Abschlussbericht zum Forschungs- und Entwicklungsvorhaben 201 43 250 auf dem Gebiet des Umweltschutzes „Anwendung modellgestützter Beurteilungssyteme für die bundeseinheitliche Umsetzung der EU-Rahmenrichtlinie Luftqualität und ihrer Tochterrichtlinien“

Formulation of criteria to be used for the determination of the accuracy of model calculations according to the requirements of the EU Directives for air quality – Examples using the chemical transport model REM-CALGRID

Rainer Stern

Johannes Flemming

Freie Universität Berlin
Institut für Meteorologie
Troposphärische Umweltforschung

Mai 2004

Berichts-Kennblatt

BerichtsnummerUBA-FB	2.	3.

4. Titel des Berichts

Formulation of criteria to be used for the determination of the accuracy of model calculations according to the requirements of the EU directives for air quality – Examples using the chemical transport model REM-CALGRID

- Englische Kurzversion des Berichts „Konzepte zur Formulierung von Kriterien zum Nachweis der von den EU-Richtlinien verlangten Genauigkeit von Modellrechnungen--Beispiele mit dem chemischen Transportmodell REM/CALGRID“-

5. Autor(en), Name(n), Vorname(n)
Stern, Rainer
Flemming, Johannes

8. Abschlußdatum
Mai 2004

9. Veröffentlichungsdatum
Mai 2004

6. Durchführende Institution (Name, Anschrift)

Freie Universität Berlin, Institut für Meteorologie

Carl-Heinrich-Becker-Weg 6-10

12165 Berlin

10. UFOPLAN-Nr.
201 43 250

11. Seitenzahl
25

7. Fördernde Institution (Name, Anschrift)

Umweltbundesamt, Postfach 33 00 22, D-14191 Berlin

12. Literaturangaben
25

13. Tabellen und Diagramme
1

14. Abbildungen
9

15. Zusätzliche Angaben

16. Kurzfassung

In dieser Arbeit wurde versucht, Kriterien zu formulieren, die zum Nachweis der von den EU-Richtlinien verlangten Genauigkeit von Modellrechnungen benutzt werden können. Es wurden verschiedene Maße untersucht und anhand einer europaweiten Anwendung des chemischen Transportmodells REM-CALGRID (RCG) auf das gesamte Jahr 2002 und eines Vergleichs der Modellergebnisse mit Messungen aus den deutschen Messnetzen diskutiert. Es wird vorgeschlagen, die Modellgüte nur am Mittelwert und an dem Perzentil zu bestimmen, das der Anzahl von erlaubten Überschreitungen der Kurzfristziel- oder Grenzwerte entspricht. Das Hauptproblem einer Modellevaluierung anhand von Messungen liegt aber in der Auswahl der zur Evaluierung zu benutzenden Stationen, da deren oft unbekannte oder nur unzureichend definierbare Repräsentativität große Auswirkungen auf den Modellfehler haben kann. Die Untersuchungen haben gezeigt, dass es auf Basis der gegenwärtigen EU-Richtlinienbestimmungen nicht möglich ist, eindeutige und objektivierbare Kriterien für eine Modellevaluierung anhand von Messungen zu entwickeln. Hier besteht also noch weiterer Handlungsbedarf.

17. Schlagwörter

EU-Richtlinien zur Luftqualität, Ausbreitungsrechnung, Modellfehler, Repräsentativität von Messungen, REM-CALGRID-Modell

18. Preis	19.	20.

Report Cover Sheet

Report No.UBA-FB BBBB	2.	3.

4. Report Title

Formulation of criteria to be used for the determination of the accuracy of model calculations according to the requirements of the EU directives for air quality – Examples using the chemical transport model REM-CALGRID.

Short Version in English of the report „Konzepte zur Formulierung von Kriterien zum Nachweis der von den EU-Richtlinien verlangten Genauigkeit von Modellrechnungen--Beispiele mit dem chemischen Transportmodell REM/CALGRID“

-

5. Autor(s), Family Name(s), First Name(s)

Stern, Rainer

Flemming, Johannes

6. Performing Organisation (Name, Address)

Freie Universität Berlin, Institut für Meteorologie

Carl-Heinrich-Becker-Weg 6-10

12165 Berlin

7. Sponsoring Agency (Name, Address)

Umweltbundesamt, Postfach 33 00 22, D-14191 Berlin

8. Report Date

May 2004

9. Publication Date

May 2004

10. UFOPLAN-Ref. No.

201 43 250

11. No. of Pages

25

12. No. of Reference

25

13. No. of Tables, Diagrams

1

14. No. of Figures

9

15. Supplementary Notes

16. Abstract

Several measures to define model accuracy were examined based on the results of European wide model calculations for the whole year 2002 employing the chemical transport model REM-CALGRID. Using the observations of the German measurement network, the model accuracy is discussed in context with the requirements given in the EU directives for air quality. It is proposed to calculate model accuracy for the annual mean and that percentile which corresponds to the allowed number of exceedances of the short-term limit values. However, the major problem in the determination of the model uncertainty is the selection of appropriate measurement stations, because the mostly unknown area representativity of sites has a great influence on the model accuracy. Based on the currently available guidance it seems not to be possible to develop clear and objective rules for the determination of model uncertainty.

17. Keywords

EU directives for air quality, dispersion modelling, model accuracy, representativity of stations, REM-CALGRID-model

18. Price	19.	20.

Forschungs- und Entwicklungsvorhaben 201 43 250

auf dem Gebiet des Umweltschutzes

„Anwendung modellgestützter Beurteilungssyteme für die bundeseinheitliche Umsetzung der EU-Rahmenrichtlinie Luftqualität und ihrer Tochterrichtlinien**“**

Formulation of criteria to be used for the determination of the accuracy of model calculations according to the requirements of the EU Directives for air quality – Examples using the chemical transport model REM-CALGRID

Rainer Stern and Johannes Flemming

for:

Umweltbundesamt
II 6.1
Postfach 33 00 22
14191 Berlin

Freie Universität Berlin, Institut für Meteorologie
Carl-Heinrich-Becker-Weg 6-10
12165 Berlin

Short version of the report:

Flemming, J. and Stern, R. (2004):

„Konzepte zur Formulierung von Kriterien zum Nachweis der von den EU-Richtlinien verlangten Genauigkeit von Modellrechnungen“

-Beispiele mit dem chemischen Transportmodell REM-CALGRID**“**

May 2004

Table of Content

1 Introduction

The Framework Directive 96/62/EC (FWD) on ambient air quality assessment and management explicitly designates the use of air quality modelling techniques for the assessment of air quality. According to the FWD and its daughter directives (1999/30/EC & 2002/3/EC), air quality models can be applied for the:

- Simulation of air quality parameters on all scales (international, national, urban agglomeration, street), especially for zones where concentrations of pollutants in ambient air do not exceed the upper assessment threshold or where measurements do not give sufficient information to fulfil the requirements of an air quality assessment.

Furthermore, the use of models is implicitly addressed as models are needed for the

- Prediction of the impact of air quality action plans, which have to be developed in order to ensure compliance with limit and target values.

In contrast to measurements, there is no reference methodology defined for modelling, but, as with measurements, model simulations have to meet certain accuracy standards. In the directives, an estimation of model accuracy is required with regard to the calculated annual, daily and hourly values. However, the directives give no clear guideline how to carry out the assessment of the model and modelling uncertainty.

The aim of the report is to review the EU guidelines with regard to the model quality assessment and to propose specific model quality criteria, which may be used for model accuracy assessment according to the intentions of the EU directives. Several model accuracy measures are proposed and tested utilizing a one-year simulation (2002) with a chemical transport model, REM-CALGRID (RCG). RCG is an Eulerian grid model of medium complexity that can be used on the regional, as well as the urban, scale for short-term and long-term simulations of oxidant and aerosol formation (Stern, 2004; Stern and Yamartino, 2001, 2002; Stern et al., 2003). A model overview can be found in the appendix. For this study, RCG was applied on the regional scale covering Europe with a resolution of about 25*25 km^2 for the entire year 2002. The model simulations were compared with the data from about 300 - 400 measurement sites of the German air quality network.

The focus of the study is on O_3, PM10, SO_2 and NO_2. The limit and target values for these species as well as the respective data quality objectives for measurement and modelling are given in the daughter directives 2003/3/EC and 1999/30/EC. RCG-model results for those species were used to demonstrate the significance and impact of the proposed measures. The accuracy measures are defined in such a way that they can be used for any air quality model which is able to simulate air quality concentrations at a one-hour time resolution for at least one year. The demand on the hourly resolution is necessary as some of the limit and target values are defined for hourly values. The proposed accuracy measures can be applied to model output for all scales, although they are discussed here only in the context of a regional scale application.

