Atmos. Chem. Phys., 13, 7153–7182, 2013
www.atmos-chem-phys.net/13/7153/2013/
doi:10.5194/acp-13-7153-2013

Atmospheric Chemistry and Physics
Open Access

*E pluribus unum**: ensemble air quality predictions

S. Galmarini[1], **I. Kioutsioukis**[1,2], **and E. Solazzo**[1]

[1]European Commission, Joint Research Center, Institute for Environment and Sustainability, Ispra (VA), Italy
[2]Region of Central Macedonia, Thessaloniki, Greece

*****One out of many

Correspondence to: S. Galmarini (stefano.galmarini@jrc.ec.europa.eu)

Received: 22 November 2012 – Published in Atmos. Chem. Phys. Discuss.: 8 January 2013
Revised: 27 March 2013 – Accepted: 30 May 2013 – Published: 29 July 2013

Abstract. In this study we present a novel approach for improving the air quality predictions using an ensemble of air quality models generated in the context of AQMEII (Air Quality Model Evaluation International Initiative). The development of the forecasting method makes use of modelled and observed time series (either spatially aggregated or relative to single monitoring stations) of ozone concentrations over different areas of Europe and North America. The technique considers the underlying forcing mechanisms on ozone by means of spectrally decomposed previsions. With the use of diverse applications, we demonstrate how the approach screens the ensemble members, extracts the best components and generates bias-free forecasts with improved accuracy over the candidate models. Compared to more traditional forecasting methods such as the ensemble median, the approach reduces the forecast error and at the same time it clearly improves the modelled variance. Furthermore, the result is not a mere statistical outcome depended on the quality of the selected members. The few individual cases with degraded performance are also identified and analysed. Finally, we show the extensions of the approach to other pollutants, specifically particulate matter and nitrogen dioxide, and provide a framework for its operational implementation.

1 Introduction

Multi-model ensembles (MME) is the practice according to which results obtained from a somehow arbitrary collection of modelling systems and applied to a common case study, are statistically treated in an attempt to capture more effectively the variability of the observational data and to improve the final results (e.g., Galmarini et al., 2004; Knutti et al., 2010; Pirtle et al., 2010). The practice has been used in a wide range of applications in atmospheric and climate sciences (Galmarini et al., 2001; delle Monache et al., 2006; McKeen et al., 2005; Van Loon et al., 2007; Mallet and Sportisse, 2006; Solazzo et al., 2012a; Riccio et al., 2012; Potempski et al., 2008; Knutti et al., 2010; Tebaldi and Knutti, 2007) as well as in a range of other contexts. Over the years a large number of different approaches (Potempski and Galmarini, 2009) have been proposed from the very popular simple averaging of the result, to the construction of the median model to the application of weights derived from past skill scores or Bayesian model averaging theory (e.g., Delle Monache et al., 2006; Galmarini et al., 2004; Potempski et al., 2010; Riccio et al., 2007). In all of the aforementioned examples, MME members have been used in an all-or-nothing fashion, by considering the various model results as a complete representation of the processes or by modulating their contribution to the average by means of weights. In all those practices the model results are taken as they are, without any consideration of the reasons why a model is better than others and taking the results with all the good aspects as well as bad ones. This approach to ensemble analysis is motivated by the illusory conception that the statistical treatment would account for the process variability and by the fatal assumption that model results are independent. As illustrated by Potempski and Galmarini (2009) this assumption is unrealistic from the start and as demonstrated by Solazzo et al. (2012a) can also lead to a deterioration of the ensemble result as the number of models increases. Recent findings point toward a deeper and more thorough analysis of the model results in an attempt to identify those that, within the ensemble, represent real original contributions to the improvement of the ensemble result. Toward this end, analyses aiming at promoting true model diversity such those used by Riccio et al. (2012), Solazzo et al. (2012a), Masson and Knutti (2011) seem to go in that right direction.

Most recently, Tchepel et al. (2012) have applied a Kolmogorov-Zurbenko filter (KZ-f) (Zurbenko, 1986) to a set of an ensemble of model results to identify the capacity of the different models to simulate the various scales in which the modelled ozone time series was decomposed. Such an approach has led to the determination of weighting factors to be associated with the various models performance in the construction of the ensemble output. The originality of this approach remains in the fact that a deeper analysis of the model performance than the operational comparison model-observations (Rao et al., 2011) has been selected as discriminant in determining the role of members within an ensemble.

In the present study, we intend to take a step forward with respect to the ensemble screening and model selection, having as the final goal not only the improvement of the ensemble result on hindcast application, but also the forecasting capacity of a MME for air quality applications. The intent is to extract from an ensemble of models the best spectral components to construct a new set of results that is expected to behave better than the ensemble members rather than to use the KZ-f analysis to identify in a diagnostic way the relative contribution of all models to the final ensemble result. KZ-f will be used to dissect each model result, extract the "best components", and re-assemble them in a new set of model results. In our work KZ-f is, at all counts, an operator by which a new model set is constructed, and not just a diagnostic tool used to identify the best model. Hence, the KZ-f generated set can be seen as the outcome of a new model and not a combination of existing weighted results as in the work of Tchepel et al. (2012). In this respect, the ensemble of models still represents a pool of realisations from which, however we do not extract blindly a statistically treated result, but from which we try at best to use the best of the available information. The ensemble is therefore exploited as the set of all available information from which we expect to extract what we need, all model results are necessary a priori, but only few will be used in the end.

MME for air quality forecast is used operationally in some context like Global and regional Earth-system (Atmosphere) Monitoring using Satellite and in-situ data (GEMS) and Monitoring Atmospheric Composition & Climate (MACC) (http://www.gmes-atmosphere.eu/). In MACC, air quality predictions at the regional scale produced by several European institutions are gathered and treated in a classical ensemble fashion (Peuch et al., 2011). As stated in the review paper by Kukkonen et al. (2012): "The current operations in the GEMS and MACC projects have used a more elaborate ensemble technique, based upon the differential weighting of the individual models according to their skill over the last few days. However, a long-term improvement in Chemical Weather Forecast performances will be based on the improvement of individual models and their representation of dynamical, physical and chemical processes." While we completely agree with the final statement, we also feel that quite a lot can still be extracted from the state-of-the-art AQ models even when used in forecast mode and in the current state of development. This would not hold true for all pollutants with the same level of accuracy, but the ensemble practice and model improvement can still proceed in parallel producing interesting and relevant results. Ensemble results can still be improved using the current model predictions and a novel methodology is proposed here. The latter can be implemented straight forwardly as long as time series from several model results are available. The technique can be easily implemented and provides an important enhancement in the predicting capability of modelled ozone.

The present study will take advantage of the large selection of model results produced for the Air Quality Model Evaluation International Initiative (AQMEII) (Galmarini et al., 2012a; Rao et al., 2011). The initiative aimed at collecting regional scale air quality model results applied for the year 2006 to Europe and North America.

The paper is structured as follows: in Sect. 2 the technique is outlined to give a bird's eye view of the model treatment; in Sect. 3 the case study used to develop and test the technique is presented and in addition, the monitoring and simulated data are analysed from the spectral and the KZ-f view point; in Sect. 4 the results of the application of the forecasting technique are presented. Last, some final considerations are drawn in Sect. 5.

2 Methodology

2.1 The Kolmogorov–Zurbenko filter

The Kolmogorov–Zurbenko filter (Zurbenko, 1986) was first proposed by Kolmogorov and formalised later by Zurbenko. It is defined as an iteration of a moving average filter applied on a time-series $S(t)$:

$$\mathrm{KZ}_{m,p} = R_{i=1}^{p}\left\{ J_{k=1}^{W_i}\left[\frac{1}{m} \sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}} S\left(t_i\right)_{k,j} \right]\right\} \begin{cases} R: \text{iteration} \\ J: \text{running window} \\ W_i = L_i - m + 1 \\ L_i = \text{length of } S(t_i) \end{cases} \tag{1}$$

It is a two-parameter filter controlled by the window size (m) and the number of iterations (p). The KZ-f removes high-frequency variations from the data (with respect to the window size) and belongs to the class of low-pass filters (since it filters periods smaller than the selected cut-off period). By modifying the controlling parameters (m, p), different scales of motion can be eliminated and others retained. In particular, by taking the difference between two KZ-f corresponding to different parameters (m, p), a band-pass filter is created.

The applications of KZ-f in the field of chemical weather is expanding and includes, among others, the diagnosis of the

Table 1. Definition of time scales.

Component	From period	To period	Atmospheric Processes that contribute to O_3 fluctuations
Intra-day (ID)	…	12 h	Fast-acting local scale processes
Diurnal (DU)	12 h	2.5 d	Diurnal (day vs. night) processes
Synoptic (SY)	2.5d	21d	Changing weather patterns
Long-term (LT)	21 d	…	Slow-acting processes

meteorological and air quality measurements and model results (Rao et al., 1997; Hogrefe et al., 2000), the diagnosis of trends (Wise et al., 2005; Papanastasiou et al., 2012) and the bias adjustment of ozone forecasts (Kang et al., 2008). The filter has been proven in several occasions to be capable of capturing the fundamental time scales of regional models without having to perform a full Fourier analysis. For the case of ground-level ozone, four separate scales of motion have been defined relevant, detected by means of physical considerations and periodogram analysis (Rao et al., 1997). They are namely the intra-day component (ID), the diurnal component (DU), the synoptic component (SY) and the baseline or long-term component (LT). The hourly time series of ozone can therefore be decomposed as:

$$\mathrm{S}(t) = \mathrm{ID}\,(t) + \mathrm{DU}(t) + \mathrm{SY}(t) + \mathrm{LT}(t) \tag{2}$$

$$\mathrm{ID}\,(t) = S\,(t) - \mathrm{KZ}_{3,3} \tag{3}$$

$$\mathrm{DU}\,(t) = \mathrm{KZ}_{3,3} - \mathrm{KZ}_{13,5} \tag{4}$$

$$\mathrm{SY}\,(t) = \mathrm{KZ}_{13,5} - \mathrm{KZ}_{103,5} \tag{5}$$

$$\mathrm{LT}\,(t) = \mathrm{KZ}_{103,5} \tag{6}$$

Table 1 summarises the periods associated to the components and the parts of the time series spectra they represent. We shall further notice that the separation of scales does not imply independence neither between the processes within each scale nor among the four spectral components. In other words, the KZ-f does not ideally separate the spectral components, but there is some interaction especially for the neighbour components (Hogrefe et al., 2003). The total error of the decomposed by Eq. (2) time-series is propagated through the spectral components and takes the form:

$$\begin{aligned}
&\mathrm{RMSE}^2\,(\mathrm{O_3}) = \mathrm{error(ID)} + \mathrm{error(DU)} + \mathrm{error(SY)} + \mathrm{error(LT)} \\
&= \mathrm{RMSE}^2(\mathrm{ID}) + \langle \Delta\mathrm{ID} * \Delta\mathrm{DU}^\mathrm{T}\rangle + \langle \Delta\mathrm{ID} * \Delta\mathrm{SY}^\mathrm{T}\rangle + \langle \Delta\mathrm{ID} * \Delta\mathrm{LT}^\mathrm{T}\rangle \\
&+ \mathrm{RMSE}^2(\mathrm{DU}) + \langle \Delta\mathrm{ID} * \Delta\mathrm{DU}^\mathrm{T}\rangle + \langle \Delta\mathrm{DU} * \Delta\mathrm{SY}^\mathrm{T}\rangle + \langle \Delta\mathrm{DD} * \Delta\mathrm{LT}^\mathrm{T}\rangle \\
&+ \mathrm{RMSE}^2(\mathrm{SYD}) + \langle \Delta\mathrm{ID} * \Delta\mathrm{SY}^\mathrm{T}\rangle + \langle \Delta\mathrm{SY} * \Delta\mathrm{SY}^\mathrm{T}\rangle + \langle \Delta\mathrm{SY} * \Delta\mathrm{LT}^\mathrm{T}\rangle \\
&+ \mathrm{RMSE}^2(\mathrm{LT}) + \langle \Delta\mathrm{ID} * \Delta\mathrm{LT}^\mathrm{T}\rangle + \langle \Delta\mathrm{LT} * \Delta\mathrm{SY}^\mathrm{T}\rangle + \langle \Delta\mathrm{ID} * \Delta\mathrm{LT}^\mathrm{T}\rangle
\end{aligned} \tag{7}$$

where Δ denotes the difference between the observed and modelled component, "$*$" denotes the matrix multiplication and T is the transpose operator. The error from each spectral component consists of four error terms: the component contribution (diagonal terms) and its interaction with the other components (off-diagonal terms). The magnitude of the covariance terms of the error matrix determines the degree of association of the spectral components derived from the KZ-f.

2.2 The proposed ensemble strategy and the kz model

The methodology we put into place is explained as follow. Equal lengths of the observed and the time series of ozone obtained from all ensemble members are decomposed into four components by the KZ-f. The modelled spectral components are evaluated against the observed ones and the models producing each one of the four best components are identified. Then, future time series (i.e., a time series with the same length as the historic time series that is shifted to include a future horizon) of the identified models are KZ-f decomposed and for each spectral component the respective one is taken. Finally, a new model (kz model) is built by adding the respective future components. For the historic period, if the spectral components were independent (i.e., the off-diagonal terms of the covariance matrix would be zero), the kz model skill would outperform any other model skill according to Eq. (4). However, since the components are not independent and in addition, the interest lies in the forecast period (that is, kz forecast skill), the idea needs to be evaluated.

Hence, the technique that is proposed is based on the following simple ingredients:

- A time series of ozone measurements at station level or aggregated at regional or sub-regional scale and results from a multi-model ensemble are required.

- The model results can be multi-model in the widemost sense also using different emission inventories or boundary conditions.

- Model results should be available for a minimum of 3 months plus a week of prediction.

Given these elements, the following steps are then taken:

Hindcast step: H-Step

1. Three months (past period: from t_0 -90 days to t_0) of measurements are decomposed according to the KZ-f in the four modes listed in Table 1;

Table 2. Hindcast Ranking (provision for forecast) versus Forecast Ranking (real).

	EXTRACT week (hindcast)				PREDICT week (forecast)			
Model Rank with respect to the component RMSE	ID	DU	SY	LT	ID	DU	SY	LT
1st	3	8	2	12	3	9	9	12
2nd	7	5	5	1	7	8	5	7
3rd	13	3	8	7	6	12	6	1
4th	11	9	7	10	12	4	2	4
5th	6	7	4	4	4	6	11	3
6th	8	6	1	13	8	3	7	6
7th	2	13	9	11	13	5	12	10
8th	5	2	11	6	2	7	10	13
9th	1	11	6	3	11	2	13	11
10th	12	4	13	9	9	10	4	2
11th	9	10	12	2	1	11	1	8
12th	10	1	3	8	10	13	8	9
13th	4	12	10	5	5	1	3	5

2. The individual ensemble members results for the same three months period are also decomposed with KZ-f;

3. The four spectral time series derived from each member are compared with the measurements four spectral time series, respectively, to identify the best match. The best match is based on standard statistical indicators over the last week (from t_0 −7 days to t_0) such as the RMSE;

Forecast step: F-Step

1. The four spectral modes from the best-match models of the previous step are recalculated over a period of equal length that incorporates a forecast week (from t_0 −83 days to $t_0 + 7$ days) and recombined in what is defined here as the kz model which constitutes a brand new model set and the result of the ensemble analysis;

2. The prediction for the coming week (from t_0 to $t_0 + 7$ days) of the kz model are used as a forecast and compared with measurements (when available);

3. A new iteration is generated by shifting the time series window (from t_0 −90 days to $t_0 + 7$ days) by one day;

The novelty of this approach remains is that the ensemble result is no longer a mere statistical treatment of the outcome of model results, but it is diagnosed in the fundamental aspects that constitute each member which are then *re-ensembled* to constitute the only model set used for forecasting.

The technique presented above has been applied to the AQMEII phase 1 (Rao et al., 2011) case study as described in the next section. In the case study, the one year of simulation and data (for 2006) have been used in blocks of weekly forecast condition for the period from 1 April to 30 September (ozone reporting period). The total number of iterations fitting in this period is 175. Therefore, we have applied and tested the methodology over a total of 175 weeks forecast. Figure 1 provides the scheme of how the technique was used at each iteration.

