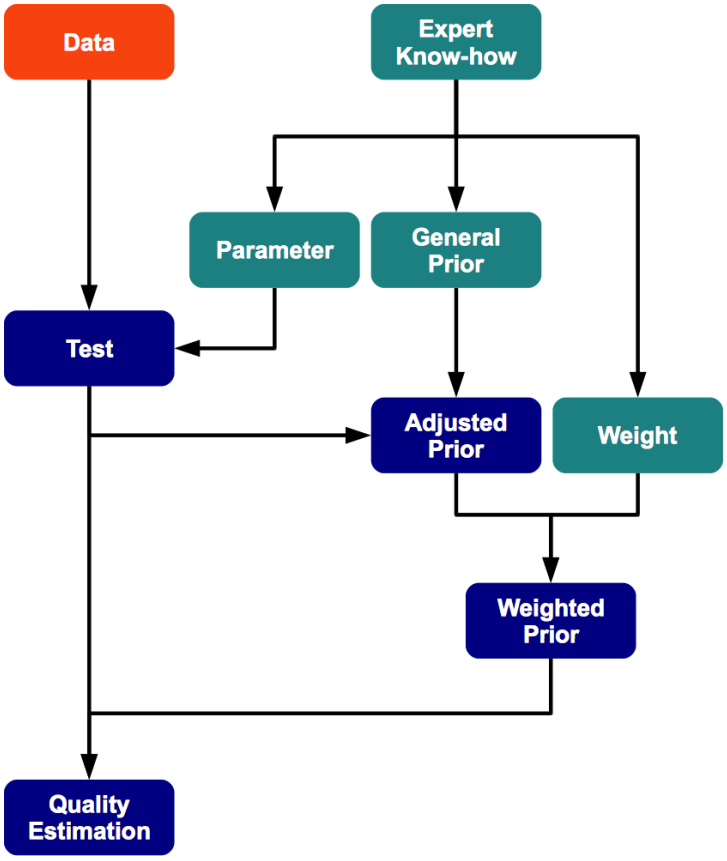

Figure 1. Structure of the implementation of the quality evaluation procedure. The elements influenced by the data only are in red, those influenced solely by the expert are in green, and the blue elements are influenced by both.

As mentioned above, there are two main prerequisites for the quality evaluation procedure. The first requires that a quality assurance test be completely defined by its set of parameters. This is simple to fulfil. The second prerequisite, that the tests used must deliver a probabilistic result, is at first glance not always possible to achieve. Nevertheless, approaches such as those described by Dose and Menzel (2004) might help to transform results of existing quality tests into probability statements.

3 APPLICATION

One advantage of the quality evaluation method is its ability to work as an effective analysis tool for larger datasets. To demonstrate this capability, the procedure is used to evaluate the not quality controlled datasets of 15 physical variables of the climatological weather station at the University of Hohenheim (Germany) during the years 2007 and 2008 (Wulfmeyer & Henning-Müller, 2006). The collection of data was part of the General Observing Period (GOP) and the field experiment Convective and Orographically-induced Precipitation Study

(COPS) of Priority Program 1176 of the Deutsche Forschungsgemeinschaft DFG in 2007 (Crewell et al., 2008; Hense & Wulfmeyer, 2008; Wulfmeyer et al., 2006). The analysed variables are found in Table 1. The dataset consists of measurements made over 421 days and has a temporal resolution of 30 seconds (between 01/01/2007 and 26/02/2008, the data for 03/04/2007 are not available).

Table 1. Measured variables at the climate station in Hohenheim

Abbreviation	Meteorological Parameter	Unit
TimeFromStart	Measurement Time	s
T200	Air Temperature 2m above ground	°C
RH200	Relative Humidity 2m above ground	%
T005	Air Temperature 5cm above soil	°C
WD1000	Wind from Direction 10m above ground	°
WS1000	Wind Speed 10m above ground	m/s
GR200	Global Radiation 2m above ground	W/m^2
RR200	Reflected short Radiation 2m above ground	W/m^2
NR200	Longwave Radiation Budget 2m above ground	W/m^2
RAIN	Rain-Collector 1m above ground	mm
ST002	Soil Temperature at 2cm under bare soil	°C
ST005	Soil Temperature at 5cm under bare soil	°C
ST010	Soil Temperature at 10cm under bare soil	°C
ST020	Soil Temperature at 20cm under bare soil	°C
ST050	Soil Temperature at 50cm under bare soil	°C

As a basis for the checks, the procedure described by Meek and Hatfield (1994) is used. It covers three tests types, which are defined as follows:

LIM checks for illegal or questionable values (check on limits)
ROC checks for questionable temporal evolutions (check on rate of change)
NOC checks for instrument "hang up" (check on no change)

The LIM check tests every value independently to see if it exceeds a given maximum or minimum limit. When this is the case, the value is marked. The same is done in the ROC test. The difference between two consecutive values is calculated and marked if it exceeds given limits. For purposes of demonstration, the NOC check is defined in this paper as a test on a change in consecutive values that is too small for a given time span. When this is the case, all values in the given time span are flagged. The configuration of the checks was gathered from WMO (2007) and used as far as applicable (the parameters used can be found in the appendix).