2 Definition of model uncertainty according to the EU Directives

The description of the model quality assessment procedure is given in the respective annexes of the EU directives as follows: „The uncertainty for modelling and objective estimation is defined as the maximum deviation of the measured and calculated concentration levels, over the period for calculating the appropriate threshold, without taking into account the timing of the events." The phrase „without taking into account the timing of the events" indicates that the match between the modelled and observed frequency distribution of either hourly or daily values ought to be the main model quality criteria. The model quality objectives for the allowed uncertainty are given as a relative uncertainty without clear guidance on how to calculate this relative uncertainty. The stipulated accuracy bounds vary between ±30% for annual averages of SO_2 and NO_2, ±50% for PM10 and ±50% to 60% for daily and hourly averages, respectively. To obtain a relative uncertainty, it can be assumed that the respective measured value shall be used to normalize the absolute difference between the maximum deviation of the measured and calculated concentration levels. Another possibility would be to take the maximum relative deviation. However, such an approach could shift the emphasis to the very low measured concentration ranges where usually the largest relative deviations between observations and calculations occur. Starting from those prepositions, several accuracy measures were examined with and without taking into account the timing of the events. „With timing" means that the model errors are determined in the usual way at corresponding hours of the observed and calculated time series, whereas „without timing" means that the errors are the difference of the observed and calculated concentration values at the same percentile. The latter approach can be described more easily as taking the differences between the highest observed and the highest calculated value, between the second highest observed and calculated value, and so on down to the lowest value in each pair of time series. The question of timing is then only relevant for those limit or target values which are defined as a number of allowed exceedances of a given threshold concentration. This is the case for the following limit and target values and their associated frequency percentiles:

- SO_2, 1-h mean of 350 µg/m^3, not to be exceeded more than 24 times a calendar year (99.73-percentile),
- SO_2, 24-h mean of 125 µg/m^3, not to be exceeded more than 3 times a calendar year (99.18-percentile),
- NO_2, 1-h mean of 200 µg/m^3, not to be exceeded more than 18 times a calendar year (99.79-percentile),
- PM10, 24-h mean of 50 µg/m^3, not to be exceeded more than 35 times a calendar year, stage 1, (90.41-percentile)
- O_3, maximum daily 8-h mean of 120 µg/m^3, not to be exceeded more than 25 times a calendar year, averaged over three years, (93.15-percentile).

For the limit values that are defined as annual means, the question of timing is irrelevant. Starting from those prepositions, the following accuracy measures where examined:

- Relative maximum error with timing, $rel\,max\,err\_t$, and without timing, $rel\,max\,err\_p$:

$$rel\,max\,err\_t = \frac{\max\left(|\,o_t - r_t\,|\right)}{o_{t_{\max(o_t - r_t)}}}$$

$$rel\,max\,err\_p = \frac{\max\left(|\,o_p - r_p\,|\right)}{o_{p_{\max(o_p - r_p)}}}$$

- Maximum relative error with timing, $max\,rel\,err\_t$, and without timing, $max\,rel\,err\_p$:

$$max\,rel\,err\_t = \max\left(|\frac{o_t - r_t}{o_t}|\right)$$

$$max\,rel\,err\_p = \max\left(|\frac{o_p - r_p}{o_p}|\right)$$

- Relative error without timing at the percentile which corresponds to the allowed number of exceedings of the limit value concentration threshold, $rel\,per\,err\_p_{LV}$:

$$rel\,per\,err\_p = \frac{|\,o_p - r_p\,|}{o_p},\ p = p_{LV}$$

where p_{LV} is taken at the 99.79th-percentile of the hourly NO_2 values, the 90.41th-percentile of the daily mean PM10 values, or the 93.15th-Percentile of the maximum 8-h daily means of ozone, respectively.

- RMSE with timing, $rmse\_t$, and without timing, $rmse\_p$:

$$rmse\_t = \sqrt{\frac{1}{n_t}\sum_{t=1}^{n_t}(o_t - r_t)^2}$$

$$rmse\_p = \sqrt{\frac{1}{n_p}\sum_{p=1}^{n_p}(o_p - r_p)^2}$$

With

$p = 1, n_p$ Percentiles ; $t = 1, n_t$ hours ; $p_1 < p_2 \longleftrightarrow o_{p_1} \le o_{p_2}$ and $r_{p_1} \le r_{p_2}$

o_p hourly observations ordered by magnitude, o_t observation at hour t

r_p hourly calculations ordered by magnitude, r_t calculation at hour t

The model quality measure described in the EU directives is interpreted as the relative maximum error without timing, $rel\,max\,err\_p$, which is the largest concentration difference of all percentile differences normalized by the respective measured value.

The error measures were calculated based on the hourly observed and modelled time series and were determined for each available station separately. The daughter directives say nothing about a selection criteria for stations to be used for the determination of model accuracy. Using all available stations leads to the problem that many stations are not representative for the model's resolution (e.g. those monitors next to major roadways or embedded within a busy urban area) and therefore could cause a model failure in terms of the accuracy

objectives. It is almost obvious that modelling with a 10 - 50 km horizontal grid resolution will have problems to simulate a measurement site influenced by local traffic. Using all stations, the observation strategy, which defines the number of the stations in certain air quality regimes, will have an impact on the determined overall model accuracy, i.e. a large scale model evaluation utilizing the observations of a network with the majority of the stations in urban areas will be worse than a model evaluation utilizing the observations of a network with the majority of the stations in rural areas.

A solution to this problem might be a pre-selection of stations to be used for model evaluation. This pre-selection relied on prior knowledge of the air quality regime of the measurement sites. This regime can be determined either by an assessment by the local authorities or by the use of an objective classification method. Such an objective method, which classifies the German sites into 6 different air quality regimes by means of hierarchical clustering, was used in this study (see Flemming, 2003a). Table 1 shows this species-dependent classification.

species							
Ozone	Mountain B	Rural R	Sub-urban U1	Urban U2	Urban-street, polluted U3	Street S	
NO_2	Rural R		Sub-urban	Urban U2	Urban-street, polluted U3	Street S	Street, severely polluted: S2
PM10	#1		#2	#3	#4	#5	
SO_2	#1		#2	#3	#4	#5	

Table 1: Air quality regimes derived via hierarchical clustering. The classification is based on the daily mean and the daily variation of each pollutant (see Flemming, 2003a).

To conform to the existing classification scheme for ozone used by the German Umweltbundesamt, the names for the regimes “mountain”, “rural”, “urban” and “street”, were maintained in this objective classification. However, one of the objectives of the classification is to resolve more precisely the large group of urban time series. Therefore, these time series are sub-clustered into regimes of “suburban” plus a transition regime “urban–street “, which means that, in total, six different ozone regimes were determined. The NO_2 classification is similar to the ozone classification, except for the mountain regime, which is not applicable for NO_2. Substantial testing and cross-comparison between the ozone and NO_2 classifications revealed that at stations which belong to the ozone regime “street”, a wide range of rather high NO_2 levels is observed. To take into account these very high NO_2 observations, an additional regime “severely polluted street” was introduced. The number of PM10 and SO_2 clusters was chosen to be five, which correspond to the number of ozone regimes without the mountain-regime. PM10 and SO_2 air quality regimes are simply labelled from #1 to #5, indicating an increase in the mean pollution level.

3 Model accuracy for annual means

For the annual means, the model accuracy requirements of the EU directives are rather unambiguous. The accuracy for the calculated annual averages should be 30% for NO_2 and SO_2, and 50% for PM10. A distinction between a paired-in-time and an unpaired-in-time intercomparison, or between the relative maximum error and the maximum relative error is not necessary.

Figure 1 shows the scatter diagram of the observed annual mean concentrations and the appropriate relative model error for NO_2, PM10 and SO_2 at all available stations in Germany. The observed annual mean values of NO_2 cover a wide range which extends from low concentrations below the lower assessment threshold (LAT) to very high concentrations far beyond the limit value. Low annual means are observed at the rural (labelled R) and sub-urban (labelled U1) stations, the highest concentrations are found at the street stations (labelled S and S2). It is obvious that the model error is larger than 30% at many stations. The large scale model application fulfils the accuracy requirements for the annual mean of NO_2 at less than 50% of the rural stations, i.e. at stations which are supposed to be representative for the model scale of 25 to 30 km resolution (Figure 2). It is not surprising that the compliance rate decreases with increasing pollutant levels. At most of the street stations, the model error is larger than 30%. The observed annual mean NO_2 concentrations at the rural stations are low and well below the lower assessment threshold. Therefore, a small absolute deviation of the calculation from the observation can create a large relative error.

For the annual PM10 mean values, the model error increases in most cases with increasing observed mean values (Figure 1, middle). However, at most of the stations the model error is smaller than the allowed 50% (see also Figure 2, middle). It is interesting to note, that errors larger than 50% occur at some rural stations with rather low observed PM10 levels and at the severely polluted street stations.

The largest relative model errors are calculated for the annual SO_2 mean values (Figure 1, below). This is not surprising because at most of the stations the observed annual means are rather low, therefore, a small absolute deviation creates a large relative error.

Overall, it can be stated that the large scale model application described here does not completely fulfil the accuracy requirements given by the EU Directives for the annual mean values, even if only the rural stations which are supposed to be appropriate for a model resolution of 25 to 30 km, are considered. The main reason for this non-compliance with the accuracy requirements for the annual mean is the fact, that rather small absolute model deviations can create large relative errors at stations with low observed annual mean values. It can be expected that this fact is also a problem for other models as well. The comparison of the RCG results with the results of the other models of the GLOREAM aerosol study shows that RCG, as a model of medium complexity, performs well in comparison to other models of equal or even higher complexity (Hass et al., 2003). Also the RCG performance within the TOR study (Roemer et al., 2003), the EURODELTA[1] , the CITYDELTA[2] or the TFMM[3] model intercomparison studies clearly show that RCG results are comparable to those of other models.