For the sake of a better explanation of the methodology, we present the calculated four spectral components of the observations and all deterministic models, using a three-month time-series (from t_0 – 90 days to t_0). The one presented here is a single case extracted from the available data; we postpone to later the statistical evaluation of all examined cases. In Fig. 2 (left column) the calculation of the models four components of the signal together with those of the measurements over the period t_0 −90, t_0 is shown. Figure 2 (middle column) zooms into the last week where the determination of the models producing the best four components takes place (step 3). The results of step 4 are shown in Fig. 2 (right column) where the kz model is applied to the forecast week. In the same figure, we also plot the real (in red) best components of the forecast week, after validation with the observed components. As shown in the figure for ID and LT the models selected were the same, whereas for the other two components they turned out to be different. The differences between the components are marginal, however the diurnal variation for the daily signal and the bell shape for the synoptic are nicely captured by the identified kz model.

Table 2 shows the difference in model performance for the past (t_0 −7, t_0) and future (t_0, t_0 +7) week and the role of the models in determining the various components for these periods. As shown above, the kz model was obtained in the past week from models 3, 8, 2 and 12, and for the future week the best performance was obtained by 3, 9, 9 and 12. The table shows that the forecast was made with a suboptimal spectral component quartet (ID, DU, SY, LT) with rankings of 1, 2, 4 and 1, respectively. However, even in this case, the kz model outscores any other model (presented in the

Table 3. Participating models and their features. The presented order does not correspond to the numbers presented in the figures/tables as the models are used anonymously throughout the text.

Domain	Model		Res (km)	No. Vertical layers	Emissions	Chemical BC
	Met	AQ				
European	MM5	DEHM	50	29	Global emission databases, EMEP	Satellite measurements
	MM5	Polyphemus	24	9	Standard[1]	Standard
	PARLAM-PS	EMEP	50	20	EMEP model	From ECMWF and forecasts
	WRF	CMAQ	18	34	Standard[1]	Standard
	WRF	WRF/Chem	22.5	36	Standard[1]	Standard
	WRF	WRF/Chem	22.5	36	Standard[1]	Standard
	ECMWF	SILAM	24	9	Standard anthropogenic In-house biogenic	Standard
	MM5	Chimere	25	9	MEGAN, Standard	Standard
	ECMWF	Lotos-EUROS	25	4	Standard[1]	Standard
	COSMO	Muscat	24	40	Standard[1]	Standard
	MM5	CAMx	15	20	MEGAN, Standard	Standard
	GEM	GEM-AQ	25	28 (up to 10 mb)	Standard over AQMEII region; Global EDGAR/ GEIA over the rest of the global domain	Global variable grid setup (no boundary conditions)
	COSMO-CLM	CMAQ	24	30 (up to 100 hPa)	Standard[1]	Standard
North[2] American	GEM	AURAMS	45	28	Standard[3]	Climatology
	WRF	Chimere	36	9	Standard	LMDZ-INCA
	MM5	CAMx	24	15	Standard	LMDZ-INCA
	WRF	CMAQ	12	34	Standard	Standard
	WRF	CAMx	12	26	Standard	Standard
	WRF	Chimere	36	9	Standard	standard
	MM5	DEHM	50	29	global emission databases, EMEP	Satellite measurements
	COSMO-CLM	CMAQ	24	30 (up to 100 hPa)	Standard	Standard

[1] Standard anthropogenic emission and biogenic emission derived from meteorology (temperature and solar radiation) and land use distribution implemented in the meteorological driver (Guenther et al., 1994; Simpson et al., 1995).
[2] Standard inventory for NA includes biogenic emissions (see text).
[3] Standard anthropogenic inventory but independent emissions processing, exclusion of wildfires, and different version of BEIS (v3.09) used.

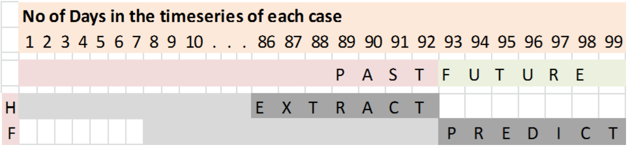

Fig. 1. Chart on computational strategy. Each examined case consists of two steps, an H-step and an F-step. H-step: H denotes the past period (last three-month time-series: from t_0 −90 days to t_0) where each modelled time-series is decomposed into its spectral components and EXTRACT denotes the last 7-day period (from t_0 −7 days to t_0) of H where the spectral ensemble is validated against the observed spectral components, with respect to the RMSE, to identify the models that produced the optimal spectral components. F-step: F denotes the shifted-by-one-week-period including a forecast week at the end (from t_0 −83 days to t_0 +7 days) where the spectral components of the model id's identified in the H-step of the case are re-calculated and summed up (kz model) and finally the kz model prevision during the PREDICT week (from t_0 to t_0 + 7 days) is validated against the observations over the future 7-day period.

next paragraph). This shows how the methodology captures in essence the model behaviours and is conservative with respect to the quality of the results.

Finally, to conclude the explanatory part of the technique, in Fig. 3 the kz model ozone time series for one of the 175 weeks of forecast is shown, as example (Table 2). The panels show the individual model results (panel 1–13) together with the results of the kz model as well as the following ensemble products:

- the median model (mm): defined as the median values obtained considering the complete distribution of model results
- the spectral median model (sm): the model assembled by combining the components (ID, DU, SY and LT) like in the case of the kz model the difference being that the four selected in this case are the median value of all model components rather than the best

We will refrain here from judging the quality of this result, though apparent, postponing it to a systematic analysis of the quality of the methodology to Sect. 4.

In the sections that follow, we investigate and address the questions:

- Do the observed and modelled KZ-f decomposed time-series have similar properties?

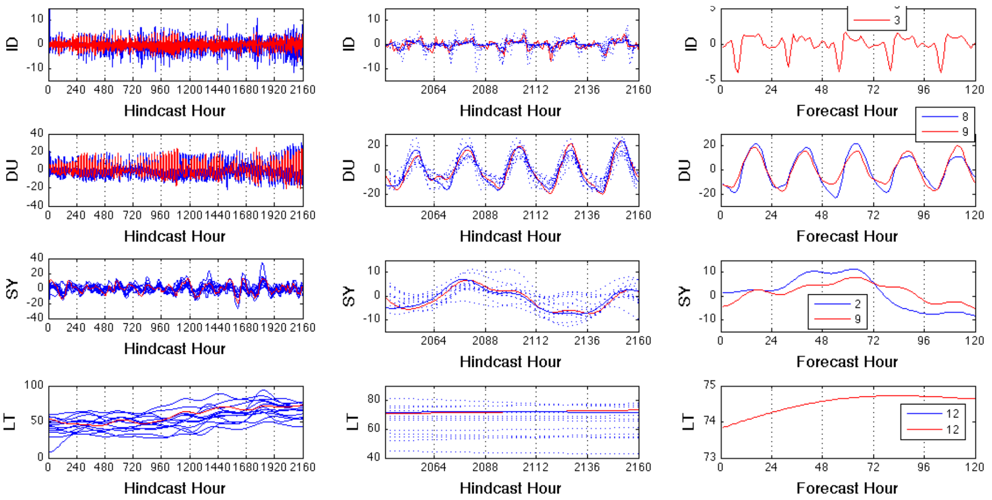

Fig. 2. Illustrative example of the computational strategy for one case (of the 175). Left column: the spectral components of the observations and all deterministic models (H-step). Middle column: the spectral components of the observations and all deterministic models during the EXTRACT week (H-step). In this example, the model id's that produced the least RMSE (shown in thick blue) in the spectral components are: ID(3), DU(8), SY(2), LT(12). Right column: the spectral components of the deterministic models identified in the EXTRACT week are re-calculated for the F period and shown for the PREDICT week (in blue) together with the actual (in red) optimal components (F-step).

Table 4. The characteristics of the working domains.

Sub-Region	Longitude		Latitude		Ensemble Members	Number of receptors in the aggregation		
	from	to	from	to		U	S	R
EU1	-10	5	42	60	13	205	117	85
EU2	5	25	46	56	13	202	176	260
EU3	7	15	43	46	13	47	19	24
EU4	−2	22	37	42	13	14	25	29
NA1	−125	−112	31	42	8	45	79	59
NA2	−104	−90	25	37	8	22	52	37
NA3	−85	−69	36.5	48.5	8	38	53	80

- Which spectral component dominates the error?
- Can the best spectral components be forecasted from a multi-model ensemble?

3 The case study: observations and the ensemble members

3.1 The data and study domains

The test case for the kz model is ozone simulation at a regional scale over four European and three North-American sub-regions, and uses the outcomes of the AQMEII activity (Rao et al., 2011), as mentioned in the introduction.

AQMEII was started in 2009 as a joint collaboration of the EU Joint Research Centre, the US-EPA and Environment Canada with the scope of bringing together the North American and European communities of regional scale air quality models. Within the initiative the two-continent model evaluation exercise was organised which consisted in having the two communities to simulate the air quality over north America and Europe for the year 2006 (full detail in Galmarini et al., 2012b). Data of several natures were collected and model evaluated (Galmarini et al., 2012c). The community of the participating models is presented in Table 3, which forms a multi-model set in terms of meteorological driver, air quality model, emission and chemical boundary conditions. The models of Table 3 have been subject of evaluation against measurements in terms of individual model (model-to-observation) as well as of ensemble (ensemble-to-observation) comparison, for a range of pollutants and meteorological fields (Solazzo et al., 2012a, b; Vautard et al., 2012). The model settings and input data are described in detail in Solazzo et al. (2012a, b), Schere et al. (2012), Pouliot et al. (2012), where references about model development and history are also provided.

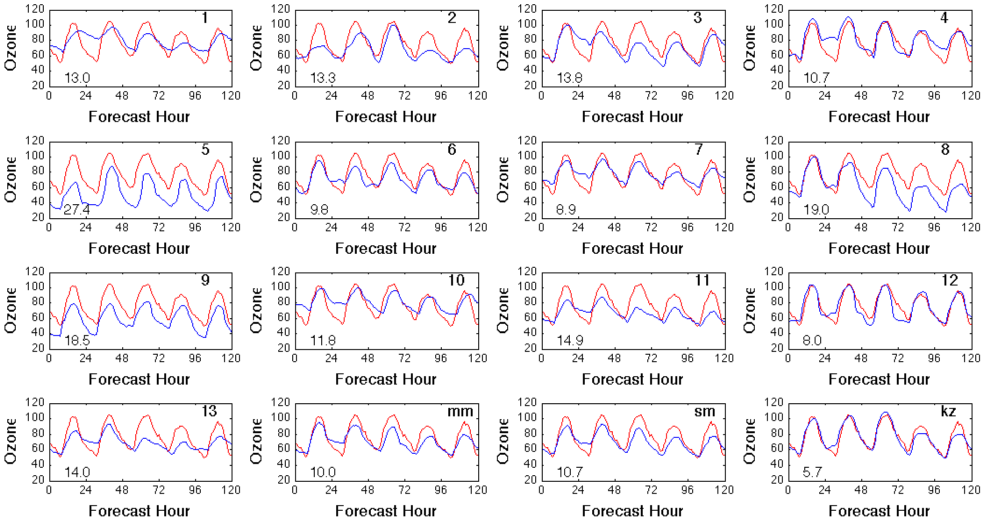

Fig. 3. The skill of the weekly provision of the kz model is validated, with respect to the RMSE, against the observations over the PREDICT period (from t_0 to t_0 + 7 days). Numbers 1–13 correspond to the id of the deterministic models, mm is the median model and sm is the spectral median model.

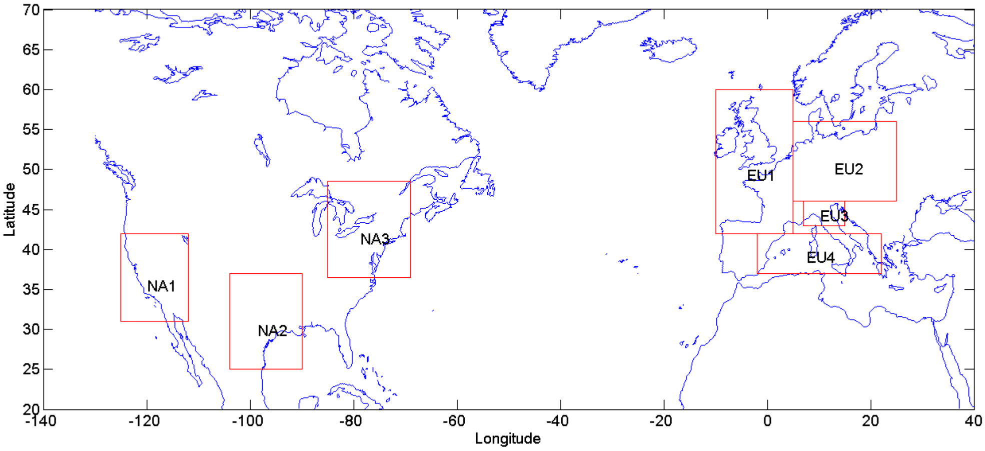

Fig. 4. Visualisation of the working domains.

The European and North American sub-regions used for analysis are shown in Fig. 4, and extensions are given in Table 4, where the number of the selected monitoring stations (selection criterion: availability of at least 75 % of measurements over the analysed period, grouped according to rural, urban, and sub-urban categories, as described by the metadata provided by the monitoring networks (Solazzo et al., 2012c)) are also reported. These regions were chosen to correspond to those used in the other AQMEII evaluation studies dealing with the ensemble of models of Table 3. They represent a variety of conditions in terms of emissions (Pouliot et al., 2012), weather (Vautard et al., 2012), chemical regimes (Solazzo et al., 2012a, b), boundary conditions (Schere et al., 2012) that constitute an important bench test for the technique proposed. The hourly time series for each working domain have been generated as the spatial average of the model outputs interpolated at each receptors/grid points. The evaluation period (forecast mode) is from 1 April to 30 September 2006, for a total of 4392 h). An analysis of the kz model performance will also be presented at individual stations.

3.2 Extraction and analysis of the temporal components of ozone: observations

The analysis of the observations starts from a detailed Fourier transformation to which the KZ-f will be associated in an attempt to identify the relevance of the components splitting in the power spectrum. We analyse hourly data over a 6-month period. Hence, the resolved periods range from 2 h to 60–90 days. The results presented here relate to EU1 only but also apply to all other sub-regions. The power spectrum of the

Table 5. Identification numbers of the deterministic models contributing most frequently to the kz model.

	URBAN				SUBURBAN				RURAL			
	ID	DU	SY	LT	ID	DU	SY	LT	ID	DU	SY	LT
EU1	3	8	5	8	7	8	5	8	3	8	5	4
EU2	3	5	5	5	3	5	5	8	6	5	5	4
EU3	10	5	7	8	1	5	12	13	1	5	12	6
EU4	7	8	12	9	3	8	12	12	7	6	12	3
NA1	7	2	7	2	7	2	3	2	6	3	3	2
NA2	8	5	1	5	1	5	1	1	1	1	1	1
NA3	2	5	1	5	2	5	1	5	2	2	1	8

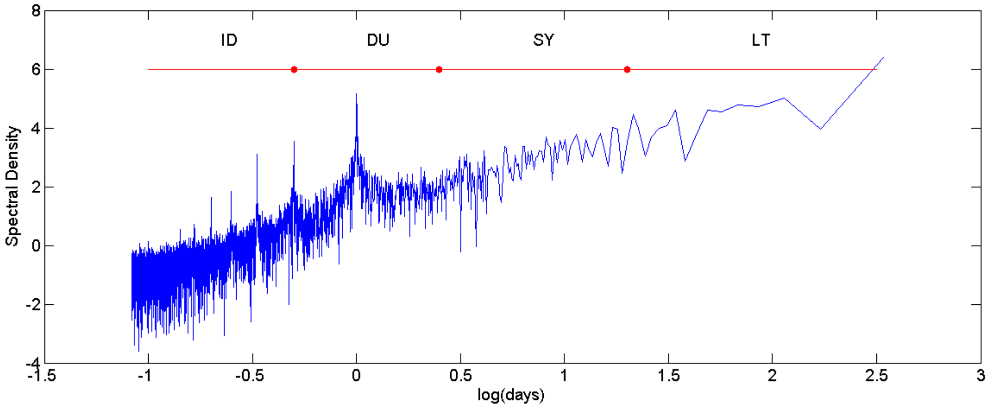

Fig. 5. Periodogram of the observed ozone concentrations (aggregated over rural stations) in the EU1 subdomain.

observations (Fig. 5) indicates that the largest forcing in the time series has a 24 h period (diurnal range). Other frequencies with high energy lay in the intra-day, synoptic and long-term range. Many peaks are particularly evident for small periods, with the most intense at 12:00 and 08:00 LT. Those peaks clearly identify an intra-day and a diurnal cycle. A synoptic cycle is also added to the analysis, to distinguish the changing weather patterns from the slow acting processes. As explained in Sect. 2, the selected cycles and their physical interpretation are given in Table 1. Clearly the forcing identified by the power spectrum relates to the periodicity of the meteorological phases that regulate the dispersion of the emissions in the boundary layer and the exchanges from the latter to the free atmosphere. Superimposed to that, the large scale forcing which relates to the transport of ozone from other areas according to the timescale represented on the time axis. As typically occurs with scalar tracers (gasses, heat and moisture) (Galmarini et al., 2000) the power spectrum shows monotonically increasing variance for large scales indicating the absence of clear scale separation between the synoptic, meso, and boundary layer scales as it happens for dynamic variables (e.g., the vertical velocity). This is due to concurring contributions of processes of different nature and scale. At short scale, the diurnal variation of the boundary layer growing and collapsing regulates most of the variance and the inter-diurnal variability of emission precursors can also be a contributing factor to the determination of the total variance.