For all variables these settings consist of up to four checks. The first check is a LIM check for which the values of the dataset are strongly required to fulfil. Second and third checks are ROC checks, one with a strong and another with a weak requirement. The fourth check is a NOC check, which is given a weak requirement. The result of each check is a flag vector, which includes a flag for every suspect value under the given rule. Because the quality evaluation procedure requires a probabilistic result for each quality check under investigation, the flag vector has to be translated to such a result. This is done in this case by calculating the percentage of non-flagged observations relative to the total number available.

To take the different requirements for the checks into account, the general prior is defined accordingly. In each case it consists of a function, which for every possible outcome of the check (0 to 100%) defines the appropriate translated prior value. If all values are flagged, the prior has a value of 0% and if none is flagged of 100%. The two needed priors differentiate in the steepness of the slope from 100% of the prior value to 0%. For a strong prior, 0% will be reached at 99%; for a weak prior, 0% will be at 90% of the result. Both settings are rather strict but can be of course altered and might take a more complicated shape when necessary.

The last step is to define the weighting of each test in the quality estimation. For purposes of demonstration, the weighting of the LIM check receives a weighting of 2, and the other checks receive a 1. To generate quality estimates in the range of 0 and 1, the weighting is normalised for the analysis of each time series so that all weights sum up to 1. With this definition the settings are complete, and the checks can be executed and their results evaluated.

The quality estimation for each variable is summed up, and the mean over all variables for each day is shown in Figure 2. The vast majority of the days have a mean quality estimate of over 90%. Indeed the median of all of the days is 95.8%. Nevertheless, there are several days with a low quality estimation underlined by an interquartile range (IQR) of 6.0%. The scheme allows us to select the cases that have low quality estimation and take a look at them. It is possible to detect the variables with the largest problems and determine the causes. For this, instead of looking at the several thousand test results that have been used in this evaluation, we have to evaluate only a few to find the most severe problems. The results of the quality evaluation help to select the best time series for detecting these problems.