[1] http://rea.ei.jrc.it/netshare/thunis/eurodelta

[2] http://rea.ei.jrc.it/netshare/thunis/citydelta

[3] www.nilu.no/projects/ccc/tfmm/index.html

More detailed guidance on the determination of model uncertainty should consider whether the very low observed concentration levels should be excluded from a model evaluation exercise.

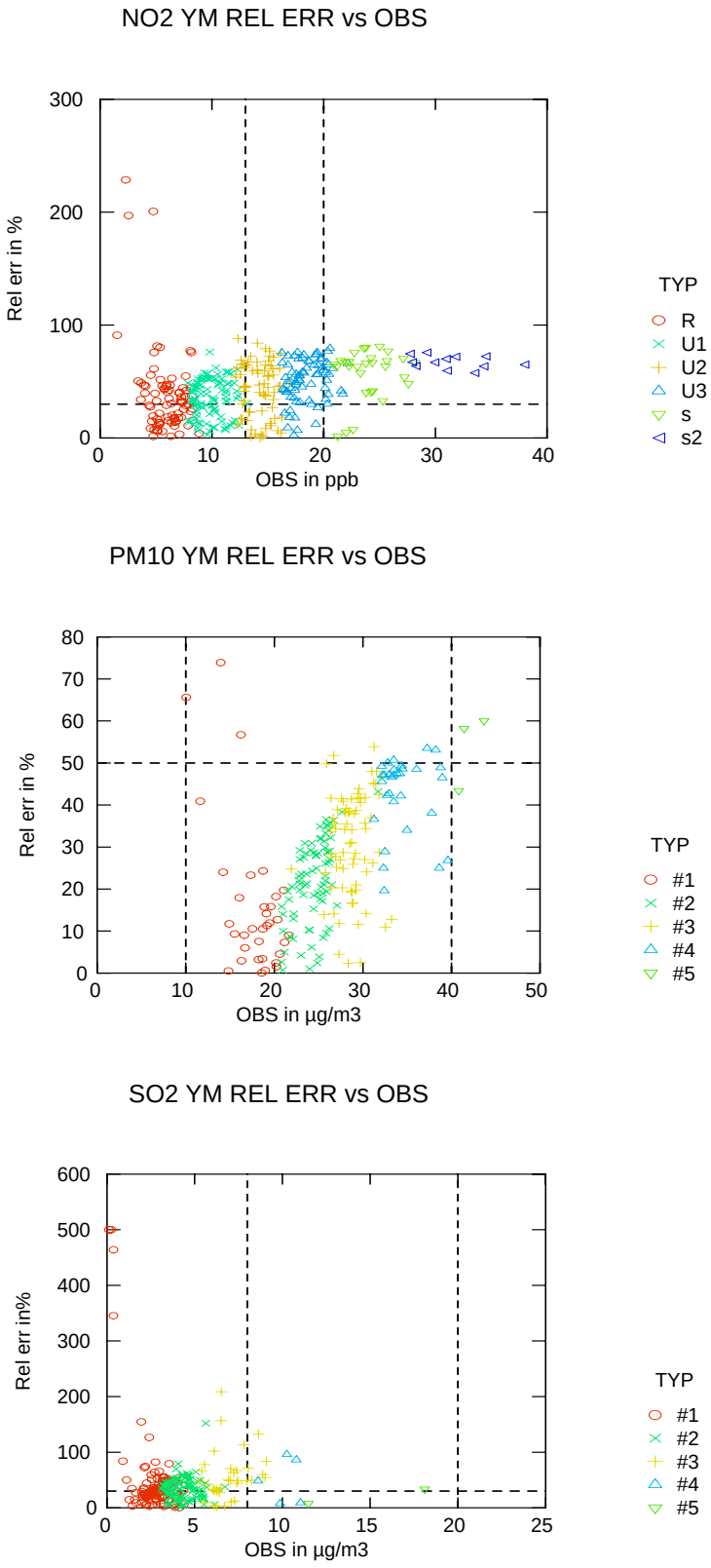

Figure 1: Scatter diagram of the observed annual mean values at German sites (x-axis) and the relative model error in % (y-axis) for NO_2, PM10 and SO_2. The horizontal dashed line indicates the model accuracy requirement of the EU directives (30% for SO_2, NO_2, 50% for PM10). The vertical dashed lines indicate the lower assessment threshold and the limit value. The station type indicates the air quality regime, see Table 1. For further explanation see text.

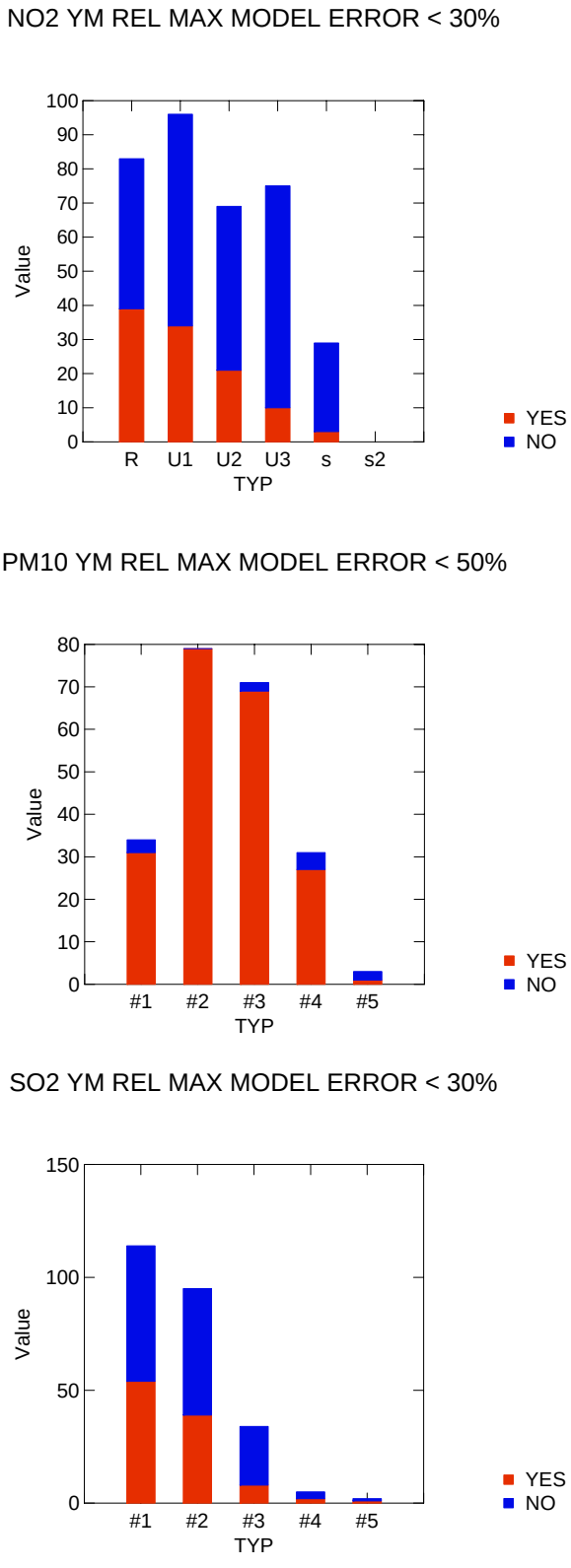

Figure 2: Number of stations in Germany where the model results fulfil (YES) or do not fulfil (NO) the accuracy requirements for the annual mean. For the definition of the station types see Table 1.

4 Model accuracy for daily and hourly averages

Most of the problems with the interpretation of the model accuracy requirements according to the EU Directives occur in the case of the comparison of the hourly and daily concentrations. This chapter shows first the differences between a comparison with and without timing of the events. The next paragraph investigates the consequences of the error measure as defined in the EU directives and proposes an alternative error measure. Finally, an overview of the model performance utilizing all proposed relative error measures is presented.

4.1 Paired-in-time versus unpaired-in-time

Overall, the modelling errors for hourly and daily averages decrease when the model accuracy is calculated without regard to the timing of the events, rather than for paired-in-time results. This can be seen in Figure 3 which shows Box-Whiskers charts[4] of the distribution of the RMSE calculated, paired-in-time and unpaired-in-time results for all stations in Germany. It is obvious that the median of the RMSE distribution calculated without regard to the timing of the events, is smaller for all species than the median of the RMSE distribution calculated for paired-in time cases. Because the EU directives do not demand a paired-in-time comparison, the main emphasis in the remaining parts of this report is placed on the measures without regard to the timing of the events.

[4] The Box-Whisker graph illustrates the spread of data groups around their medians, using a “box” and “whiskers” to break down each data group by percentile. The center line within the box gives the median of the distribution of the data. The top and bottom box gives the 25th- and 75th-percentiles, respectively. The top and bottom of the whisker (marked with horizontal crossbars) indicate the farthest points that are not outliers (i.e. that are within 3/2 times the inter-quartile range. Outliers are marked by dots.