As explained earlier, the spectral components are composed by three signals with zero mean (ID, DU and SY) and one slow varying signal (LT). The ID, DU, and SY signals are zero mean fluctuations about the smoothed time series (LT). In terms of their relative strength, the amplitude of the zero mean signals is highest for the DU component and lowest for the ID component. Figure 6 shows how the variance is distributed across the four components of the KZ-f measurements aggregated in the seven sub-regional domains over the two continents. The variance distribution has been calculated for urban, sub-urban and rural monitoring stations. The total explained variance from the four (ID, DU, SY, LT) spectral components (single contributions + interactions) identified generally similar importance rankings across the sub-regions and aggregation types. From Fig. 6 it can be inferred the following:

- The DU component drives the ozone variability and accounts for more than half of its variance. Generally, its importance is weakened (but still dominating) moving from the urban to the rural aggregation possibly due to the reduction in photochemical activity.

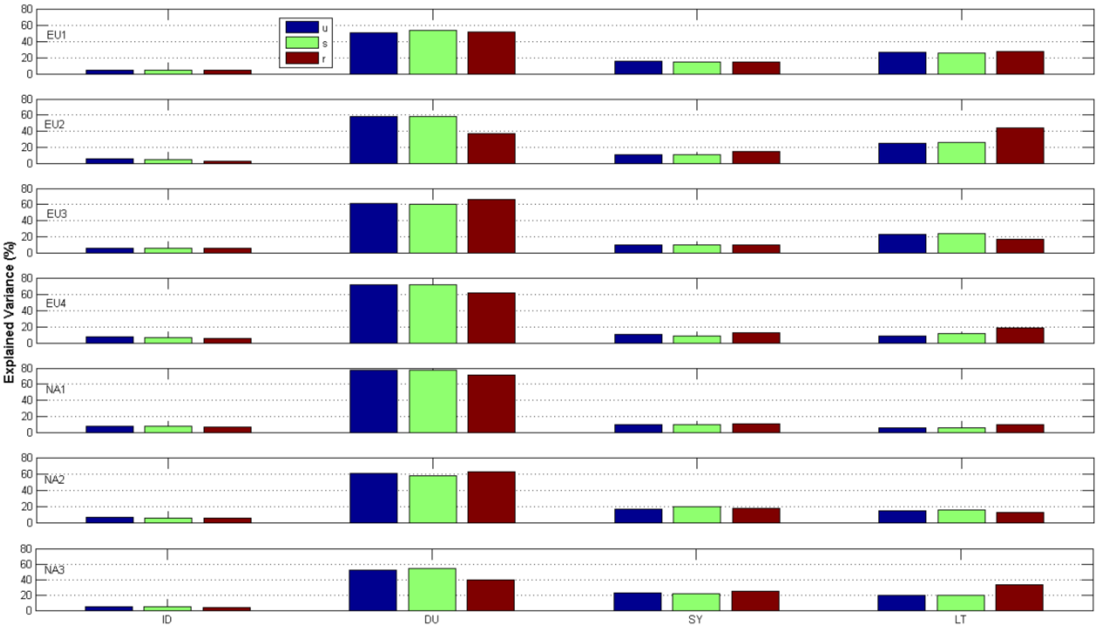

Fig. 6. The explained variance from the four (ID, DU, SY, LT) spectral components (single contributions + interactions) for the observation time series, for all seven subdomains and three ozone aggregation types.

- The LT and SY components are ranked in the 2nd and 3rd position in terms of their explained variance. The SY component has a directional dependence that is generally stronger in the NA and weaker in the EU. The opposite is true for the LT component that is probably explained by the existence of a prevailing direction for the large-scale transport patterns.

- The ID component explains the least amount of the ozone variability due to the small magnitude of its fluctuations.

The explained variance from the single contributions of the four components accounts for roughly 80 % of the total variability, implying an imperfect separation and higher-order interactions between the different scales. Specifically, the explained variance by single contributions of the four spectral components accounts for approximately 74–81 % of the total variance lumping the rest to the interactions between the components. Although different sets of sub-region specific parameters for the KZ-f optimised the explained variance, for the sake of comparison the same values identified by Hogrefe et al. (2000) were selected and applied to all sub-regions in Europe and North America.

Finally, the analysis of the magnitude of the spectral components with respect to the ozone levels (not shown) yielded results similar to Hogrefe et al. (2000). In particular, the probability of high ozone concentrations is related to:

i. increased variability in the ID and DU components;

ii. increased strength in the SY component;

iii. increased strength and decreased variability in the LT component.

Hence, the distinction between episodic and non-episodic ozone conditions could be clustered through the relative magnitude of the SY and LT components.

3.3 Extraction and analysis of the temporal components of ozone: models vs. observations

The KZ-f components extracted from all ensemble members and from the observations are compared in Fig. 7. Overall, the observed scale separation and the accounted variance of the individual components was replicated satisfactory by the ensemble members. In Fig. 7a, the variance captured by the ensemble of models for the four components of the time series is presented in coloured lines, where each colour represents a different model. The variance of the median model is displayed with an unfilled square while the red circles indicate the variance of the observation. Finally, the filled square illustrates the variance of the kz model, built from the least RMSE spectral components over the entire 6-month period. The decomposed observed and modelled time-series generally reveal similar patterns. Spectral decomposition does not distort the allocation of variance between the components and hence maintains their relative importance. Moreover, this decomposition results at equal amounts of explained by individual components variance as seen in the last column of

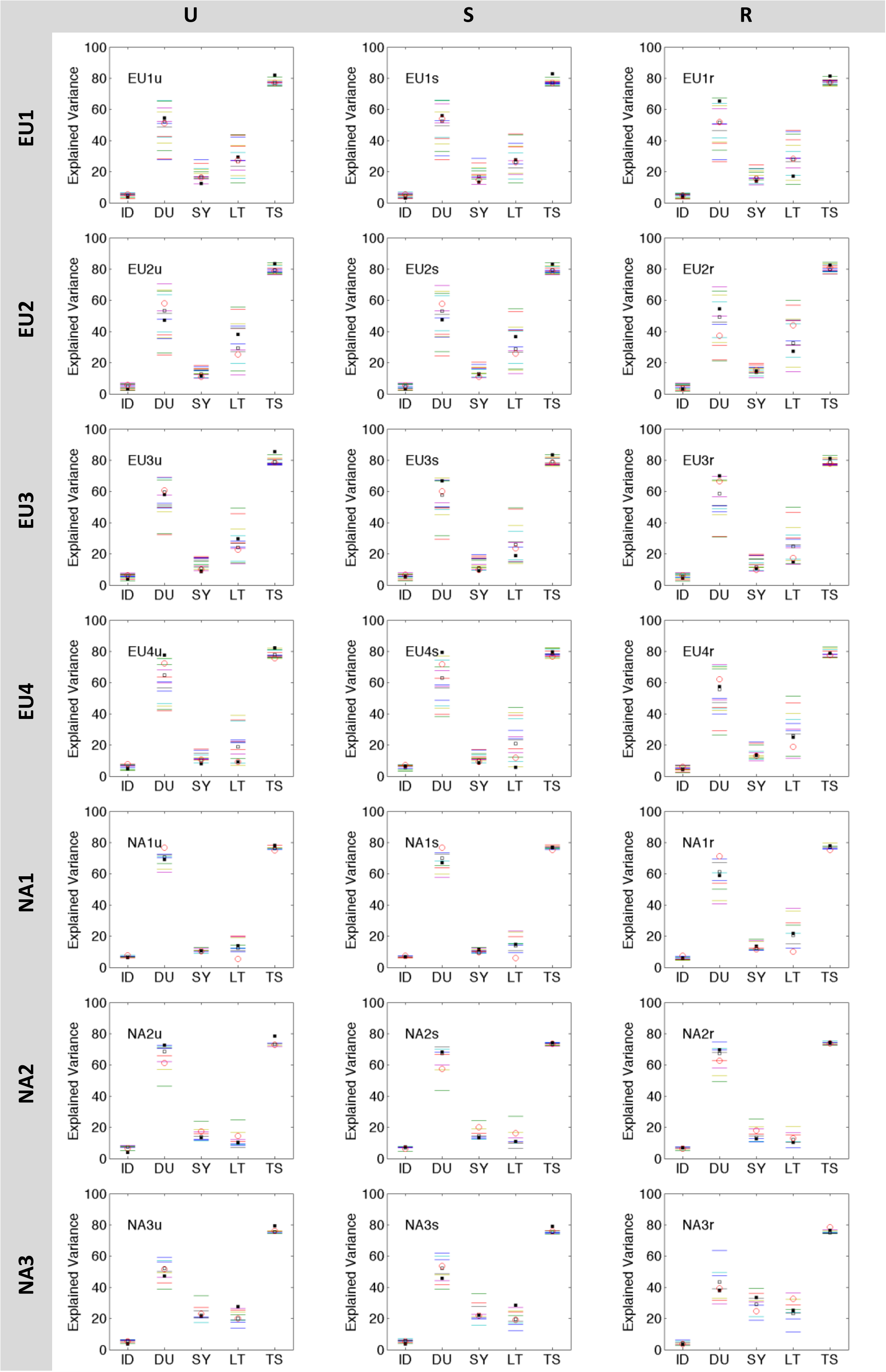

Fig. 7a. Properties of the spectrally decomposed modelled time-series of ozone, into four (ID, DU, SY, LT) components, versus their observed counterparts. Results are shown for all seven subdomains and three ozone aggregation types. For all plots, the square marker reflects the mm, the filled square the kz, the red circle the observations and the coloured lines the individual models. **(a)** Explained variance: the total explained variance is similar between models and observations, despite the dissimilar allocations seen for many models that tend to allocate less variance into the DU component, especially in the EU domain, and more variance to the LT component.

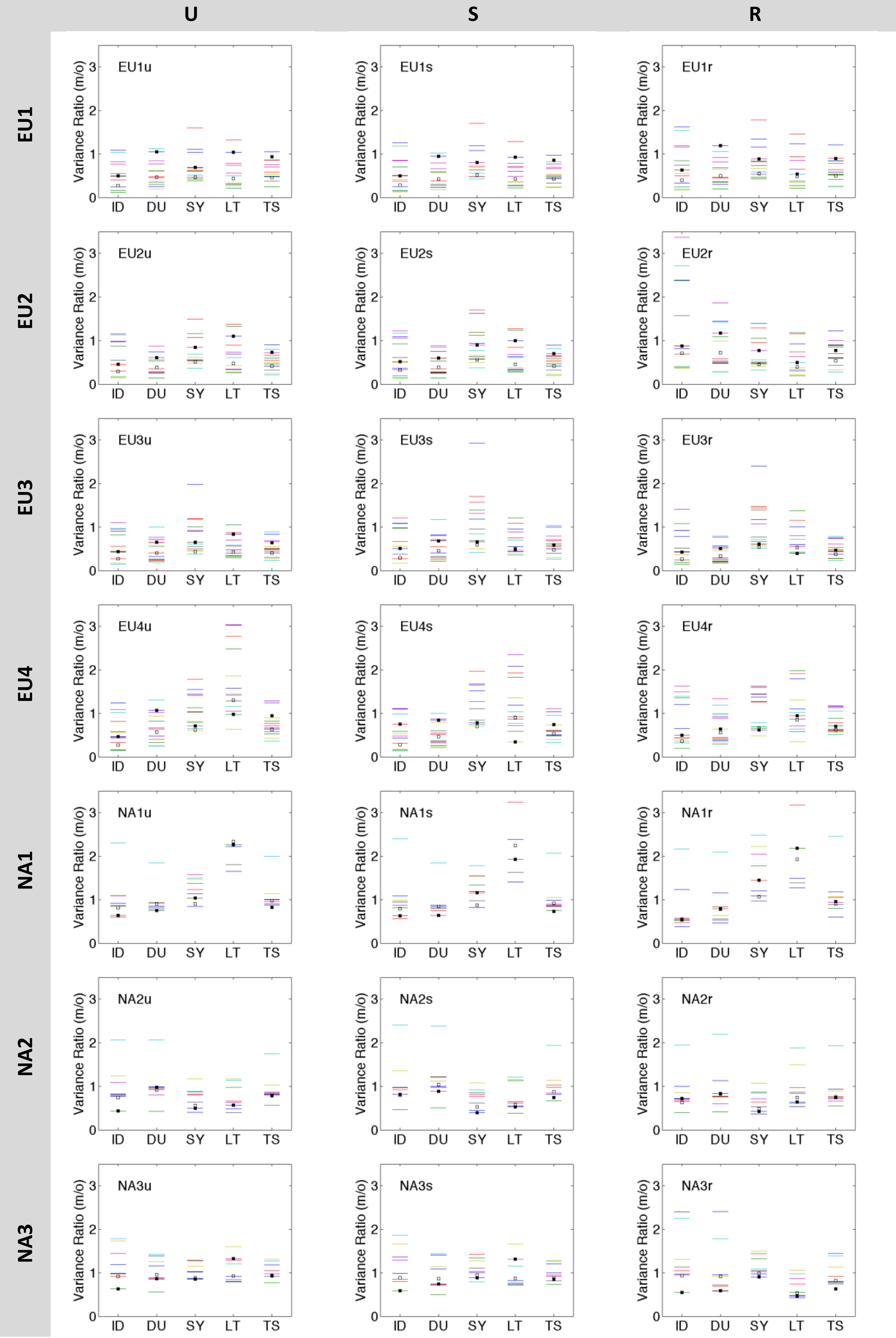

Fig. 7b. Properties of the spectrally decomposed modelled time-series of ozone, into four (ID, DU, SY, LT) components, versus their observed counterparts. Results are shown for all seven subdomains and three ozone aggregation types. For all plots, the square marker reflects the mm, the filled square the kz, the red circle the observations and the coloured lines the individual models. **(b)** Variance ratio: the modelled explained variance is expressed in terms of the observed explained variance by the use of their ratio. The variance ratio for the mm (squares) is variable and in many cases far from unity. Using the least RMSE spectral component results in a clear improvement of the variance ratio.