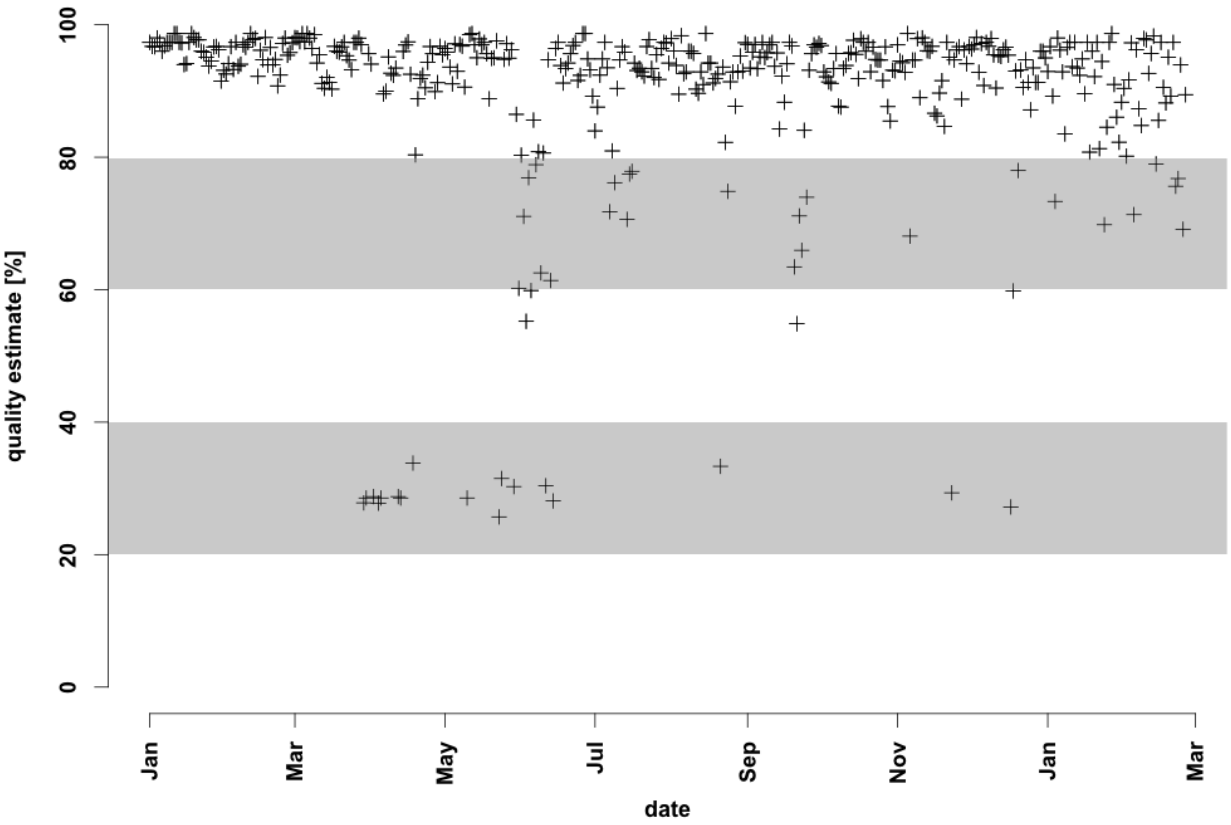

Figure 2. Mean quality estimate of all variables over time at the climatological weather station in Hohenheim between January 2007 and February 2008.

As an example, we chose the wind direction at 10m, a variable with one of the lowest medians of probability (83.7%, IQR 48.7%). The time series of the quality estimates for this variable is shown in Figure 3. It can be seen that several days have a low quality estimate for wind direction. To identify the reasons for this, a day with an average quality estimate is chosen, 18 February 2008. The overall estimate for that day is 57.3%. This is basically determined by the evaluation of two quality checks, a LIM and a NOC test. The probability of the test result is 99.9% for the LIM test, but taking the strict prior together with a high weighting, the quality measure is reduced to an overall value of 57.3%. The NOC test has a test result of 84.1%. This does not change the overall quality estimate for this variable, even with the application of the weak prior.

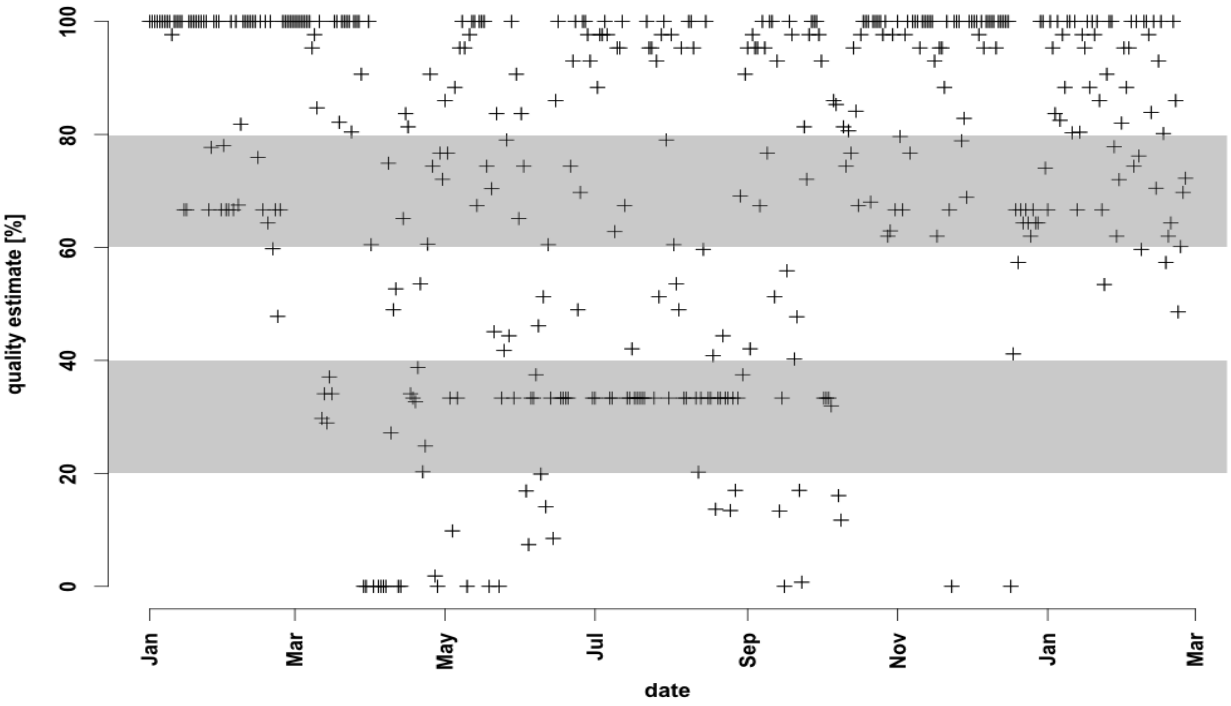

Figure 3. Quality estimate of the wind direction in 10m height over time at the climatological weather station in Hohenheim between January 2007 and February 2008.