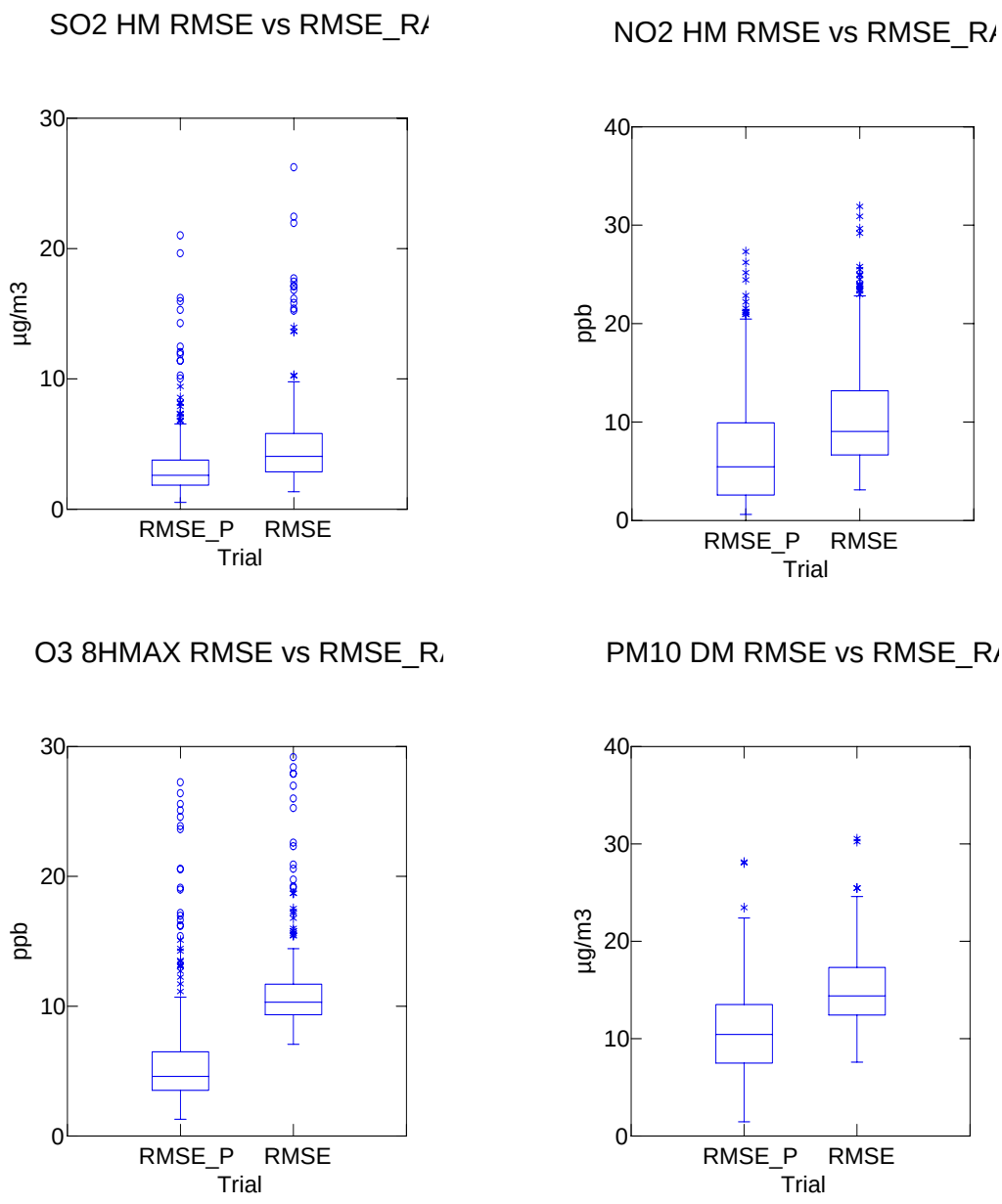

Figure 3: Box-Whisker charts of the distribution of the RMSE for the hourly SO_2 concentrations (upper left), the hourly NO_2 concentrations (upper right), the maximum daily 8-h-mean of O_3 (lower left) and the daily mean values of PM10 (lower right). RMSE_P: unpaired-in-time. RMSE: paired-in-time. All German stations.

4.2 The maximum deviation without taking into account the timing of the events

The maximum deviation without taking into account the timing of the events is the absolute error measure according to the EU directives[5] and is defined as follows:

$$max\,err\_p = \max\left(|o_p - r_p|\right)$$

As already described in Chapter 2, the maximum deviation is defined as the largest difference out of all differences between the observed and calculated values of equal rank in the frequency distribution.

The investigation of this error measure at the German stations reveals that at many stations the absolute maximum error $max\,err\_p$ is calculated at the highest percentile, i.e. at the highest measured value (Figure 4). In those cases the assessment of the model accuracy depends on the model performance in a concentration range associated with an extremely small probability. This means also, that the model accuracy assessment can be based on an outlier concentration caused by an error of the monitoring unit or an extreme weather situation. On the other hand, if it is assumed that the maximum deviation is the maximum relative error without timing,

$$max\,rel\,err\_p = \max\left(|\frac{o_p - r_p}{o_p}|\right),$$

one encounters the problem that, as in the case of the annual mean values, a small measured value can cause a very large relative error, that easily is in the range of 100% or higher. Using this measure, the model would most likely be judged on the model's ability to predict properly in the low concentration ranges.

Because of these problems, the alternative model error measure

$$rel\,per\,err\_p = \frac{|o_p - r_p|}{o_p}, p = p_{LV}$$

is proposed, which defines the concentration difference at the percentile that corresponds to the allowed number of exceedances of the limit value normalized by the observation value at that percentile. This measure is more robust than the error $rel\,max\,err\_p$ and also evaluates the model performance in the high concentration ranges, but without the sensitivity to outliers. Because the model accuracy is examined in the concentration range of the limit values there is also a direct link to the EU Directives. This measure was already used by Stedman et al. (2003) for air quality assessment in the UK.

The comparison of the error measure $rel\,max\,err\_p$ with the measure $rel\,per\,err\_p_{LV}$ shows that the number of stations simulated within the required accuracy range is higher for the accuracy measure linked to a fixed percentile (Figure 5, Figure 6, Figure 7, Figure 8). Overall, the model performance is better at stations which belong to an air quality regime that can be appropriately simulated by a large scale model application. However, even some stations belonging to regimes which should be resolved by the model cannot be simulated with the required accuracy. This may point to an incomplete model formulation or to uncertainties in

[5] In the interpretation of the authors

the input data bases, but also may indicate limitations of the air quality regime classification for the measurement sites. Sub-grid phenomena, such as the influence of local sources or the re-suspension of dust by local gusts, may be further reasons for large model errors at some sites. In addition, ozone at the mountain stations (labelled “B”) is simulated with rather large errors because only surface layer concentrations were used for this model evaluation instead of the more-well-suited concentrations predicted in the higher model layers.

Based on the error measure $rel\ per\ err\_p_{LV}$, daily averaged PM10 and ozone simulations comply with the accuracy requirement of 50% at almost all stations in Germany that are suited for a large scale model application. Compliance is reached at about three-fourths of all stations for the hourly NO_2-concentrations and at about one-third of the all stations for hourly SO_2-concentrations. The rather high number of stations with insufficient model accuracy for SO_2 points out the problem of setting relative accuracy targets. In the case of SO_2 the targets are difficult to meet because of the large number of low observed SO_2 concentrations.

4.3 Overview of the relative error measures used for hourly and daily averages

Figure 9 provides an overview of model performance utilizing the different error measures, defined earlier in Chapter 2. The largest relative model errors are seen to be calculated using the maximum relative error with timing, $max\ rel\ err\_t$, and without timing, $max\ rel\ err\_p$. For those measures, nearly none of the observed time series is modelled with to the requisite level of accuracy. The major reason for the large errors is the strong emphasis of this measure on the errors connected with the low observed concentrations. In particular, small absolute deviations from the observed concentration can create very large relative errors, if the model overestimates the low observations. Large errors are also produced by the error measure $rel\ max\ err\_t$, which is based on a paired-in-time comparison. Using the measure $rel\ max\ err\_p$, which is interpreted as being the measure defined in the EU Directives, the model fulfils the accuracy requirement of 50% at about 75% of the ozone stations, at about 50% of the NO_2 stations and at about 25% of the PM10 stations. The best model performance for the hourly and daily mean values is realized based on the measure $rel\ per\ err\_p_{LV}$, which is linked to the percentile that corresponds to the allowed number of exceedances of the limit value. However, as already discussed in the last chapter, the model is not able to reproduce the observations at all stations, even if only those stations are considered that are best-suited for the evaluation of large scale model results.