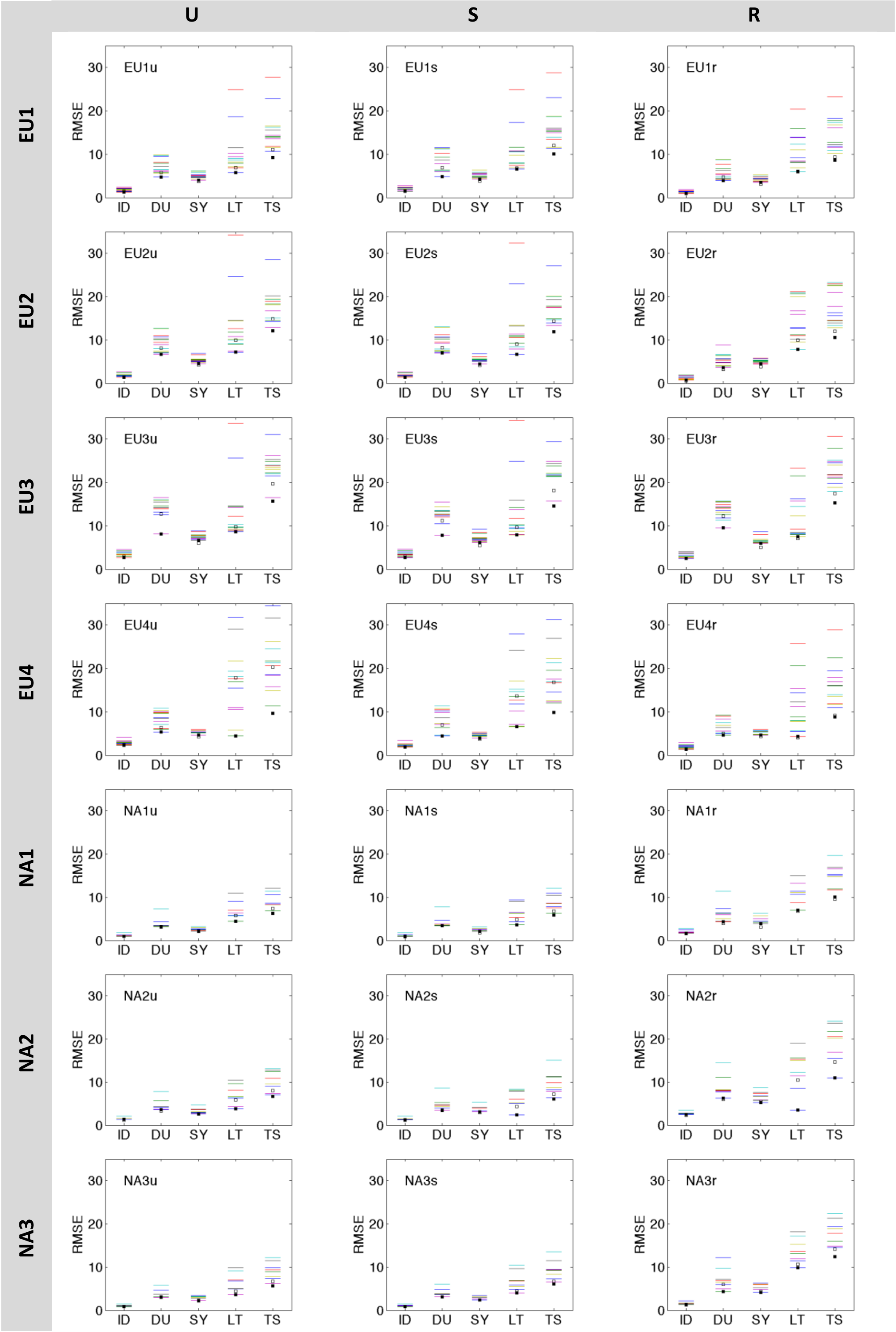

Fig. 7c. Properties of the spectrally decomposed modelled time-series of ozone, into four (ID, DU, SY, LT) components, versus their observed counterparts. Results are shown for all seven subdomains and three ozone aggregation types. For all plots, the square marker reflects the mm, the filled square the kz, the red circle the observations and the coloured lines the individual models. **(c)** RMSE: spectral RMSE for all ensemble members and total time series RMSE (TS). Building a model from different spectral components (filled square) results in lower RMSE than the documented ensemble mean (unfilled square).

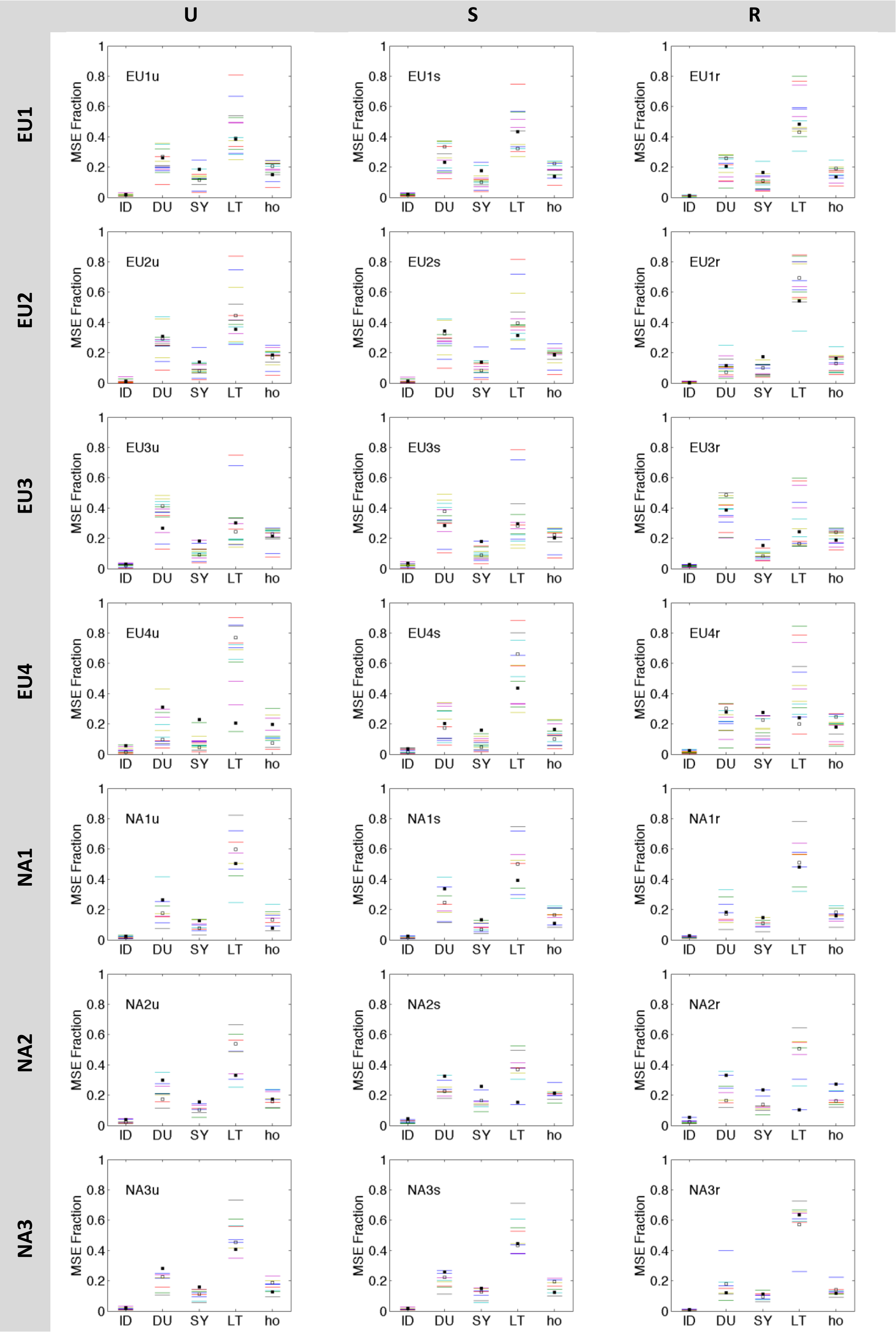

Fig. 7d. Properties of the spectrally decomposed modelled time-series of ozone, into four (ID, DU, SY, LT) components, versus their observed counterparts. Results are shown for all seven subdomains and three ozone aggregation types. For all plots, the square marker reflects the mm, the filled square the kz, the red circle the observations and the coloured lines the individual models. **(d)** MSE fraction: the decomposition of the mean square error of the modelled time-series in terms of the fraction each of the four spectral components accounted for. For mm, LT represents the principal error driver followed by DU (only exception is EU3). Using optimal components, kz has more balanced error distribution.

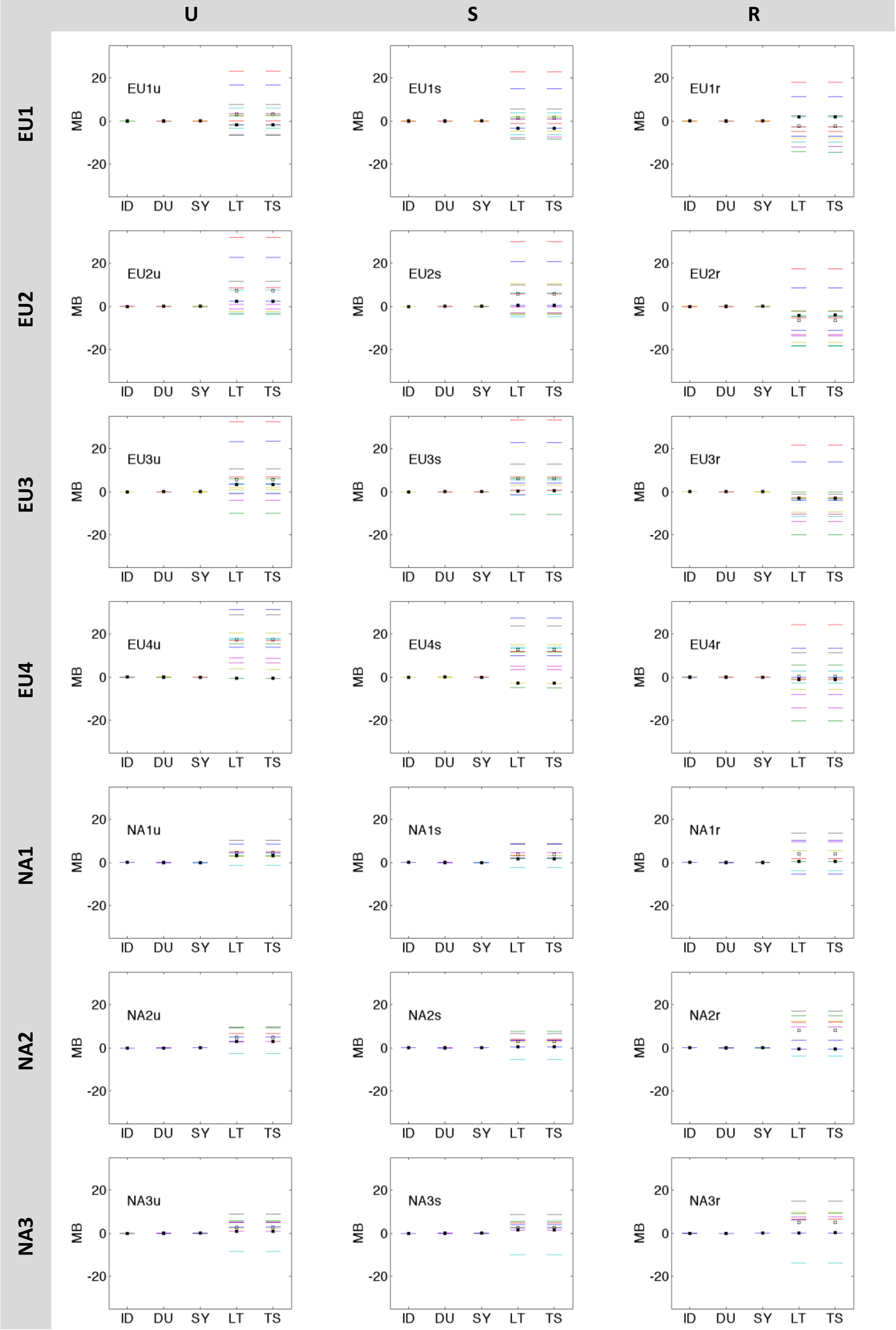

Fig. 7e. Properties of the spectrally decomposed modelled time-series of ozone, into four (ID, DU, SY, LT) components, versus their observed counterparts. Results are shown for all seven subdomains and three ozone aggregation types. For all plots, the square marker reflects the mm, the filled square the kz, the red circle the observations and the coloured lines the individual models. **(e)** MB: decomposition of the mean bias error of the modelled time-series. For all models, the bias error equals the LT component bias. In comparison to mm, kz simulations have lower absolute bias.

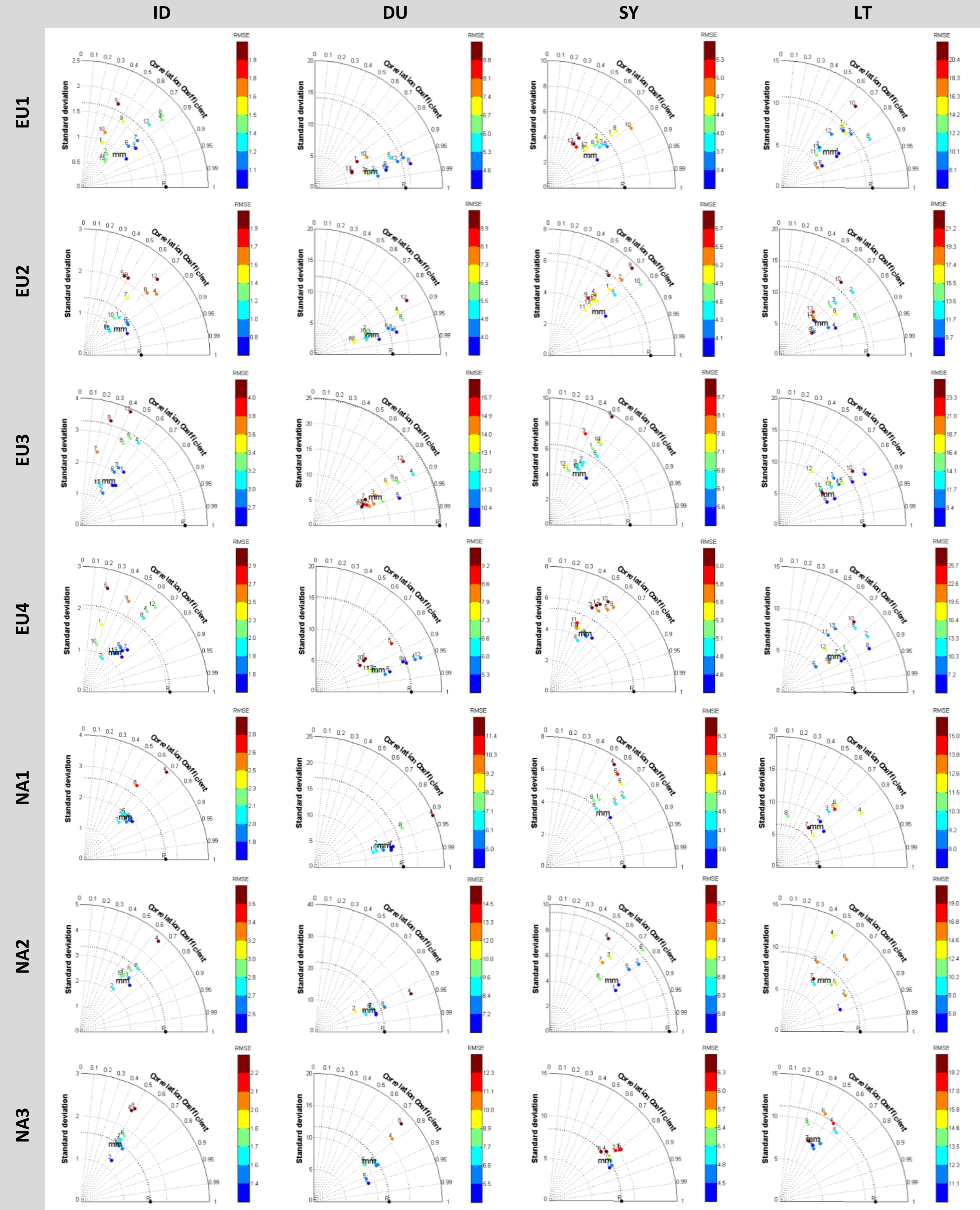

Fig. 7f. Properties of the spectrally decomposed modelled time-series of ozone, into four (ID, DU, SY, LT) components, versus their observed counterparts. Results are shown for all seven subdomains and three ozone aggregation types. For all plots, the square marker reflects the mm, the filled square the kz, the red circle the observations and the coloured lines the individual models. **(f)** Taylor diagrams of the modelled ID, DU, SY, LT spectral components (rural stations). Modelled ID is less successful in capturing the pattern (correlation), followed by SY. The amplitude (variance ratio, error) is generally conceived by the spectral components, with regional variability of the modelled skill.

each graph. There exist individual models however, whose kz decomposition entails similar total explained variance to the observations but with a systematical apportionment of dissimilar to the observed pattern allocations between short and long scales. While this should be of interest to the model developers, to investigate which process in their model is not represented accurately, here we do not filter out those models as a result of an ensemble pre-processing, but rather leave this task to the extraction algorithm presented in Sect. 2.2.

The portion of the observed range of spectral fluctuations captured by the models is explored through the modelled-to-observed variance ratio (Fig. 7b) and the component error graph (Fig. 7c). In the NA sub-regions the average variance falls very closely to the measured one, while in the EU sub-regions, a variance ratio close to unity is only for a few models. Overall, the variance ratio of the kz model is usually close to unity, demonstrating a clear improvement over the mm model that exhibits a variable behaviour with a tendency towards the underestimation of the ratio. Similar is the dominance of kz over mm in view of the RMSE. In addition, we can identify:

1. the dominant scales in the observations;
2. whether individual models as well as the ensemble are able to capture the variance at the right scale;
3. the components that drive the output error;
4. information to improve the use of the ensemble.