The reasons for this behaviour can be found in a plot of the raw data of that day, which is shown in Figure 4. In the first half of the day, there are two longer sections where the measurements do not show enough variability. This behaviour is detected by the NOC test as suspicious. In the second half of the day, there are two values of 360.1°, which are above the limit of 360°. Additionally, there are two occurrences of a value that is several orders of magnitudes over the limit and can be regarded as a non-documented missing value identifier.

As in a peer review scheme for journals, the findings can now be presented to the data authors. To simulate this step, we presented our results and their explanations to the scientists at Hohenheim. Concerning the values, which are several magnitudes above the expected range of the wind direction, it was found that no missing or fill value parameter was set during the data conversion to the netCDF format. An explanation for the values that marginally exceed the threshold of 360 degrees is the specifics of the electronics of the instrument and its potential measurement uncertainties. Furthermore, it was recognised that the instrument has a threshold of 1 m/s of wind speed before it reacts to wind direction changes. It was acknowledged that at the location of the measurement station, the wind speed could be lower for a longer time span during specific weather situations. Within the discussions, it was also proposed by the data authors to change the testing parameters so that they better represent the specifics of the investigated instruments.

In this manner, the dataset can now be analysed step-by-step, and severe problems can be quickly identified, analysed, and documented. As a result the quality of the dataset is enhanced and more important information for the data re-user delivered.

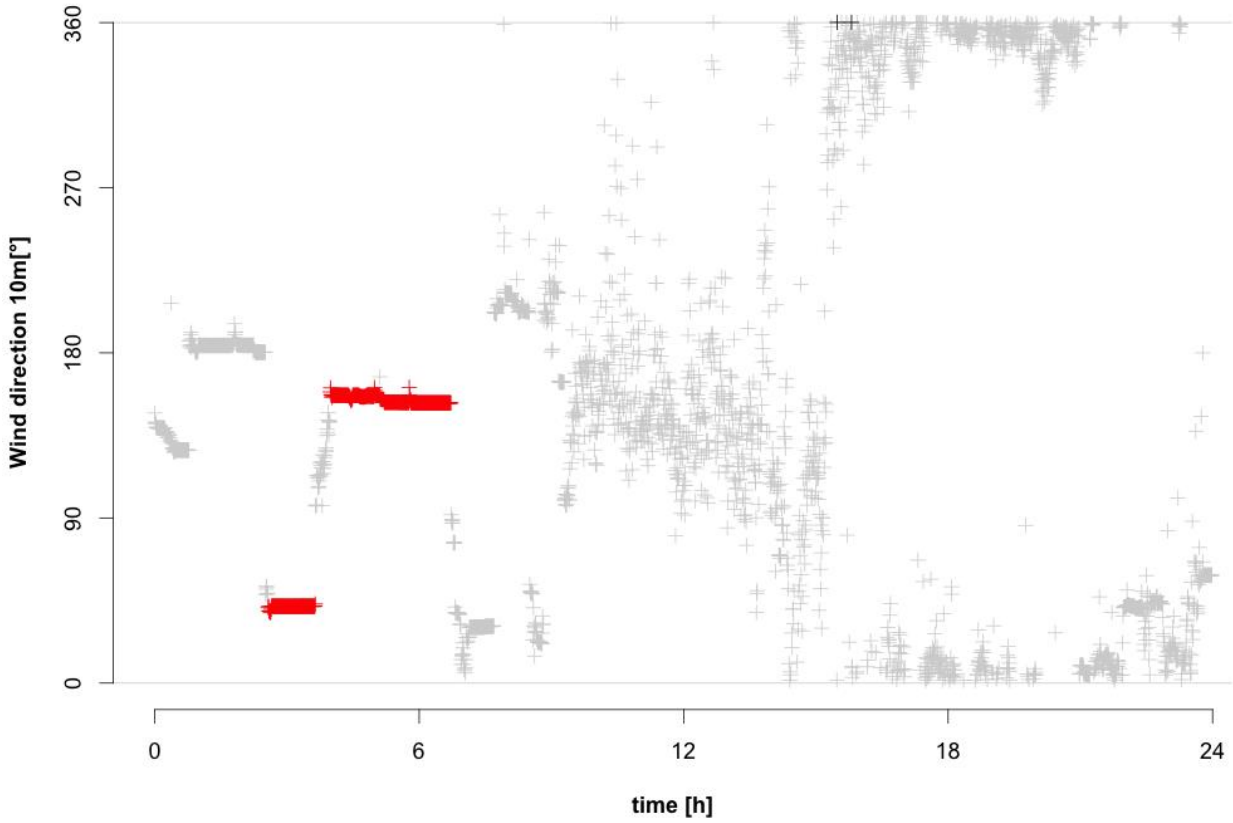

Figure 4. Time series of the wind direction at a height of 10m at the climatological weather station in Hohenheim from the 18th of February 2008. Red crosses indicate detected warnings of the NOC check, black crosses of the LIM check.