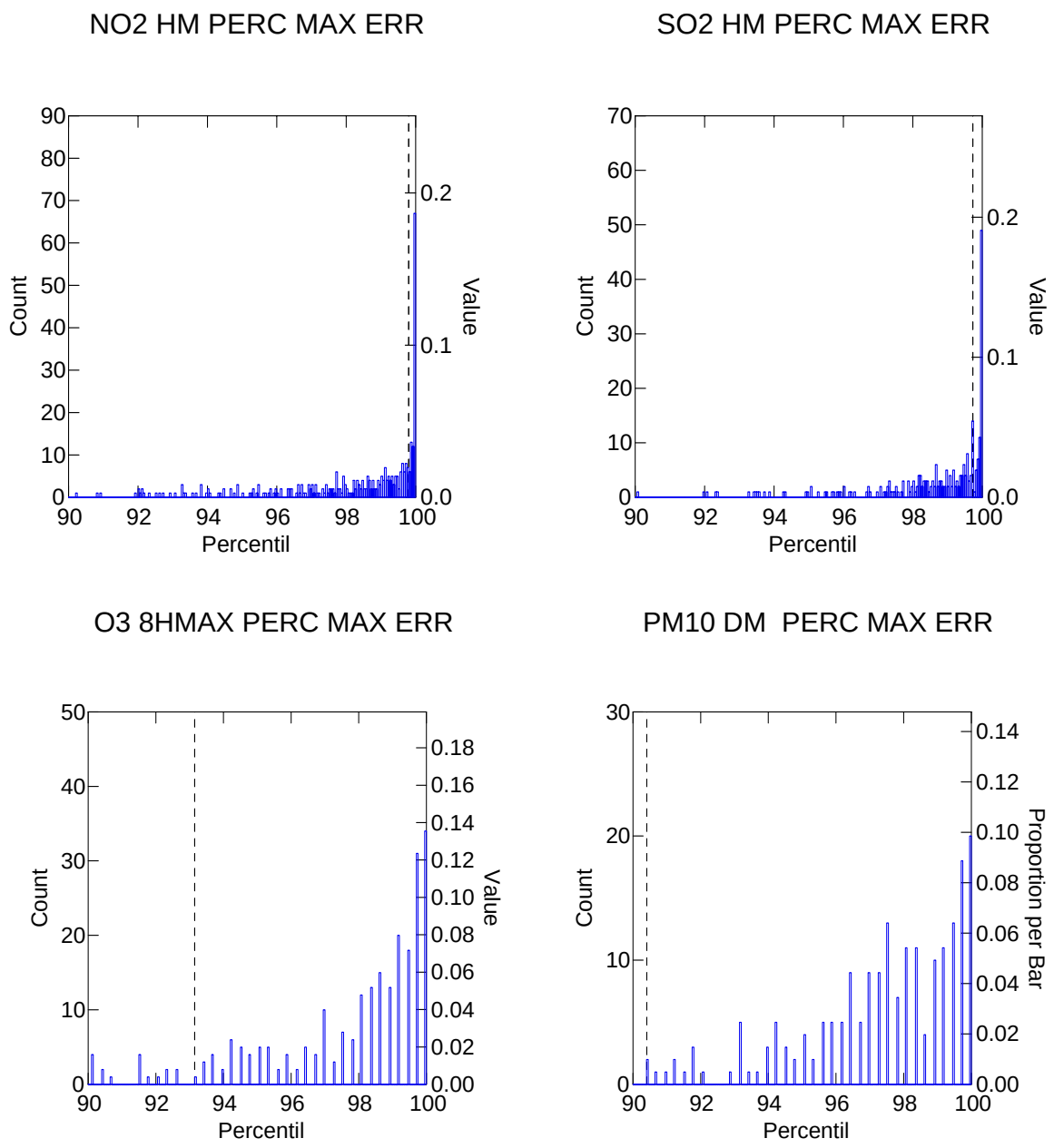

Figure 4: Histogram of the percentiles for which the largest deviation ($max\,err\_p$) is calculated. Hourly values NO_2 and SO_2, 8-h mean values O_3, daily mean values PM10. The vertical dashed lines indicate the percentile that corresponds to the allowed number of exceedances of the limit values, see Chapter 2. All German stations.

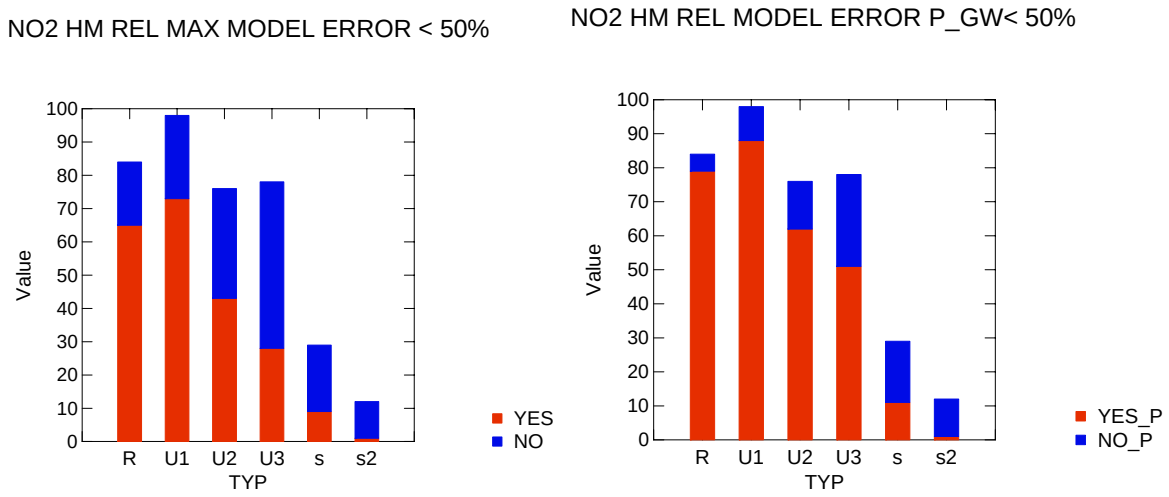

Figure 5: Number of stations in Germany at which the model results fulfil (YES) or do not fulfil (NO) the accuracy requirements of 50% for the hourly averaged NO_2 concentrations. Left hand side: error measure $rel\,max\,err\_p$, right hand side: error measure $rel\,per\,err\_p_{LV}$. For the definition of the station types see Table 1.

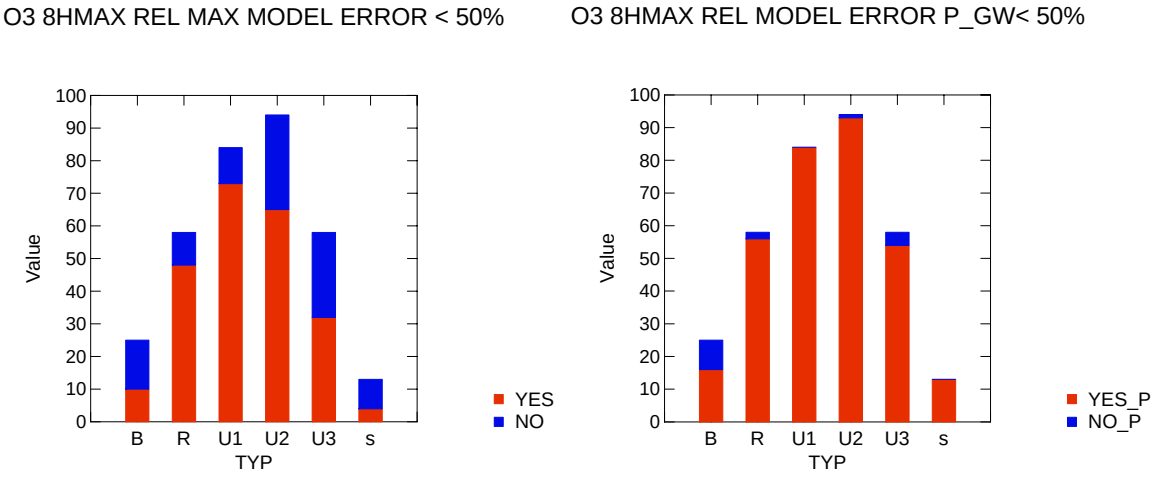

Figure 6: Number of stations in Germany at which the model results fulfil (YES) or do not fulfil (NO) the accuracy requirements of 50% for the daily 8-h mean O_3 concentrations. Left hand side: error measure $rel\,max\,err\_p$, right hand side: error measure $rel\,per\,err\_p_{LV}$. For the definition of the station types see Table 1.