The fraction of the mean-square-error (MSE), decomposed in the first order terms (ID, DU, SY, LT) as well as their interactions (ho) is shown in Fig. 7d. For the majority of the cases, LT is the dominant error component of mm accounting for 40–80 % of the MSE; DU follows with accounted error in the order 10–50 %. The kz model, besides achieving a lower RMSE (Fig. 7c), it has a more balanced error allocation due to the selection of optimal components (and especially LT). Figure 7e demonstrates another property of the KZ-f decomposed time series, namely that the mean error bias (MB) of the time series equals the mean error of the LT component.

Figure 7f shows the combined skill (correlation, variance, error) of the individual models through Taylor plots (rural stations). Compared to the other components, the ID is less successful in capturing the observed pattern (correlation), but there is also high spread between the members skill. The highest correlation is seen for the DU; the figure also shows that the variance of the modelled DU signal is reasonably represented by many models. The spread of the values in the Taylor plot is variable across the sub-regions and in general, it is correlated to the variation of the modelled shortwave radiation (Vautard et al., 2012).

The plots presented in Fig. 7 also allow identifying the advantage of using together MME and KZ-f. At individual scale level the distribution of model results can be relatively big and skewed. The cases for the central modes are frequent where the measurements fall at the edge of the distribution. These elements will reflect in the distribution of the modelled time series and in the spread of the ozone values. The selection of the best component of the signal on the other hand preserves and contains all the models behaviours and captures only the one closest to the observed component. The fact that mode-wise model performance seems much poorer than the case when the model complete signal is analysed (Solazzo et al., 2012a), indicates that *cherry picking* the best modes from the model distribution and recomposing it into the kz model signal should produce better results than the statistical treatment of all model results as averages or medians. Figure 7 also reveals the interesting feature that the mode for which the models show the widest distributions of values are DU and LT. These in fact are controlling most of the process variance and reflect the variety of the model results in determining the ozone time variation.

4 Operational evaluation of the spectral model

4.1 Sub-regional level

As explained in Sect. 2, all ensemble members are decomposed into their KZ-f components, compared with the relative component obtained from the observations and the best ones are then composed to produce the kz model. In this section, we will evaluate, for a number of cases (175 week forecasts), the performance of the kz model against each individual ensemble members, the classical ensemble product mm, the sm and observation. As shown in Fig. 1, the best components for the kz model are obtained by comparing the individual model spectral results (what with abuse of language are normally defined as deterministic results) with the observations. Evaluation metrics are used to determine the level of agreement between the results and the observations.

The first operational assessment (Dennis et al., 2010) is presented in Fig. 8, where all models are directly compared to the observations (EU1, results are similar for the other sub-regions). The scatter diagram shows super-imposition of three clouds pertaining to the comparison with observed concentrations at rural, sub-urban and urban stations. The individual models are compared in the first 13 panels and are followed by mm, sm and kz. The improvement of kz is evident with respect to all other models. The cloud is tilted upward gaining a good deal of positions even against mm and sm. The spread of the data appears slightly larger than for sm and mm because the median aggregation in those models always results in deterioration of their variance. Another reason is related to imperfect selection of the best spectral components and it will be explored later in this section. However, kz model forecasts are homogeneous throughout the range of values. From a purely visual view point, the improvement produced by the kz model are clear.

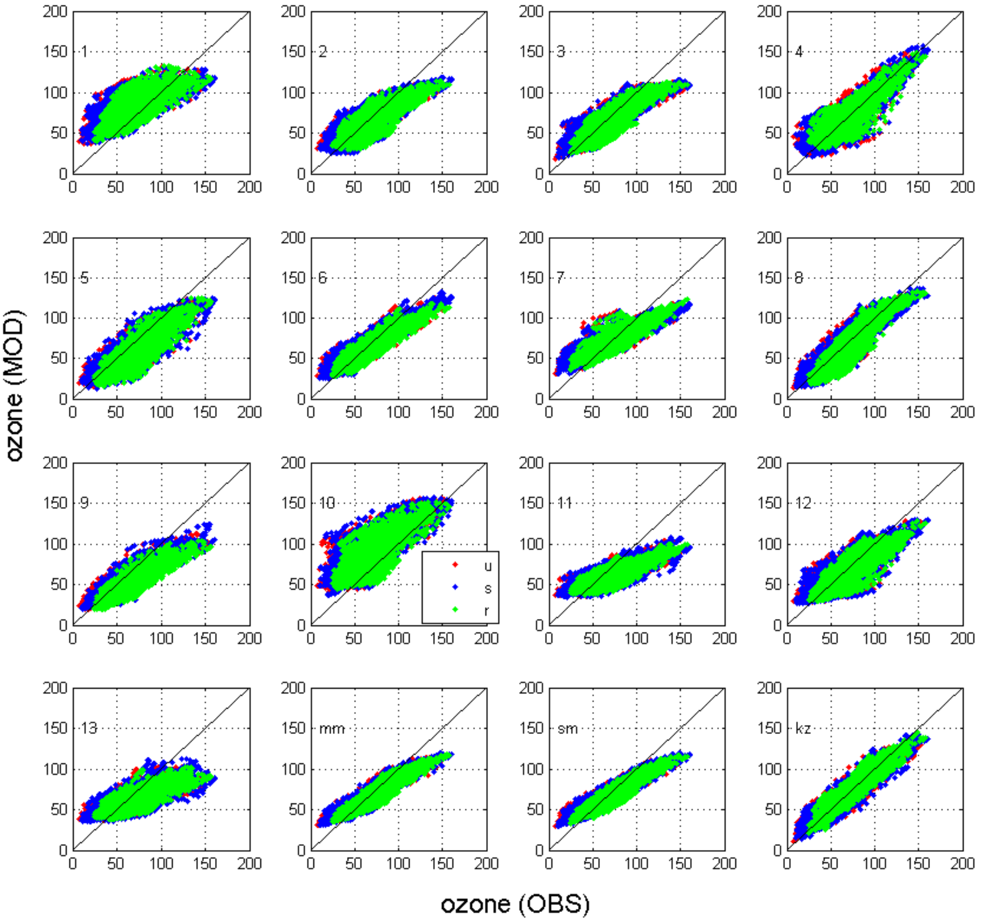

Fig. 8. Scatterplot of all examined cases corresponding to the prediction week, for the EU1 sub-domain and all three ozone aggregation types. Compared to the rest of the models, the cloud of the kz model scatter is tilted towards the diagonal.

Table 6. Decomposition of the kz model error into the spectral components (% of total error).

	URBAN				SUBURBAN				RURAL			
	ID	DU	SY	LT	ID	DU	SY	LT	ID	DU	SY	LT
EU1	6 %	40 %	27 %	27 %	6 %	39 %	29 %	26 %	5 %	36 %	28 %	31 %
EU2	5 %	41 %	25 %	29 %	5 %	42 %	24 %	28 %	2 %	25 %	28 %	45 %
EU3	8 %	39 %	31 %	22 %	7 %	40 %	30 %	22 %	8 %	47 %	27 %	18 %
EU4	9 %	41 %	27 %	23 %	10 %	41 %	30 %	18 %	6 %	38 %	37 %	19 %
NA1	7 %	38 %	22 %	33 %	7 %	43 %	19 %	31 %	7 %	30 %	23 %	39 %
NA2	7 %	38 %	20 %	36 %	8 %	43 %	29 %	19 %	9 %	42 %	28 %	20 %
NA3	5 %	42 %	27 %	26 %	4 %	38 %	29 %	29 %	4 %	25 %	30 %	40 %

The large amount of data and results forces us to condense the assessment in comprehensive graphical representation. In Fig. 9 the Taylor diagram (Taylor, 2001) is presented for all sub-regions and stations groups. The diagram relates the position of each deterministic model, mm, sm and kz model to the position of the observation on the x-axis. In all cases the kz model outscores all others: it minimises the distance from the reference point *R* (indicating high correlation and pattern match) and scores among the lower RMSE (colour scale). The mm and sm also behave better than the deterministic models, not unexpectedly. The performance of kz model in many cases is comparable to that of mm and sm (with the exception of the standard deviation ratio). The advantage is that in the case of kz model the result is obtained on the ground of a physical diagnosis of the ensemble whereas in the case of the ensemble the result is obtained

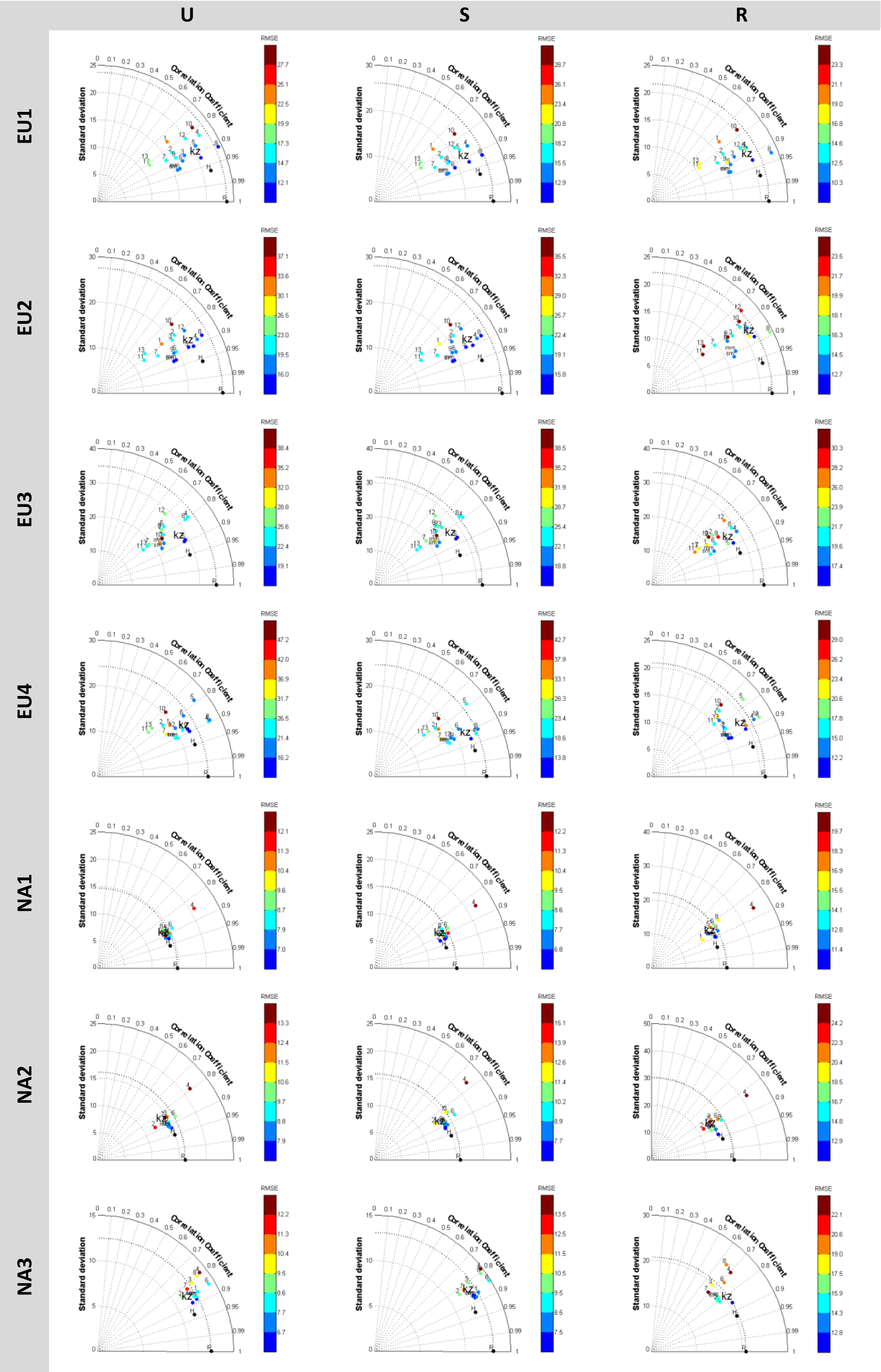

Fig. 9a. Summary statistics of all examined cases corresponding to the prediction week, for all seven subdomains and three ozone aggregation types. **(a)** Taylor diagrams: despite the different deterministic model excelling at each sub-domain, the behaviour of the kz model is homogeneous across domains achieving the least RMSE, very high PCC and STD close to the observed one.

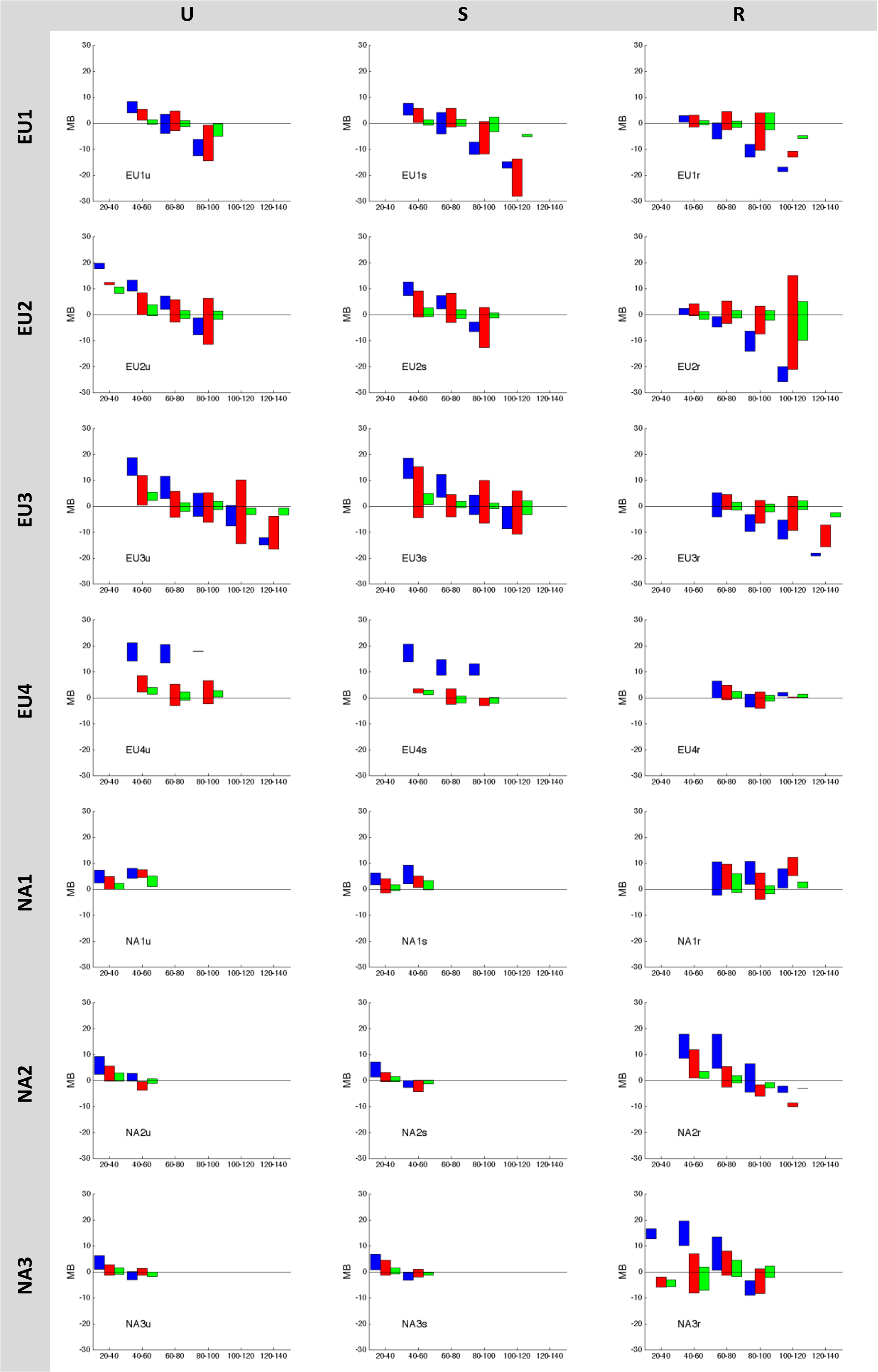

Fig. 9b. Summary statistics of all examined cases corresponding to the prediction week, for all seven subdomains and three ozone aggregation types. **(b)** Mean Bias over binned observed mean ozone mixing ratios for the prediction week, for mm (blue), kz (red) and kzH (green). The box extent is the inter-quartile range.

from a statistical treatment of an under-represented sample of model result. In fact, the lack of knowledge on the models level of dependence or correlation does not give a-priori guarantees on its success and produces wrong perception of model agreement. In the same figure with the black circle named *H* the predictability limit (upper bound of forecast skill) of the approach it is also plotted; it shows the forecast skill of the kz model if the best spectral components could be forecasted with absolute certainty.