4 INTEGRATION INTO DATA PUBLICATION PROCESSES

The integration of a data publication process into a more general scientific process was described by Quadt et al. (2012). They illustrated that for reasons of symmetry, a peer review of data has to be introduced into the publication process to make it comparable to traditional paper publications. The same paper demonstrated a software implementation of a data publication process, which included scientific quality assurance (SQA) for data and metadata. While the metadata were quality assured via a web interface, the authors stressed that a thorough documentation of the quality tests used for the data is needed. These steps of documentation could of course also be included into the workflow of a web interface. In this workflow a data author would document the steps that were performed in the quality assurance.

One important aim of a peer review is to generate the transparency and credibility of a dataset's quality. Therefore it is recommended that an independent entity such as the publishing data centre should perform the quality checks on their resources. Necessary for this are the availability and implementation of standardised quality checks, completely defined by sets of parameters, at the data centre. Furthermore, this would require that the data be available in standardised formats. An example of such a standard for meteorological-oceanographic datasets is the Climate and Forecast (CF) convention (Eaton et al., 2011). By connecting data and metadata into one file, the convention allows a better automated analysis of the datasets.

Depending on the grade of detail of the interface, the workflow step that up to this point has been used for test documentation can now be used to initialise the tests. When the author has finished the input of his/her preferred test settings, the data centre performs the computations and delivers the result to the author. Nevertheless, this

step should not be underestimated because the integration of the capacity for analysing datasets is a technical challenge for a data centre.

Likewise, interfaces for the reviewer can be constructed. Data centres have to account not only for the evaluation of the quality of datasets by using the performed checks but also for the possibility that reviewers might want to run different tests based on their own design. Also, reviewers might want to change the parameters of the performed checks and rerun the quality estimation. Depending upon the complexity of the datasets and checks, this could require additional computational time for the data centre. By integrating further typical steps of a peer review, such as the function of an editor and the communication between the reviewers and the author, it is possible to generate a system for data publication similar to the online review systems of today's journals. This system would incorporate a well documented data peer review, comparable to that of traditional publications, and allow for the advantages of the established scientific system. The main requirement for the acceptance of such a system would be the effective working time needed by the scientists for authoring or reviewing a dataset. Minimising working time and maximising the amount of generated and well presented information for a data re-user will be the main task for every future data peer review system.

5 DISCUSSION

The quality evaluation scheme as described in Section 2 allows the analysis of given datasets through quality checks and the parameterised knowledge of an expert. As a result of this analysis, the quality of a dataset is given a value between 0 and 1. The important question is whether it is generally possible to assign such a value to a dataset in a meaningful way.

To answer this question it must be known whether the tests and expert knowledge are appropriate for the user of the quality estimate. Because quality is a subjective quantity, every scientist has his/her own view on exactly what it means. A good dataset for one may not work so well for another user of the dataset. While on the technical side of the dataset, most scientists might agree on the quality, agreement on the quality of the content of the dataset may be much more problematic. Nevertheless, in the environment of a peer review, this is acceptable because for journal articles too, opinions vary. It is therefore very important that when these quality estimates are published after the peer review, the exact configuration of all input parameters of the tests and their evaluation be presented in a transparent way. A general publication of these values without explanations and potentially further reduced to a traffic light like system, would neither help the publishing author or the re-user of the data. Therefore the risk is not in creating these quality estimates but in presenting and using them incorrectly.

Nevertheless, the information generated within such a process could be of huge help for a data re-user when deciding whether or not a dataset is the correct one for his/her work. For this type of decision a simple and informative method of presentation must be found. This might require the availability of quality estimates calculated with different standardised sets of parameters and priors that reflect the views of different target groups.

In particular, the scheme presented here is advantageous in that not just the extreme value checks are used for the database analysis. This approach is common when data values are flagged depending on their quality level (You & Hubbard, 2006). By using low weighted, weaker constraints on datasets, further and more detailed information for a data re-user might be possible. Reducing this detailed information to simple flags is then made possible by implementing thresholds of the quality estimates for certain levels of quality flags. Furthermore, the generated quality estimate can be used in a statistical analysis of the dataset.

One challenge is the extension of this scheme to other types of datasets. An example of this is to analyze the metadata of the datasets when checks for metadata quality are available. In order to accomplish a metadata quality check, the definition of quality, given in Section 2, has to be adjusted. For metadata, the definition of quality is no longer the truthfulness of the data but its optimal description. From a mathematical point of view no changes to the equations are necessary. With quality estimates for both the data and the metadata, it is possible to give a quality estimate of the complete dataset. With similar adjustments to the scheme, the presented quality evaluation method can be used in a large variety of structured and unstructured data.