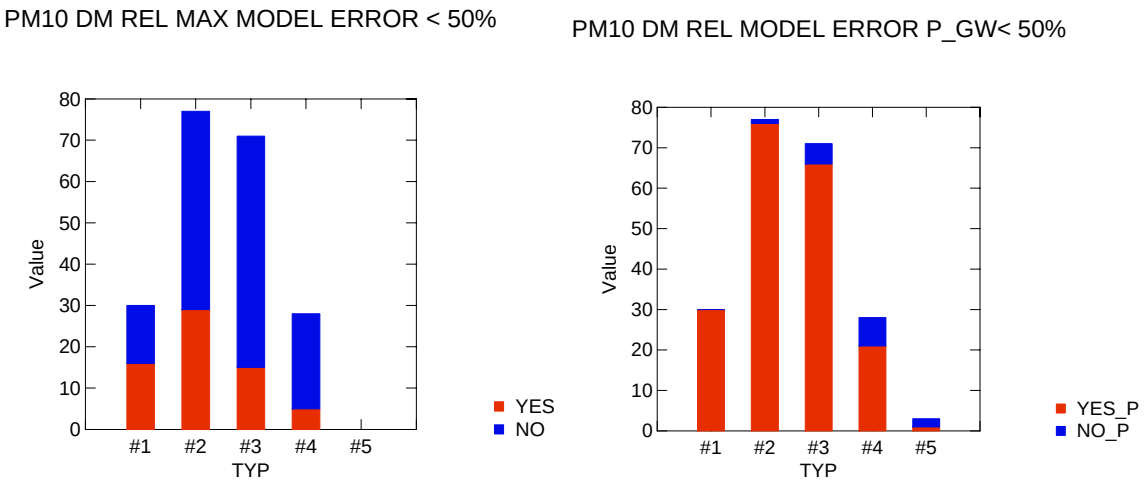

Figure 7: Number of stations in Germany at which the model results fulfil (YES) or do not fulfil (NO) the accuracy requirements of 50% for the daily mean PM10 concentrations. Left hand side: error measure $rel\,max\,err\_p$, right hand side: error measure $rel\,per\,err\_p_{LV}$. For the definition of the station types see Table 1.

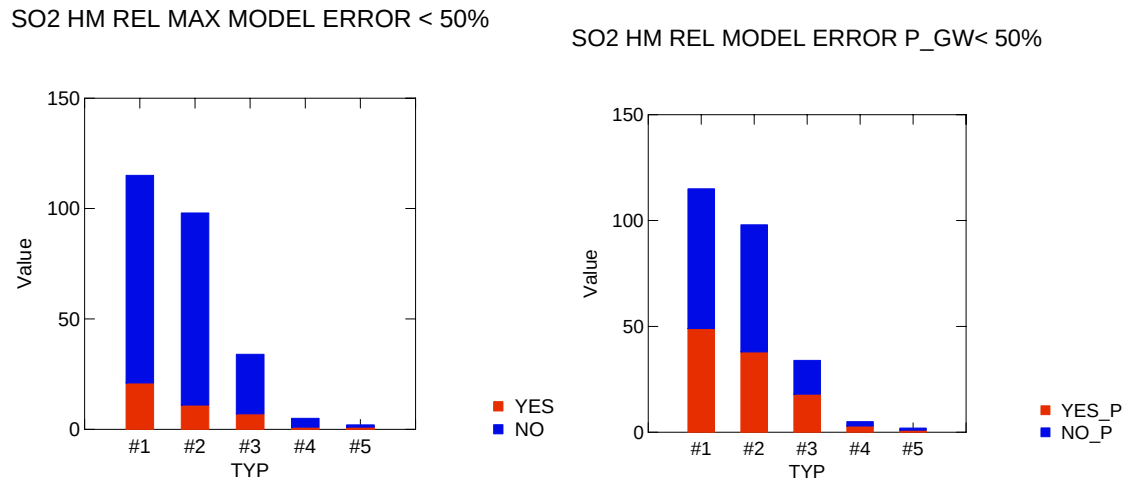

Figure 8: Number of stations in Germany at which the model results fulfil (YES) or do not fulfil (NO) the accuracy requirements of 50% for the hourly averaged SO_2 concentrations. Left hand side: error measure $rel\,max\,err\_p$, right hand side: error measure $rel\,per\,err\_p_{LV}$. For the definition of the station types see Table 1.

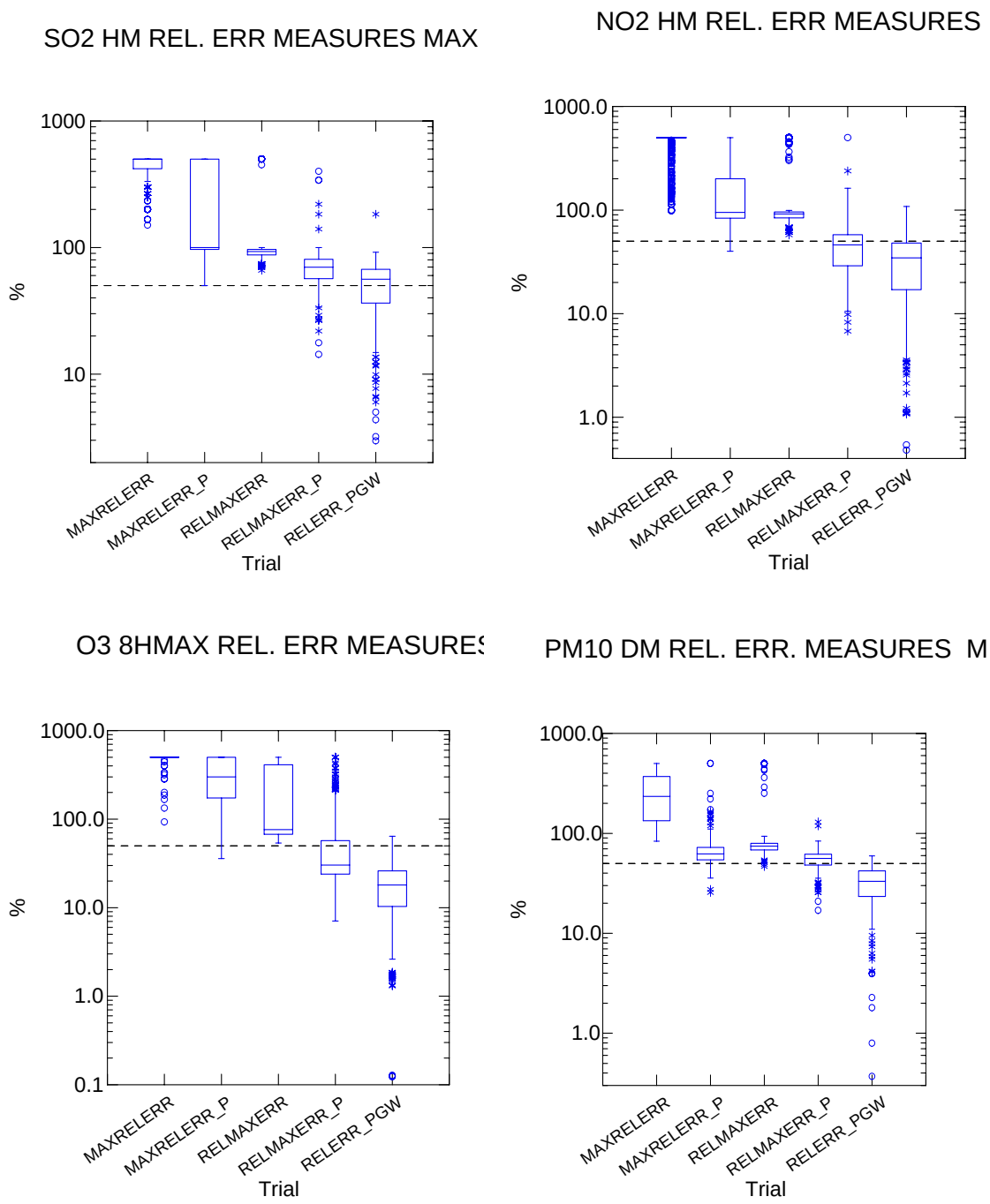

Figure 9: Box-Whisker charts of different error measures. Distributions in % of MAXRELERR: $max\,rel\,err\_t$; MAXRELERR_P: $max\,rel\,err\_p$; RELMAXERR: $rel\,max\,err\_t$; RELMAXERR_P: $rel\,max\,err\_p$; RELERR_PGW: $rel\,per\,err\_p_{LV}$. Hourly averages for NO_2, SO_2, daily 8-h-mean O3, daily mean PM10. The dashed lines indicate the accuracy requirements of the EU Directives. All stations in Germany. The measures are defined in Chapter 2.

5 Other problem areas

There is no guidance in the EU Directives on how deficiencies of the measurements, namely measurement inaccuracy, unknown representativeness of the sites, and incomplete data coverage, should be taken into account in the context of a model evaluation. In this study, several attempts were made to quantify this so called “measurement error” and to relate it to the model error measures discussed in the foregoing chapters.

A measure of the average observational area representativeness of a site is the so-called “observation error variance”, which can be estimated by interpolating spatial covariances from surrounding stations. In data assimilation theory, the observation error is a consequence of the instrument error and, more importantly, of the limited spatial representativeness of the measurements in relation to the applied model’s spatial resolution. It is defined as the spatially uncorrelated part of the measurement having zero mean. The estimation of the observation error variance is based on the observational method of Hollingsworth and Lönnberg (1986). The observation error variance at a specific station is estimated by the extrapolation of the covariance field from surrounding stations by means of a spherical covariance model. For more details see Flemming (2003b). For the German stations, the relative values of this observation error are about 3-4 times higher than the local measurement accuracy required by the EU Directives, namely 15% at a 95% confidence interval. The latter value does only account for the measurement uncertainty in comparison with a reference instrument and does not include the error induced by the unknown area representativeness of a site. Thus, the assessment of the representativeness of a site remains an untackled problem within the EU Directives.