Figure 9b illustrates the ensemble model behaviour beyond summary statistics. In particular, the mean bias error of the mm (blue) and kz (red) forecasts are shown as a function of the observed ozone mixing ratios. The green boxes correspond to the kz previsions generated with the optimal spectral components. While mm replicates the tension of the models to underestimate peaks and overestimate low concentrations, kz tends to generate predictions with a symmetric error distribution across all ozone ranges. The improvement in forecast accuracy at higher mixing ratios is one of the most notable properties of kz over mm.

Figure 10 shows the Cumulative Distribution Function (CDF) for the RMSE. The predictive skill of the ensemble is shown by the shaded area that is constructed by the RMSE of the best and the worst deterministic models. On top of those we superimpose the respective functions of the kz model and the mm. While the CDF of the mm in most cases exhibits higher RMSE than the best model, the CDF of the kz model demonstrates an extreme behaviour with the least RMSE values. The forecast skill of the kz model is further enhanced by the fact that the best deterministic model is generally different at each panel.

The individual model performance in reproducing at best the scale filtered in the observation is presented in Table 5. For all the dub-regions, station types, and components the table reports the identification number of the model showing the minimum RMSE with the filtered observed signal. It is interesting to notice that for each sub-region and all sets of station types, a limited number of models is needed to reconstruct the signal, almost independently to the station type. For the EU1 sub-region, five models are sufficient, for EU3 we would need seven, while for NA2 only 3. Three conclusions can be made here:

1. the number 4–6 as minimum set of models sufficient to reproduce the result is in agreement from the finding of Solazzo et al. (2012a) that found that the best ensemble results out of 12 models available could be obtained with 4/5 models only.

2. That the issue of model independence is very relevant in this context too and that only a handful of original contributions can be extracted even from a large ensemble and only that group will make the difference (Potempski and Galmarini, 2009)

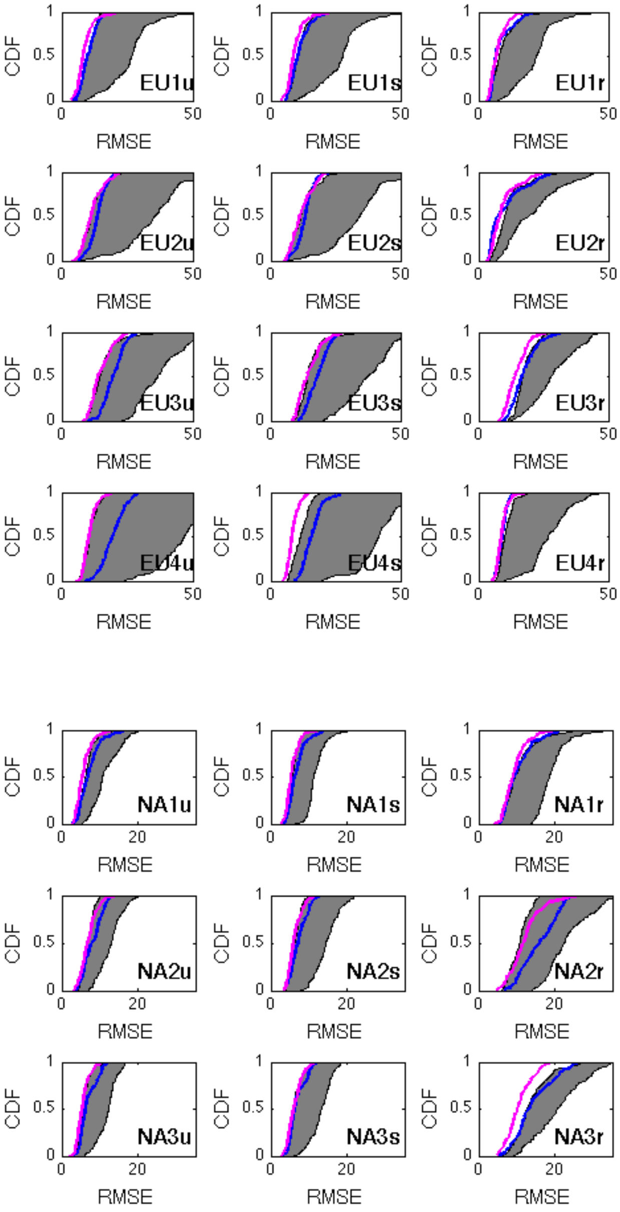

Fig. 10. The Cumulative density function of the RMSE distribution of the best and worst deterministic model is illustrated by the shaded area. The RMSE distribution of kz model (magenta) is always found in the leftmost side of the figure. The mm distribution is given in blue.

3. From Table 5 it can be seen that many models are needed to reproduce a comprehensively good result across all sub-regions and, therefore having the possibility of using a large pool of models is of essence.

This latter point is confirmed by the results in Fig. 11. The histograms provide the contribution of each model in identifying the best component. Only for a very limited number of cases the dominance of one or two models is evident, especially in the NA sub-regions. A question raises: do the kz

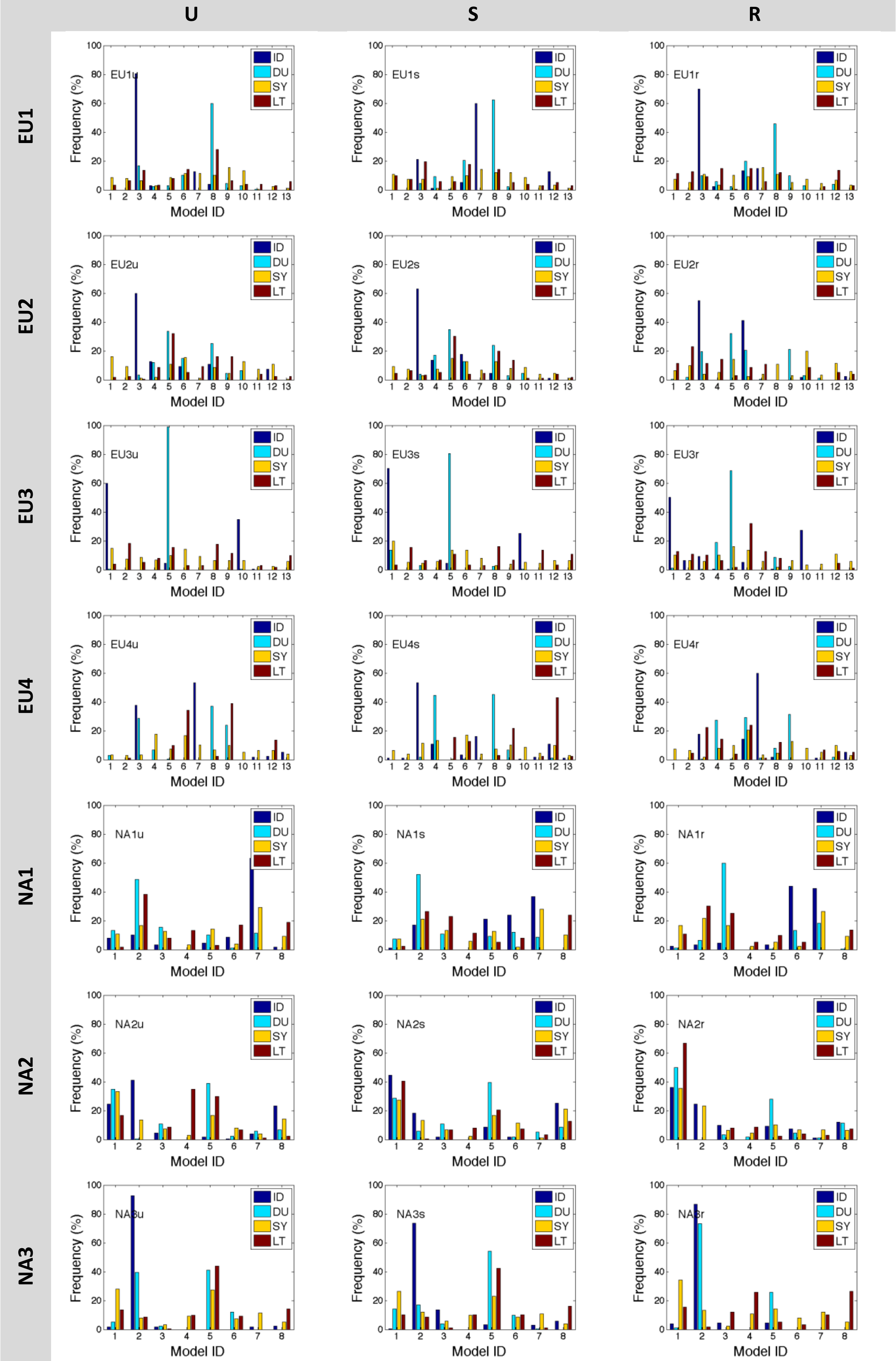

Fig. 11. The frequency of selection of each model's spectral components as elements of the kz model, for all seven subdomains and three ozone aggregation types. Generally, a couple of models dominate into the ID and DU components while the SY and LT components of the kz model make use of nearly all the ensemble members.

Table 7. Independence of spectral components versus error. The covariance of the error is averaged over all models. Using only two spectral components, being either (ID + DU + SY, LT) or (ID + DU, SY + LT), the decomposition achieves independent factors, but their corresponding kz model has ∼5 % higher RMSE (compared to the case of four spectral components).

		Average Error Covariance				RMSE (kz)			
DOMAIN	TYPE	4SPC ID DU SY LT	3SPC ID DU SY + LT	2SPC ID + DU + SY LT	2SPC ID + DU SY + LT	4SPC ID DU SY LT	3SPC ID DU SY + LT	2SPC ID + DU + SY LT	2SPC ID + DU SY + LT
EU1	u	18.8 %	12.9 %	5.9 %	6.5 %	9.3	10.5	10.7	10.7
EU1	s	19.3 %	13.7 %	5.7 %	6.9 %	10.0	10.4	11.4	10.7
EU1	r	15.5 %	9.9 %	5.7 %	5.6 %	8.6	9.6	9.4	9.8
EU2	u	17.4 %	12.6 %	4.9 %	6.6 %	12.1	12.4	12.7	12.6
EU2	s	18.1 %	13.1 %	5.1 %	7.0 %	11.9	12.8	12.5	13.0
EU2	r	12.9 %	6.2 %	6.8 %	3.8 %	10.6	11.2	11.1	11.7
EU3	u	21.4 %	17.5 %	4.1 %	8.5 %	15.7	16.2	16.1	16.4
EU3	s	21.5 %	16.7 %	4.9 %	8.4 %	14.6	15.0	15.3	15.4
EU3	r	20.4 %	16.4 %	4.0 %	8.6 %	15.3	15.6	15.8	15.9
EU4	u	13.0 %	10.3 %	2.8 %	4.9 %	9.7	10.2	10.6	10.6
EU4	s	13.8 %	10.8 %	3.1 %	5.3 %	9.9	10.2	10.5	10.4
EU4	r	17.3 %	12.5 %	5.0 %	7.4 %	8.9	10.0	9.2	10.1
NA1	u	14.6 %	10.4 %	4.2 %	5.4 %	6.3	6.6	6.6	6.7
NA1	s	16.3 %	12.1 %	4.4 %	6.3 %	5.8	6.1	6.3	6.3
NA1	r	16.0 %	10.1 %	5.9 %	5.7 %	10.1	10.3	10.1	10.3
NA2	u	17.9 %	12.9 %	5.2 %	6.4 %	6.7	6.7	7.0	7.0
NA2	s	20.8 %	14.8 %	6.0 %	8.1 %	6.1	6.1	6.4	6.4
NA2	r	18.3 %	12.8 %	5.6 %	7.0 %	11.0	11.0	11.0	11.0
NA3	u	16.3 %	12.0 %	4.5 %	6.6 %	5.7	5.8	6.2	6.0
NA3	s	16.7 %	12.3 %	4.6 %	7.1 %	6.1	6.3	6.6	6.6
NA3	r	13.4 %	9.1 %	4.4 %	5.6 %	12.4	13.6	13.2	13.6
Mean		17.1 %	12.3 %	4.9 %	6.6 %	9.9	10.3	10.4	10.5

model components shown in Fig. 11 accurately represent the distribution of the actual “best”’ components? The answer is yes, but not always with the right order. This is now explored.

It is important to examine the accurate extraction of the best spectral components at each forecast week. In Fig. 12 the CDF at each set of station type measurements and sub-region, of how the best model components identified during the past week correspond to the best ones over the next week, is shown. The plot shows, for example, that for component ID in EU1 for urban stations the selected model component over the past week for kz model was actually the best component of the future week in the 70 % of the cases, was the 2nd best in the 20 % of the cases leaving the rest to lower rankings. For all components the hit rate is very high. The only exception is the SY components as it clearly appears from the Fig. 12. In almost all sub-regions the selected SY components span linearly all the ranks indicating that the predictability of this component is rather limited. Since this component is related to weather predictability, this result is not unexpected. This is also a clear indication of where widespread fundamental deficiencies across model occur. Overall, this imperfect selection of the components caused the distance from the H point at the Taylor diagrams which however did not prevent the kz model from outperforming other models.

We will examine now the relative contribution of each kz model component to the total error. For each forecast week, we calculate the relative strength of the error terms of Eq. (4). Then we calculate the higher order error contribution of each spectral component and finally we compute the mean error per component. The result is given in Table 6. On average, the DU component generally entails the higher error fraction across all sub-regions. The only exception to this rule is found for the rural aggregation of ozone in the three most densely populated sub-regions: EU2, NA1 and NA3 (an explanation is given by Fig. 13 in the next paragraph). On the whole, the DU is responsible for roughly 40 % of the error, the SY and LT explain around 28 % of the error each, leaving the last 5 % to the ID.

In view of the operational applicability of the approach, we combine the kz model skill with the decomposition of its spectral error in order to isolate the cases where its

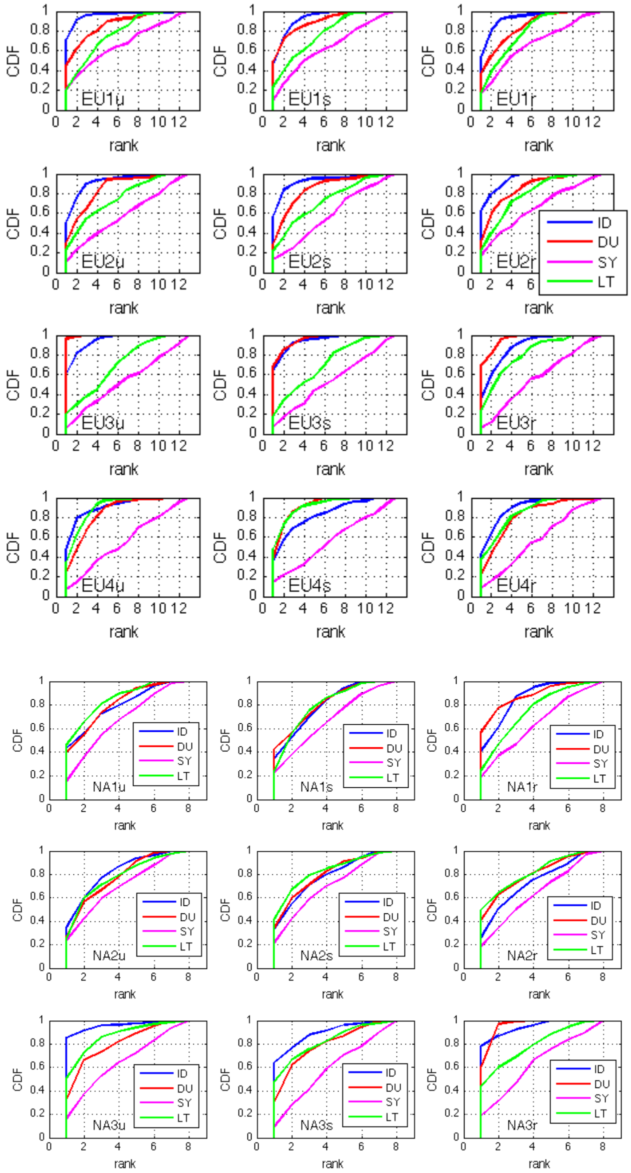

Fig. 12. The Cumulative density function of the actual rank (validated against observations) of the selected models in the PREDICT week (F-step). In diagnostic (hindcast) mode, we only have rank 1. The persistence assumption for the best modelled spectral components is strong for ID, DU and LT (in this order) and weaker for the SY component.

performance was degraded. First, for each examined case (of the 175) we rank the forecast skill of the models (deterministic and kz model) with respect to their RMSE. The primary y-axis in Fig. 13 (bar plot) shows the frequency allocated to all rankings by the kz model (rank 1 is best) while the dotted line represents the cumulative probability. Generally, for more than two thirds of the cases the kz model achieved the least or the 2nd least RMSE, across all sub-regions. This finding is conservative in the sense that the behaviour of the deterministic models was not homogeneous across the sub-regions, resulting in different rankings. On the other hand, there exist a few cases where the kz model ranking was poor. For this reason, we decomposed the kz model total error for all spectral components (averagely shown in Table 6) and rankings (Fig. 13, line plot in the secondary y-axis). We clearly observe that the low kz model rankings are caused by an improper selection of the LT component. The functioning of different selection procedures will be explored in the future.