The integration of such schemes into data centres can be complicated, especially when quality checks are calculated within the infrastructure of the centre. In data publication many different types of data and physical variables are common and have to be expected. Developing effective general quality checks for as many of these varieties as possible is essential for such schemes. Furthermore, the generation of adequate priors and

weightings might be a serious challenge at the beginning. Because the time needed for scientists to review and author a dataset is limited, ways to automate the generation of parameters, priors, and weightings might be of interest. When datasets are similar to each other, tests chosen for one dataset might be applied to a new dataset as well. To do this, well prepared metadata and standardisation are a key issue.

The standardisation of datasets and formats helps to minimise the efforts of the data centres and reviewers. Even so, the authors' efforts may become greater, depending on the scientists' workflow. The procedures proposed here are not dependent upon standardised datasets, even when the application of the scheme is simplified considerably. Alternatives to statistical tests could be the parameterised opinion of the reviewer, given in terms of probability. The quality evaluation scheme could use and weight these estimates accordingly and create the type of quality estimation shown in this paper. Nevertheless, such a procedure would drastically reduce the transparency and is therefore not recommended.

A problem for the peer review of data is the choice of reviewers. First of all it has to be expected that scientists performing a review on data have lower motivation than those reviewing for traditional publications. The advantage of being able to view a dataset before the general public might be not seen as important as being able to be the first to read a journal paper. It has also to be expected that datasets are already available for scientists before they are entered into a formal review process. Additionally, it is not easy to find the right reviewer for a dataset. This was addressed by Parsons and Duerr (2005) who discussed data stewardships for their data centre. Different target groups and therefore different requirements for the choice of tests, parameters, priors, and weightings can be challenging. Because it must be expected that not all reviewers and data re-users have a statistical background, any statistical scheme used in a review process has to be quite simple and straightforward. The scheme used in this paper fulfils this requirement, especially when the input data for the checks and evaluation are published alongside the quality estimate.

Another point is whether peer review of data is necessary for data publication at all. Many different forms of review, such as open review or the integration of a data review into the general review process of journals, have been developed and introduced. The argumentation for a thoroughly performed peer review was already given by Parsons et al. (2010). For data publication to be comparable to that of journal articles, it must include a similar peer review. Scientists expect from such a process that one or more reviewers have read the entire article and have given a report after thoroughly checking the stated facts of the article. For a publication entity consisting only of data values, this is, of course, not useful or practical. However, it is required to check the publication entity, in this case the dataset, as a whole and using different quality checks is an effective way to do this.

The question is how to introduce schemes that guarantee good control but are sensitive to the limited resources available. One option would be to allow different levels of review. The first level, labelled "approved by author", would be a procedure in which the data author would be asked to perform and document quality assurance steps. An example of such a scheme was shown in Quadt et al. (2012). The next level could be labelled "checked by qualified staff". In this phase, the staff of the data centre, who have a similar role to that of the staff or editor of a journal, performs the quality assurances appropriate to their expertise. The top level would be labelled "peer reviewed", in which selected external reviewers from the scientific community comment on the checks of the phases already performed and could also add further quality checks with the help and assistance of the data centre. A scheme such as the one presented here might be a first step towards an effective data peer review.

It has been shown that a thorough data peer review, which fulfils similar standards as reviews in the traditional publications, is possible under similar time constraints. In connection with Quadt et al. (2012), it is furthermore stressed that raw data publication can be seen as a third form of scientific publication alongside of journal papers and data papers. With the preceding arguments, the question is no longer whether a data peer review is generally possible, but whether it is needed in the different scientific communities for the advancement of science and is therefore worth the effort involved to make it happen. Nevertheless, because the introduction of data peer review also has a decisive science theoretical component, many more steps will have to follow to make data publications fully comparable to traditional paper publications.

6 CONCLUSION

The paper at hand introduces a statistical scheme that uses results from quality checks and expert knowledge to generate a quality estimation of a dataset. How this and similar schemes might be introduced into the workflow of data publication at data centres is discussed and evaluated.

The next steps to be taken could involve the introduction of this and other schemes on a larger scale and a consideration of how to deal with unstructured data. For example, information such as model codes might be evaluated similarly. This would require quality checks for this form of data. A future goal could be to have in place effective procedures for the publication of all steps of the scientific process that deliver the same credibility as those procedures used in the current journal publications. A requirement for this is a set of effective and transparent peer review schemes, similar to the one presented in this paper.

7 ACKNOWLEDGEMENT

This work was funded by the Deutsche Forschungsgemeinschaft DFG (Literatur- und Informationssysteme) under the number He1916/18-1. This work is also partly associated with the project “iGlass” of the National Environmental Research Council under the number NE/I008365/1. For providing the dataset and fruitful discussions we thank Volker Wulfmeyer and Hans-Stefan Bauer from the University of Hohenheim. Additionally, we like to thank the contributors of the R project (R Core Team, 2013). Furthermore, we would like to thank the anonymous reviewer for the constructive comments.