An attempt to quantify the impact of the spatial heterogeneity of the observations has been done by means of the so-called “perfect model with a certain resolution” - hypothesis. The perfect model result is defined as the average concentration of all stations in a grid cell. Therefore, the error of the “perfect model” depends on the resolution of the grid cell and on the existing network density, which determines the number of stations within a grid cell. The error of the “perfect model” is a measure of the deviation of a single observation from the smoothed field, with the strength of smoothing defined by the resolution. The error of the “perfect model” is an important quantity since it provides a lower limit for the error of a real model application having the same spatial resolution. Therefore, improvements of model performance due to a better physical/chemical description of the relevant processes or better input data will never lead to an error smaller than this lower limit given by the “perfect model” error. Applying the current model quality objective of the EU Directives ($rel\,max\,err\_p$) to the “perfect model” using the same resolution as used in the RCG application, shows that about 20% of all NO_2 and SO_2 stations and 5-10% of all ozone and PM10 stations in Germany could never be simulated with the required accuracy via a 25 km resolution model run, even if the RCG model would gave a perfect description of reality and all the input data was error-free. The reason for this deficiency is the pronounced heterogeneity of the observed concentration fields.

The EU Directives require a minimum of 90 % data coverage of the hourly or daily values. The application of different methods to deal with missing measured values leads to the conclusion that there is no significant impact of data voids on the model accuracy parameters for the used German data. However, one has to consider that the data coverage of the German stations was mostly above 98%. Data with only 90% coverage may require a specification of how to deal with missing observed data within the model accuracy check.

6 Summary

This report reviews the EU Directive's guidelines for a model's quality assessment, and examines specific model quality criteria, that may be used for model accuracy assessment according to the intentions of the EU directives. Several model accuracy measures are tested utilizing a one-year simulation with the chemical transport model REM-CALGRID (RCG). RCG was applied in the regional scale mode, covering Europe with a resolution of about 25*25 km^2 for the entire year 2002. The model simulation results were compared with the data from about 300 - 400 measurement sites of the German air quality network.

For the annual mean values, the large scale model application described here does not fulfil completely the accuracy requirements given by the EU Directives, even if only the rural stations, which are supposed to be appropriately representative for a model resolution of 25 to 30 km, are considered. The main reason for the non-compliance with the accuracy requirements for the annual mean stems from the fact that the accuracy requirements are formulated in a relative way. Hence, cases displaying rather small absolute model deviations can create large relative errors at stations reporting low observed annual mean values.

For hourly and daily averages, the model quality measure described in the EU directives is interpreted as the maximum error without timing, which is the largest concentration difference out of all percentile differences, normalized with the respective measured value. The determination of this error measure at the German stations reveals that at many stations the absolute maximum error is calculated at the highest percentile, i.e. at the highest measured value. In those cases, the assessment of the model accuracy depends on the model performance in a concentration range having an extremely small probability. This means also, that the model accuracy assessment can be based on an outlier concentration caused by an error of the monitoring unit or an extreme weather situation. Therefore, an alternative model error measure is proposed, which is defined as the concentration difference at the percentile corresponding to the allowed number of exceedances of the limit value normalized by the observation. This measure is more robust than the error definition of the EU directives, and also evaluates the model performance in the high concentration ranges, but without being sensitive to outliers. Based on this error measure, daily averaged PM10 and ozone simulations comply with the accuracy requirement of 50% at almost all German stations that are suited to a large scale model application. Compliance was reached at about 75% of all stations in Germany for the hourly NO_2-concentrations and at about 33% of the all stations for the hourly SO_2-concentrations. The rather high number of stations with insufficient model accuracy for SO_2 points out the problem of setting relative accuracy targets. In the case of SO_2, the targets are difficult to meet because of the large number of low observed SO_2 concentrations.

Because the medium complexity RCG model performs well in comparison to other models of equal or even higher complexity, it is anticipated that other models will encounter similar problems in fulfilling the accuracy requirements of the EU Directives. Furthermore, this study shows that the pronounced heterogeneity of the observed concentration fields, derived from the dense German network, makes it impossible for a grid model to simulate all stations with the required accuracy. That is, the large observed concentration variation within a single grid cell would preclude any one predicted concentration value from meeting the accuracy criteria at all stations within the cell.

In summary, the following problem areas were identified in the context of the evaluation of a model according to the intentions of the EU Directives:

- For the annual mean values, the current relative accuracy measure shifts the emphasis to the low concentration ranges where small absolute deviations can create large relative errors.
- For the hourly and daily mean values, the current relative accuracy measure shifts the emphasis to the very highest concentration ranges. In many cases, the model accuracy assessment is based on the highest observed concentration, which has significant potential to be an "outlier".
- There are no rules governing how many stations have to be used for the model evaluation.
- There are no rules on how to deal with the measurement uncertainty or how to determine the area representativeness of a measurement station to qualify it for use in a model evaluation.
- There is no rule of procedure for the (usual) case that a model does not fulfil the accuracy requirements at all selected stations for all relevant species and time intervals.

A more detailed guidance on the determination of model uncertainty should consider:

- whether the very low observed concentration levels should be excluded from a model evaluation exercise,
- whether the current error measure $rel\,max\,err\_p$ should be replaced by the alternative measure, $rel\,per\,err\_p_{LV}$, which defines the concentration difference at the percentile that corresponds to the allowed number of exceedances of the limit value normalized by the observed value at that percentile. This measure is more robust than the $rel\,max\,err\_p$ error measure, and also evaluates the model performance in the high concentration ranges, but without the sensitivity to possible very high concentration outliers. Given that this proposed alternative measure evaluates a model's accuracy in the significant concentration range of the limit values, a more direct and meaningful link to the EU Directives is achieved.

7 Literature

Binkowski, F. S. and Shankar, U. (1995), The Regional Particulate Matter Model, 1. Model description and preliminary results. J. Geophys. Res. 100, D12, 26191-26209

Carter, W. (1996). Condensed atmospheric photooxidation mechanisms for isoprene, Atmos. Environ. 30 (1996) 4275-4290.

Claiborn et al. (1998). Regional measurements and modeling of windblown agricultural dust: The Columbia Plateau PM10 program. J.G.R. VOL 103, D16, 19753-19767 .

Gong, S. L., Barrie, L. A. and Blanchet J.-P. (1997). Modelling sea-salt aerosols in the atmosphere. 1. Model development. J. Geophys. Res., 102, 3805-3818

Hass, H., Builtjes, P. J. H., Simpson, D., and Stern, R. (1997), Comparison of model results obtained with several European regional air quality models, Atmos. Env. 31 (1997), 3259-3279.

Hass, H., van Loon, M., Kessler, C., Matthijsen, J., Sauter, F., Stern, R., Zlatev, R., Langner, J., Fortescu, V., Schaap, M. (2003). Aerosol Modeling: Results and Intercomparison from European Regional-scale Modeling Systems. A contribution to the EUROTRAC-2 subproject GLOREAM. EUROTRAC report 2003.

Flemming, J. und Stern, R. (2004). Konzepte zur Formulierung von Kriterien zum Nachweis der von den EU-Richtlinien verlangten Genauigkeit von Modellrechnungen - Beispiele mit dem chemischen Transportmodell REM/CALGRID. Abschlussbericht im Rahmen des Forschungs- und Entwicklungsvorhaben 201 43 250 auf dem Gebiet des Umweltschutzes „Anwendung modellgestützter Beurteilungssyteme für die bundeseinheitliche Umsetzung der EU-Rahmenrichtlinie Luftqualität und ihrer Tochterrichtlinien“.

Flemming, J. (2003a). Klimatische Auswertung von modellierten und beobachteten Zeitreihen für Ozon, NO2, NO, SO2 und PM10 in Deutschland für die Jahre 1995-2001 und deren Charakterisierung nach Immissionsregimes mit einem Clusterverfahren, Abschlussbericht zum Forschungs- und Entwicklungsvorhaben 298 41 252 des Umweltbundesamts "Modellierung und Prüfung von Strategien zur Verminderung der Belastung durch Ozon.

Flemming, J. (2003b). Anwendung von Optimaler Interpolation und Kalman Filter Techniken für die Ausbreitungsrechnung mit dem chemischen Transportmodell REM/CALGRID, Abschlussbericht zum Forschungs- und Entwicklungsvorhaben 298 41 252 des Umweltbundesamts "Modellierung und Prüfung von Strategien zur Verminderung der Belastung durch Ozon

Hollingsworth, A. und Lönnberg, P. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data Part I: The wind field, Tellus, Bd. 38 A, S. 111-136.

Logan, J. (1998). An analysis of ozonesonde data for the troposphere, recommendations for testing 3-D models and development of a gridded climatology for tropospheric ozone, J. Geophys. Res. 104, 16, 1998

Loosmore and Hunt (2000). Dust resuspension without saltation. J. Geophys. Res. Vol 105, D16, 20,663 ff, August 27, 2000

Nenes, A., Pilinis, C., and Pandis, S. N. (1999). Continued Development and Testing of a New Thermodynamic Aerosol Module for Urban and Regional Air Quality Models, Atmos. Env. 33 , 1553-1560.

Reimer, E. and Scherer, B. (1992), An operational meteorological diagnostic system for regional air pollution analysis and long-term modelling, in: Air Pollution Modelling and its Applications IX., van Doop, H., ed.., Plenum Press 1992.