4.2 Station level

The performance of the kz model has been so far evaluated at sub-regional level. The performance at station level is however the only one that really matters at the end of the day. A sample of stations has been selected in order to test the validity of the approach, already seen at regional level, at discrete point locations. Different sets of stations were selected, covering all examined sub-regions, with the only criterion of representativeness being the vicinity to well-known ozonesonde sites for which observational data for ozone were available. In view of this criterion, the stations presented are taken from nearly all sub-regions and are namely: IE1 (Uccle), CH1 (Payerne), IT1 (Motta Visconti – Po valley), ES2 (Saragossa), US1 (San Diego), US3 (Springfield). Results are shown in Fig. 14 in the form of Taylor diagrams, CDF plots of RMSE, scatterplots as well as time series. For the majority of the examined rural stations, the kz model forecasts provide improved RMSE distribution over the best deterministic model while at the same time they maintain one of the highest correlations and account well for the observations variance. As seen before, the scatterplot of the kz model forecasts is again tilted towards the diagonal. At the same graph we also plotted the time-series of ozone predictions for the Payerne station during the week with the highest mean level between the cases. In terms of the kz model, the persistence assumption was found true only for the DU component in this case. This result clearly shows that a good forecast can also be produced with elements of the least skill ensemble members. Last, the kz model provisions were of high quality even at the urban stations in Paris and Vienna (not shown).

4.3 Anticipating the application to other pollutants: NO_2 and PM_{10}

The methodology adopted and applied for ozone is extended for the case of Nitrogen Dioxide (NO_2) and coarse Particulate Matter (PM_{10}). Although this work is ongoing, results in the form of scatterplots for NO_2 (Fig. 15) and PM_{10} (Fig. 16) clearly show that the presented approach is not bound by the physical, chemical and dynamical nature of ozone formation and can be easily extended to other pollutants. Detailed results for NO_2 and PM_{10} will be published separately.

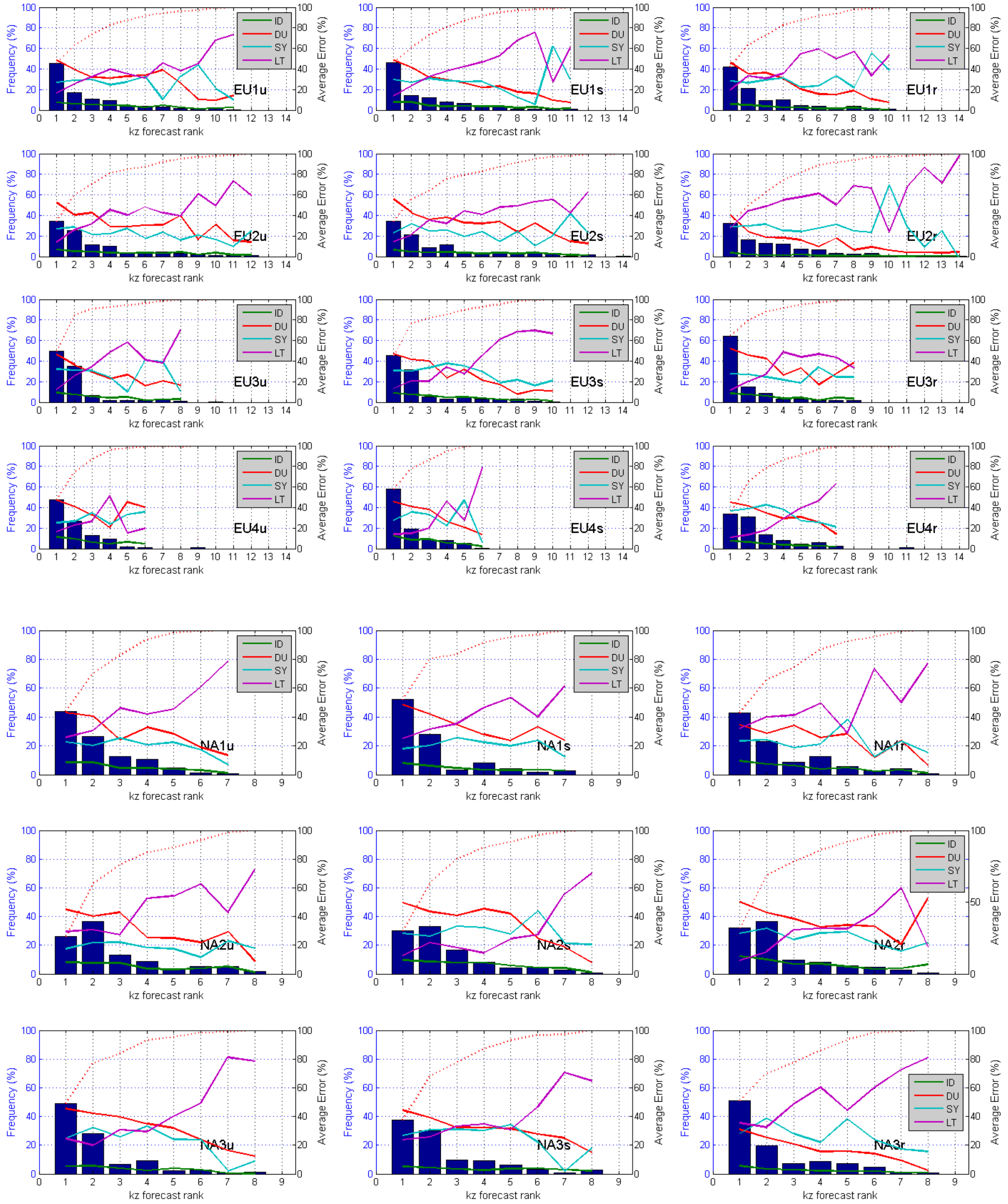

Fig. 13. Ranking frequency of the kz model together with the average spectral error at all rankings. Top is EU, bottom is NA.

4.4 Final considerations

We will close the Results section with a discussion on two important issues of the kz filter, in particular the component independence and the distortions.

In previous sections we have seen that the four selected spectral components are not independent; there is roughly 20 % variability that is explained by their interactions (and especially between neighbouring spectral bands, ID and DU, DU and SY, SY and LT). Although the component selection

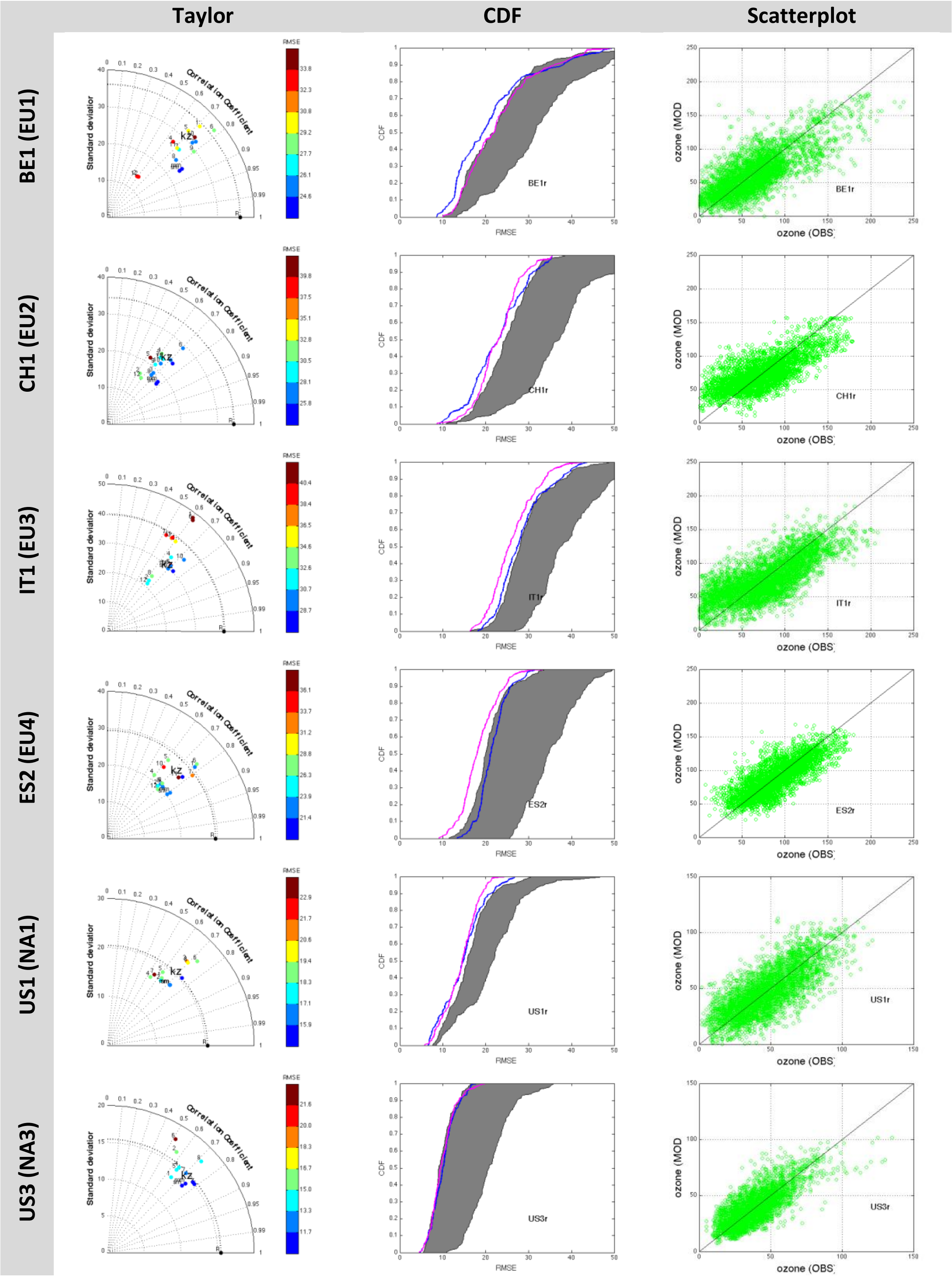

Fig. 14a. Indicative results at the station level. **(a)** (left column) Taylor diagrams for the examined stations corresponding to the prediction week; (middle column) the cumulative density function of the RMSE distribution of the best and worst deterministic model (shaded area) together with the distribution of kz model (magenta) and mm (blue); (right column) scatterplot of all examined cases corresponding to the prediction week.

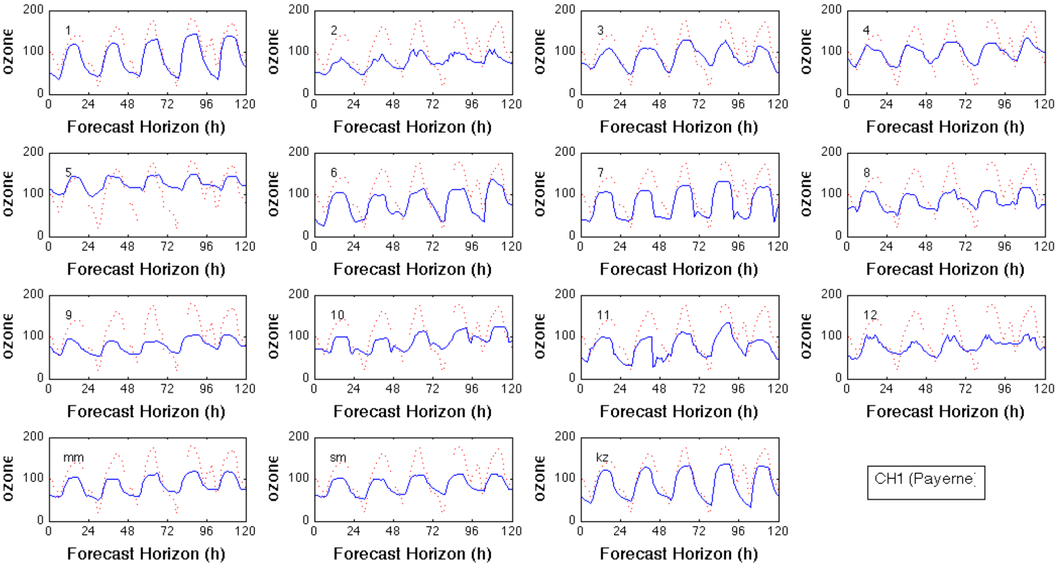

Fig. 14b. Indicative results at the station level. **(b)** Time-series of ozone predictions for the Payerne station during the week with the highest levels. The persistence assumption was true only for the DU component.

was done principally on the basis of physical considerations, we hereby explore the properties of other combinations of the four spectral bands. The additional cases examined correspond to wider spectral bands. Table 7 contains links to component independence and error of other spectral combinations generated with the KZ-f. The separation between short and long term (i.e., using only two components) gives more independent components, but results in forecasts with higher error (due to the negative error covariance in the case of dependent components but also to the coarser grouping of the processes replication).

In addition, filtered values by KZ-f near the end of the time series do not have the same statistical properties as those away from the end. This applies especially for the last half-length of the KZ-f. Those edge effects are responsible for around 10 % of RMSE to the kz and arise from lower persistence in the SY and LT components (not shown). Such distortions particularly affect the last two forecast days of the SY and LT signals and the last six hours of the seventh forecast day of the DU signal. For this reason, one may only consider the first 5 forecast days of the forthcoming week to minimise such distortions. However, a combined SY + LT signal will limit the distortions to the DU range (i.e., last six hours). If we combine this property with the independence discussed in the previous paragraph, we could argue towards the use of two components (ID + DU, SY + LT) as the envisaged extension of the presented approach.

5 Conclusions and outlook

The individual forecasts of a multi-model ensemble consisting of 13 air quality models have been spectrally decomposed together with the respective observations over multiple European and North American sub-regions. The modelled spectral components have been evaluated against their observed counterparts for coherence and accuracy. It was found that the composite model built from the best spectral elements outscores all the ensemble members as well as the ensemble median. In order to check the operational implementation of the method, we investigated whether the best spectral components could be known in advance. A persistence criterion was employed on the basis of the skill of the modelled spectral components during the last 7 days.

The evaluation against observational ground level ozone concentration gathered in AQMEII clearly showed that the forecast skill of the new model was superior to any individual ensemble member in terms of some of the most applied error metrics (correlation coefficient, mean-square-error, variance). Overall, its forecasts were bias-free, with mean-square-error not depending on the concentrations. In two-third of the examined cases across multiple sub-regions and aggregation types, it was ranked either first or second. The dominance of the new model was also witnessed by comparing the time-series of all models vs. the observations, for episodic and non-episodic conditions. Finally, following a detailed analysis of the new model forecast errors and their roots, it was found that there exist a few cases when its skill is degraded due to improper selection of the long-term spectral component. Different selection approaches are currently examined to eliminate this issue.