8 REFERENCES

Crewell, S., Mech, M., Reinhardt, T., Selbach, C., Betz, H.-D., Brocard, E., Dick, G., O’Connor, E., Fischer, J., Hanisch, T., Hauf, T., Hünerbein, A., Delobbe, L., Mathes, A., Peters, G., Wernli, H., Wiegner, M. & Wulfmeyer, V. (2008) The general observation period 2007 within the priority program on quantitative precipitation forecasting: Concept and first results. *Meteorologische Zeitschrift* 17(6), 849-866.

Costello, M. J. (2009) Motivating Online Publication of Data. *BioScience 59(5), 418-427.*

Dose, V. & Menzel, A. (2004) Bayesian analysis of climate change impacts in phenology. *Global Change Biology* 10(1), S. 259-272.

Eaton, B., Gregory, J., Drach, B., Taylor, K. E. & Hankin, S. (2011) NetCDF Climate and Forecast (CF) Metadata Conventions. Retrieved November 25, 2013 from the World Wide Web: http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.6/cf-conventions-multi.html

Gandin, L. S. (1988) Complex Quality Control of Meteorological Observations. *Monthly Weather Review* 116(5), 1137-1156.

Henneken, E. A. & Accomazzi, A. (2011) Linking to Data - Effect on Citation Rates in Astronomy. *arXiv.org* (1111.3618) Retrieved November 25, 2013 from the World Wide Web: http://arxiv.org/abs/1111.3618

Hense, A. & Wulfmeyer, V. (2008) The German Priority Program SPP1167 “Quantitative Precipitation Forecast”. *Meteorologische Zeitschrift* 17(6), 703-705.

InterAcademy Council (2010) InterAcademy Council: Climate change assessments – Review of the processes and procedures of the IPCC. Retrieved November 25, 2013 from the World Wide Web: http://www.ipcc.ch/pdf/IAC_report/IAC%20Report.pdf

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Höck, H., Lautenschlager, M, Schindler, U., Sens, I. & Wächter, J. (2006): Data publication in the open access initiative. Data Science Journal 5, 79-83.

Lawrence, B., Jones, C., Matthews, B., Pepler, S. & Callaghan, S, (2011) Citation and Peer Review of Data: Moving Towards Formal Data Publication. *The International Journal of Digital Curation* 6, 4-37.

Meek, D. W. & Hatfield, J. L. (1994) Data Quality Checking for Single Station Meteorological Databases. *Agricultural and Forest Meteorology* 69(1-2), 85-109.

National Science Foundation (2013) The National Science Foundation proposal and award policies and procedures Guide; Part I – Grant Proposal Guide. Retrieved November 25, 2013 from the World Wide Web: http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpgprint.pdf

Parsons, M. A. & Duerr, R. E. (2005) Designating User Communities for Scientific Data: Challenges and Solutions. *Data Science Journal* 4, 31-38.

Parsons, M. A., Duerr, R. & Minster, J.-B. (2010) Data Citation and Peer Review. Eos, *Transactions American Geophysical Union* 91(34), 297-299.

Piwowar, H. A. & Vision, T. J. (2013) Data reuse and the open data citation advantage. *PeerJ* 1(e175).

Quadt, F., Düsterhus, A., Höck, H., Lautenschlager, M., Hense, A. V., Hense, A. N. & Dames, M. (2012) Atarrabi – A Workflow System for the Publication of Environmental Data. *Data Science Journal* 11, 89-109.

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. Published in Vienna, Austria

Research Councils UK (2012) RCUK Common Principles on Data Policy. Retrieved November 25, 2013 from the World Wide Web: http://www.rcuk.ac.uk/research/datapolicy/

World Meteorological Organization (2007) Guidelines for quality control procedures applying to data from automatic weather stations. In Guide to the Global Observing System, WMO-No. 488.

Wulfmeyer, V & I. Henning-Müller (2006) The climate station of the University of Hohenheim: Analyses of air temperature and precipitation time series since 1878. *International Journal of Climatology* 26, 113-138.