Roemer, M., Beekmann, M., Bergström, R., Boersen, G., Feldmann, H., Flatøy, F., Honore, C., Langner, J., Jonson, J., Matthijsen, J., Memmesheimer, J., Simpson, D., Smeets, P., Solberg, S., Stern, R., Stevenson, D., Zandveld, P. ,Zlatev, Z. (2003). Ozone trends according to ten dispersion models. A contribution to the EUROTRAC-2 subproject TOR. EUROTRAC report 2003.

Schell, B., I.J. Ackermann, H. Hass, F.S. Binkowski, A. Ebel (2001). Modeling the formation of secondary organic aerosol within a comprehensive air quality modeling system. J. Geophys. Res., 106, 28275 - 28293.

Simpson, D., A. Guenther, C.N. Hewitt, R. (1995). Steinbrecher: Biogenic emissions in Europe, 1. estimates and uncertainties, Journal of Geophysical Research, Vol. 100, No. D11, pp. 22,875-22,890

Stedman, John R., Bush, Tony J, Vincent, Keith J and Baggott, Sarah (2003). UK air quality modelling for annual reporting 2002 on ambient air quality assessment under Council Directives 96/62/EC and 1999/30/EC, report AEAT/ENV/R/1564 Issue 1.

Stern R. **(1994)**.Entwicklung und Anwendung eines dreidimensionalen photochemischen Ausbreitungsmodells, Meteorologische Abhandlungen Serie A, Band 8, Institut für Meteorologie der FU-Berlin.

Stern, R. (2004). Weitere Entwicklung und Anwendung des chemischen Transportmodells REM-CALGRID für die bundeseinheitliche Umsetzung der EU-Rahmenrichtlinie Luftqualität und ihrer Tochterrichtlinien. Abschlussbericht im Rahmen des Forschungs- und Entwicklungsvorhaben 201 43 250 auf dem Gebiet des Umweltschutzes „Anwendung modellgestützter Beurteilungssyteme für die bundeseinheitliche Umsetzung der EU-Rahmenrichtlinie Luftqualität und ihrer Tochterrichtlinien“.

Stern, R. Yamartino, R. J. and Graff, A. (2003). Dispersion Modelling within the European Community's Air Quality Framework Directive: Long Term Modelling of O3, PM10 and NO2. Presented at the 26th ITM on Air Pollution Modelling and its Application. May 26-30, 2003, Istanbul, Turkey

Stern, R. and R.J. Yamartino (2002). Development and first results of a new photochemical model for simulating ozone and PM-10 over extended periods. 25th Conference on Agricultural and Forest Meteorology, 20-24 May 2002 in Norfolk, VA

Stern, R. and R. J. Yamartino (2001). Dispersion Modelling within the European Community's Air Quality Framework Directive: Long Term Modelling of PM10, Fifth GLOREAM Workshop, Wengen, Switzerland, Sept. 24-26. (2001)
http://people.web.psi.ch/keller_j/GLOREAM/WS2001/WS01_frameset_proceedings.html

Walcek, C.J.(2000). Minor flux adjustment near mixing ratio extremes for simplified yet highly accurate monotonic calculation of tracer advection, J. Geophy. Res., 105, D7 (2000), 9335-9348.

Yamartino, R.J., J. Scire, G.R. Carmichael, and Y.S. Chang (1992). The CALGRID mesoscale photochemical grid model-I. Model formulation, Atmos. Environ., 26A (1992), 1493-1512.

8 Appendix: REM-CALGRID model overview

The REM-CALGRID model is an urban/regional scale model development designed to fulfil the requirements of the ambient air quality framework directive 96/62/EC of the European Commission (Stern et al., 2003). Rather than creating a completely new model, the urban-scale photochemical model CALGRID (Yamartino et al., 1992) and the regional scale model REM3 (Stern, 1994; Hass et al., 1997) were used as the starting point for the new urban/regional scale model, REM-CALGRID (RCG). The premise was to design an Eulerian grid model of medium complexity that can be used on the regional, as well as the urban, scale for short-term and long-term simulations of oxidant and aerosol formation.

The model includes the following features:

- A generalized horizontal coordinate systems, including latitude-longitude coordinates;
- A vertical transport and diffusion scheme that correctly accounts for atmospheric density variations in space and time, and accounts for all vertical flux components when employing either dynamic or fixed layers;
- A new methodology to eliminate errors totally from operator-split transport and ensure correct transport fluxes, mass conservation, and that a constant mixing ratio field remains constant;
- Inclusion of the recently improved and highly-accurate, monotonic advection scheme developed by Walcek (2000). This fast and accurate scheme has been further modified to exhibit even lower numerical diffusion for short wavelength distributions;
- Updated releases of the SAPRC-93 and CBM-IV photochemical reaction schemes including Carter's (1996) 1-product isoprene scheme and SO_2 oxidation to SO_4;

- Two equilibrium aerosol modules, that treat the thermodynamics of inorganic aerosols;
- An equilibrium aerosol module, that treat the thermodynamics of organic aerosols
- Simple modules to treat the emissions of sea salt aerosols and wind blown dust particles
- A simple wet scavenging module based on precipitation rates;
- An emissions data interface for long term applications that enables on-the-fly calculations of hourly anthropogenic and biogenic emissions, and greatly facilitates emissions reduction scenario studies.

RCG uses the bulk approach for the aerosol modelling. Aerosol dynamics are not considered. The model employs two different equilibrium aerosol modules: a) The MARS-A module (Binkowski and Shankar, 1995) that treats the thermodynamics of the inorganic sulfate, nitrate, ammonium aerosols and water, and b) the ISORROPIA module that additionally treats sodium and chloride aerosols (Nenes et al., 1999). Secondary organic aerosols (SOA) that are formed by condensation of biogenic and anthropogenic hydrocarbon oxidation products are considered employing the SORGAM module developed by Schell et al. (2001).

The sea-salt aerosol emissions are parameterized according to Gong et al. (1997) as a function of size and wind speed. For the calculation of land-use dependent wind blown dust emissions the following release mechanisms are treated: Direct release of small dust particles by the wind (Loosmore and Hunt, 2000), and indirect release by collisions with bigger soil grains, that are lifted by the wind but return to the surface because of their weight („saltation“ process, Claiborn et al.,1998).

The aerosol components are assigned to two size fractions: fine mode PM2.5 and coarse mode PM10-PM2.5. The anthropogenic emissions data base has to provide the primary emissions split into these two size categories. The anthropogenic PM emissions are allocated to the following model species: The coarse mode species $PMCO_{prim}$ = PM10-PM2.5 and the fine mode species EC (Elemental Carbon), OC (Organic Components) and MC (Mineral Components). The anthropogenic fine mode emissions EC, OC and MC are derived from total PM 2.5 emissions employing source-group-dependent split factors. The photochemical mechanisms include the oxidation of SO_2 to gaseous sulphate, which is assumed to be gaseous sulphuric acid. The aerosol modules treat the thermodynamics of the inorganic aerosols, depending on the gas phase concentrations of nitric acid, ammonia, sulphuric acid, HCl, humidity and temperature. In RCG, all secondary aerosols are assigned to the PM2.5 fraction. All aerosols are transported and subjected to wet and dry deposition. PM10 concentrations are then defined as the sum of primary PM10 and secondary organic and inorganic aerosols via the relationship:

$$PM10 = EC+OC+MC+PMCO_{prim}+SO_4+NO_3+NH_4+SOA+Na+Cl+WBDF+WBDC$$

WBDF and WBDC are the wind blown dust particles in the fine and coarse mode, respectively, and are calculated on-the-fly. RCG uses a resistance-based model (aerodynamic resistance, viscous sub-layer resistance, surface resistance) for the computation of dry deposition rates as a function of geophysical parameters, micrometeorological conditions, and gaseous or particle pollutant parameters including the gravitational settling speed of particles. Species dependent deposition velocities are calculated for each land-use class within each grid cell, and deposition fluxes are computed by summing over the fractional land-use terms.

Meteorological data needed by RCG at hourly intervals consist of layer-averaged gridded fields of wind, temperature, humidity and density, plus 2-d gridded fields of mixing heights, several boundary layer and surface variables, precipitation rates and cloud cover. All this meteorological data is produced employing a diagnostic meteorological analysis system based on an optimum interpolation procedure on isentropic surfaces. The system utilizes all available synoptic surface and upper air data (Reimer and Scherer, 1992).

The RCG model also requires annual emissions of VOC, NO_x, CO, SO_2, CH_4, NH_3, PM10, and PM2.5, split into point and gridded area sources. Mass-based, source group dependent NMVOC profiles are used to break down the total VOC into the different species classes of the chemical mechanisms. Hourly emissions are derived during the model run using source-group dependent, month-of-the-year, day-of-week and hour-of-the-day emissions factors. Biogenic VOC-emissions are derived using the E94 emissions factors for isoprene and OVOC (Other VOCs) as described in Simpson et al. (1995). Terpene emissions factors are taken from the CORINAIR emission hand-book. These biogenic calculations are based on the land-use data for deciduous, coniferous, mixed forests and crops. Light intensity and temperature dependencies are also considered.

Monthly varying lateral and top boundary conditions for ozone are taken from climatological background data (Logan, 1998). Boundary data for all other species are chosen as typical background values.