Wulfmeyer, V., Behrendt, A., Kottmeier, Ch., Corsmeier, U., Barthlott, C., Craig, G.C., Hagen, M., Althausen, D., Aoshima, F., Arpagaus, M., Bauer, H.-S., Bennett, L., Blyth, A., Brandau, C., Champollion, C., Crewell, S., Dick, G., Di Girolamo, P., Dorninger, M., Dufournet, Y., Eigenmann, R., Engelmann, R., Flamant, C., Foken, T., Gorgas, T., Grzeschik, M., Handwerker, J., Hauck, C., Höller, H., Junkermann, W., Kalthoff, N., Kiemle, C., Klink, S., König, M., Krauss, L., Long, C. N., Madonna, F., Mobbs, S., Neininger, B., Pal, S., Peters, G., Pigeon, G., Richard, E., Rotach, M. W., Russchenberg, H., Schwitalla, T., Smith, V., Steinacker, R., Trentmann, J., Turner, D.D., van Baelen, J., Vogt, S., Volkert, H., Weckwerth, T., Wernli, H., Wieser, A. & Wirth, M. (2011) The Convective and Orographically Induced Precipitation Study (COPS): The Scientific Strategy, the Field Phase, and First Highlights. *Quarterly Journal of the Royal Meteorological Society* 137, 3-30.

You, J. & Hubbard, K. G. (2006) Quality Control of Weather Data during Extreme Events. *Journal of Atmospheric and Oceanic Technology* 23, 184-197.

9 APPENDIX

Table 2. Measured variables at the climate station in Hohenheim. Shown are the tests used with their parameters and the median for the quality evaluation (QE) contribution for each test. To gain the QE results for a meteorological variable the results for all tests of this variable have to be summed up.

Abbreviation	Test	Parameter 1	Parameter 2	Weight	Median QE (IQR)
TimeFromStart	ROC	downward value: 0		1	100% (0%)
T200	LIM	minimum value: -90	maximum value: 70	2	40% (1.4%)
T200	ROC	downward value: 3	upward value: 3	1	20% (0.2%)
T200	ROC	downward value: 10	upward value: 10	1	20% (0.1%)
T200	NOC	minimum variation: 0.1	time frame: 60	1	20% (9.3%)
RH200	LIM	minimum value: 0	maximum value: 100	2	40% (1.4%)
RH200	ROC	downward value: 15	upward value: 15	1	20% (1.4%)
RH200	ROC	downward value: 10	upward value: 10	1	20% (0.2%)
RH200	NOC	minimum variation: 1	time frame: 60	1	0% (9.7%)
T005	LIM	minimum value: -80	maximum value: 60	2	40% (1.4%)
T005	ROC	downward value: 10	upward value: 10	1	20% (1.4%)
T005	ROC	downward value: 5	upward value: 5	1	20% (0.2%)
T005	NOC	minimum variation: 0.1	time frame: 60	1	20% (0%)
WD1000	LIM	minimum value: 0	maximum value: 360	2	62.0% (34.9%)
WD1000	NOC	minimum variation: 10	time frame: 60	1	33.3% (17.2%)
WS1000 (2 min mean)	LIM	minimum value: 0	maximum value: 75	2	34.4% (16.8%)
WS1000 (2 min mean)	ROC	downward value: 20	upward value: 20	1	20% (5.6%)
WS1000 (2 min mean)	ROC	downward value: 10	upward value: 10	1	20% (0.6%)
WS1000 (2 min mean)	NOC	minimum variation: 0.5	time frame: 60	1	20% (0%)
GR200	LIM	minimum value: -1600	maximum value: 1600	2	100% (3.5%)
RR200	LIM	minimum value: -1600	maximum value: 1600	2	100% (3.5%)
NR200	LIM	minimum value: -1600	maximum value: 1600	2	100% (3.5%)
RAIN (1 min sum)	LIM	minimum value: 0	maximum value: 40	2	100% (7.0%)
ST002	LIM	minimum value: -50	maximum value: 50	2	100% (3.5%)
ST005	LIM	minimum value: -50	maximum value: 50	2	50% (1.8%)
ST005	ROC	downward value: 1	upward value: 1	1	25% (1.8%)
ST005	ROC	downward value: 0.5	upward value: 0.5	1	25% (0.2%)
ST010	LIM	minimum value: -50	maximum value: 50	2	50% (1.8%)
ST010	ROC	downward value: 1	upward value: 1	1	25% (1.8%)
ST010	ROC	downward value: 0.5	upward value: 0.5	1	25% (0.2%)
ST020	LIM	minimum value: -50	maximum value: 50	2	50% (1.8%)
ST020	ROC	downward value: 1	upward value: 1	1	25% (3.5%)
ST020	ROC	downward value: 0.5	upward value: 0.5	1	25% (0.2%)
ST050	LIM	minimum value: -50	maximum value: 50	2	50% (1.8%)
ST050	ROC	downward value: 0.5	upward value: 0.5	1	25% (1.8%)
ST050	ROC	downward value: 0.3	upward value: 0.3	1	25% (0.2%)

The only defined radiation variable within WMO (2007) is solar irradiance. To apply this definition to the three radiation variables in this dataset (GR200, RR200 and NR200), only the limits test was applied. Since the direction of these parameters is not intrinsically defined, only the maximum in both directions is tested here.

(Article history: Received 19 February 2014, Accepted 1 May 2014, Available online 5 June 2014)