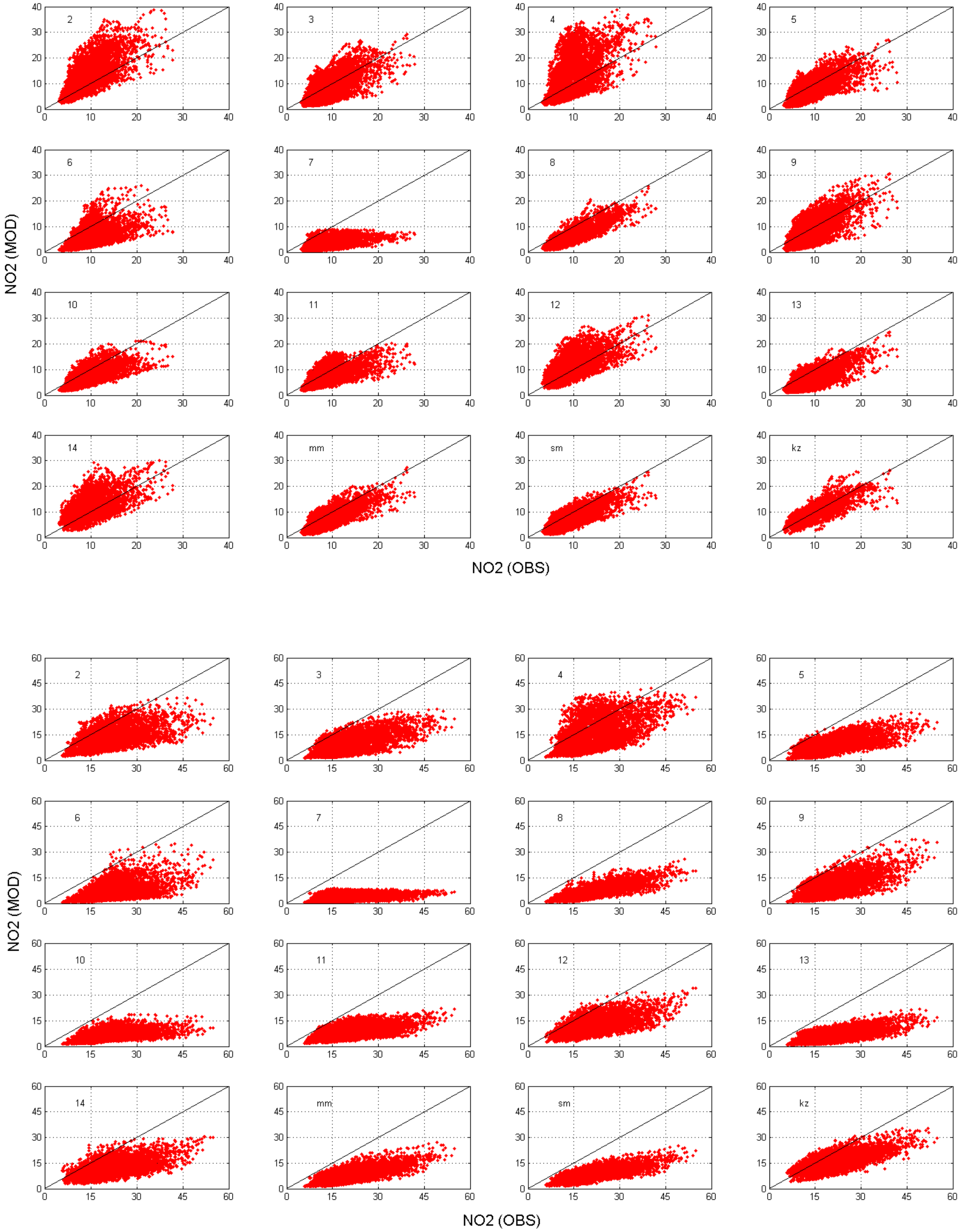

Fig. 15. Indicative results for NO_2. Scatterplot of all examined cases corresponding to the prediction week, for the rural (top) and urban (bottom) concentrations of the EU2 sub-domain.

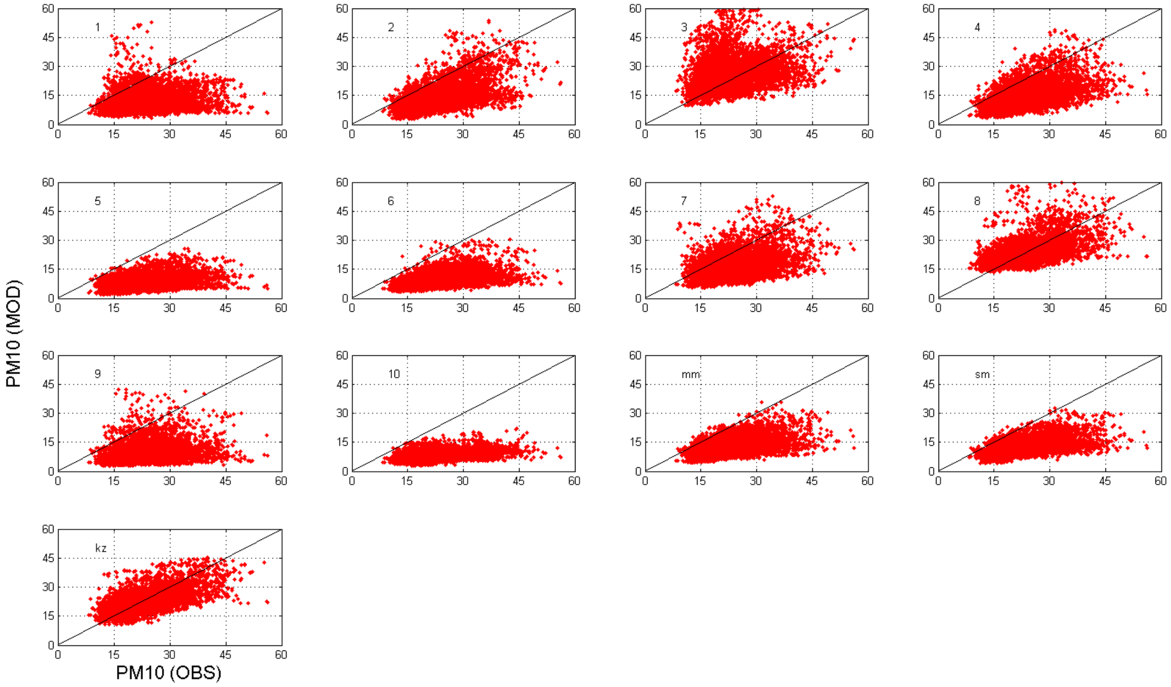

Fig. 16. Indicative results for PM_{10}. Scatterplot of all examined cases corresponding to the prediction week, for the rural concentrations of the EU1 sub-domain.

The forecasting methodology we introduce is new and represents a new approach to multi-model ensembles. In fact it still requires the availability of different model results but the diagnosis of the performance of the latter on a hindcast period allows the selection and combination of only those that are considered satisfactory with respect to a precise set of parameters. The diagnosis that is performed at scale levels gives many more guaranties on the performance in forecast mode than any classical multi-model statistical treatment.

The approach is adaptive and screens each time the ensemble members to extract the best spectral components. The persistence approach is a simple way to extract members on the basis of the most accurate recent representation of the observed state. Its advantages were demonstrated for seven sub-regions (four in EU, three in NA) at all aggregation types (urban, suburban, rural) but also at the station level. As the skill of the models varies with sub-region, synoptic conditions and chemical conditions, more sophisticated approaches (utilising e.g., synoptic clustering) are expected to further improve the forecast skill. Although the analysis was restricted to ozone, it was also seen that it can be extended to other pollutants such as NO_2 and PM_{10}. In view of its applicability, the technique is rather easy to implement at point locations. Similarly, it can be extended to spatial domains through the use of a multi-dimensional cost function (e.g., MSE of all stations) for use in operational forecasting with MME.

Acknowledgements. We would like to thank the two anonymous reviewers and S. Potempsky for their meaningful comments that helped us to clarify the parts of the manuscript.

Edited by: J. Brandt

References

Delle Monache, L., Deng, X., Zhou, Y., and Stull, R.: Ozone ensemble forecasts: 1. A new ensemble design. J. Geophys. Res., 111, D05307, doi:10.1029/2005JD006310, 2006.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D., and Venkatram, A.: A framework for evaluating regional-scale numerical photochemical modelling systems, Environ. Fluid Mech., 10, 471–489, 2010.

Galmarini, S., Michelutti, F., and Thunis, P.: Estimating the Contribution of Leonard and Cross Terms to the Subfilter Scale from Atmospheric Measurements, J. Atmos. Sci., 57, 2968, doi:10.1175/1520-0469(2000)057<2968:ETCOLA>2.0.CO;2, 2000.

Galmarini, S., Bianconi, R., Bellasio, R., and Graziani, G.: Forecasting consequences of accidental releases from ensemble dispersion modelling, J. Environ. Radioact., 57, 203–219, 2001.

Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J. C., Bellasio, R., Bompay, F., Buckley, R., Bouzom, M., Champion, H., D'Amours, R., Davakis, E., Eleveld, H., Geertsema, G. T., Glaab, H., Kollax, M., Ilvonen, M., Manning, A., Pechinger, U., Persson, C., Polreich, E., Potemski, S., Prodanova, M., Saltbones, J., Slaper, H., Sofiev, M. A., Syrakov, D., Sørensen, J. H., Van der Auwera, L., Valkama, I., and Zelazny, R.: Ensemble dispersion forecasting, Part 1: Concept, Approach and indicators, Atmos. Environ., 38, 4607–4617, 2004.

Galmarini, S., Rao, S. T., and Steyn, D. G.: AQMEII: An International Initiative for the Evaluation of Regional-Scale Air Quality Models – Phase 1, Atmos. Environ., 53, 1–3, 2012a.

Galmarini, S., Rao, S. T., and Steyn, D. G.: Preface, Atmos. Environ., 53, 1–3, 2012b.

Galmarini S., Bianconi, R., Appel, W., Solazzo, E., Mosca, S., Grossi, P., Moran, M., Schere, K., and Rao, S. T.: ENSEMBLE and AMET: Two systems and approaches to a harmonized, simplified and efficient facility for air quality models development and evaluation, Atmos. Environ., 53, 51–59, 2012c.

Guenther, A., Zimmerman, P., and Wildermuth, M.: Natural volatile organic compound emission rate estimates for U.S. woodland landscapes, Atmos. Environ., 28, 1197–1210, doi:10.1016/1352-2310(94)90297-6, 1994.

Hogrefe, C., Rao, S. T., Zurbenko, I. G., and Porter, P. S.: Interpreting the information in ozone observations and model predictions relevant to regulatory policies in the eastern United States, B. Am. Meteorol. Soc., 81, 2083–2106, doi:10.1175/1520-0477(2000)081<2083:ITIIOO>2.3.CO;2, 2000.

Hogrefe, C., Vempaty, S., Rao, S. T., and Porter, P. S.: A comparison of four techniques for separating different time scales in atmospheric variables, Atmos. Environ., 37, 313–325, 2003.

Kang, D., Mathur, R., Rao, S. T., and Yu, S.: Bias Adjustment Techniques for Improving Ozone Air Quality Forecasts, J. Geophy. Res., 113, D23308, doi:10.1029/2008JD10151, 2008.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G.: Challenges in Combining Projections from Multiple Climate Models, J. Climate, 23, 2739–2758, 2010.

Kukkonen, J., Olsson, T., Schultz, D. M., Baklanov, A., Klein, T., Miranda, A. I., Monteiro, A., Hirtl, M., Tarvainen, V., Boy, M., Peuch, V.-H., Poupkou, A., Kioutsioukis, I., Finardi, S., Sofiev, M., Sokhi, R., Lehtinen, K. E. J., Karatzas, K., San José, R., Astitha, M., Kallos, G., Schaap, M., Reimer, E., Jakobs, H., and Eben, K.: A review of operational, regional-scale, chemical weather forecasting models in Europe, Atmos. Chem. Phys., 12, 1–87, doi:10.5194/acp-12-1-2012, 2012.

Mallet, V. and Sportisse, B.: Ensemble-based air quality forecasts: A multimodel approach applied to ozone, J. Geophys. Res., 111, D18302, doi:10.1029/2005JD006675, 2006.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, L08703, doi:10.1029/2011GL046864, 2011.

McKeen, S., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Hsie, E.-Y., Gong, W., Bouchet, V., Menard, S., Moffet, R., McHenry, J., McQueen, J., Tang, Y., Carmichael, G. R., Pagowski, M., Chan, A., Dye, T., Frost, G., Lee, P., and Mathur, R.: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, J. Geophys. Res. D Atmos., 110, 1–16, 2005.

Papanastasiou, D. K., Melas, D., and Bartzanas, T.: Estimation of Ozone Trend in Central Greece, Based on Meteorologically Adjusted Time Series. Environ. Model. Assess., 17, 353–361, 2012.

Peuch V.-H. and Razinger, M.: ENSEMBLE: Products, Quality and Background Information, MACC Report D_R-ENS_3.1.6, 2011.

Pirtle, Z., Meyer, R., and Hamilton, A.: What does it mean when climate models agree? A case for assessing independence among general circulation models. Environ. Sci. Pol., 799, 2010.

Potempski, S. and Galmarini, S.: *Est modus in rebus*: analytical properties of multi-model ensembles, Atmos. Chem. Phys., 9, 9471–9489, doi:10.5194/acp-9-9471-2009, 2009.

Potempski, S., Galmarini, S., Addis, R., Astrup, P., Bader, S., Bellasio, R., Bianconi, R., Bonnardot, F., Buckley, R., D'Amours, R., van Dijk, A., Geertsema, G., Jones, A., Kaufmann, P., Pechinger, U., Persson, C., Polreich, E., Prodanova, M., Robertson, L., Sørensen, J., and Syrakov, D.: Multi-model ensemble analysis of the ETEX-2 experiment, Atmos. Environ., 42, 7250–7265, 2008.

Potempski, S., Galmarini, S., Riccio, A., and Giunta, G.: Bayesian model averaging for emergency response atmospheric dispersion multimodel ensembles: Is it really better? How many data are needed? Are the weights portable?, J. Geophys. Res., 115, D21309, doi:10.1029/2010JD014210, 2010.

Pouliot, G., Pierce, T., Denier van der Gon, H., Schaap, M., Moran, M., and Nopmongcol, U.: Comparing Emissions Inventories and Model-Ready Emissions Datasets between Europe and North America for the AQMEII Project, Atmos. Environ., 53, 4–14, 2012.

Rao, S. T., Zurbenko, I. G., Neagu, R., Porter, P. S., Ku, J. Y., and Henry, R. F.: Space and time scales in ambient ozone data, B. Am. Meteorol. Soc., 78, 2153, doi:10.1175/1520-0477(1997)078<2153:SATSIA>2.0.CO;2, 1997.

Rao, S. T., Galmarini, S., and Puckett, K.: Air quality model evaluation international initiative (AQMEII): Advancing the state of the science in regional photochemical modelling and its applications, B. Am. Meteorol. Soc. 92, 23–30, doi:10.1175/2010BAMS3069.1, 2011.

Riccio, A., Giunta, G., and Galmarini, S.: Seeking for the rational basis of the Median Model: the optimal combination of multi-model ensemble results, Atmos. Chem. Phys., 7, 6085–6098, doi:10.5194/acp-7-6085-2007, 2007.

Riccio, A., Ciaramella, A., Giunta, G., Galmarini, S., Solazzo, E., and Potempski, S.: On the systematic reduction of data complexity in multi-model ensemble atmospheric dispersion modelling. J. Geophys. Res., 117, D05314, doi:10.1029/2011JD016503, 2012.

Schere, K., Flemming, J., Vautard, R., Chemel, C., Colette, A., Hogrefe, C., Bessagnet, B., Meleux, F., Mathur, R., Roselle, S., Hu, R.-M., Sokhi, R. S., Rao, S. T., and Galmarini, S.: Trace gas/aerosol boundary concentrations and their impacts on continental-scale AQMEII modeling domains, Atmos. Environ., 53, 38–50, 2012.

Simpson, D., Guenther, A., Hewitt, C. N., and Steinbrecher, R.: Biogenic emissions in Europe 1. Estimates and uncertainties, J. Geophys. Res., 100, 22875–22890, 1995.

Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Denier van der Gon, H., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Jeričević, A., Kraljević, L., Miranda, A. I., Nopmongcol, U., Pirovano, G., Prank, M., Riccio, A., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S., and Galmarini, S.: Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 60–74, 2012a.

Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M. D., Wyat Appel, K., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Miranda, A. I., Nopmongcol, U., Prank, M., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Operational model evaluation for particulate matter in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 75–92, 2012b.

Solazzo, E., Bianconi, R., Pirovano, G., Moran, M. D., Bellasio, R., and Galmarini, S.: Air Quality Ground-Based Observational Data for Evaluating Regional-Scale Models: The AQMEII Experience, Environmental Manager July 2012, 12–20, 2012c.

Taylor, K. E.: Summarizing multiple aspects of model performance in a simple diagram, J. Geophys. Res. 106, 7183–7192, 2001.

Tchepel, O., Monteiro, A., Ribeiro, I., Carvalho, A., Sá, E., Ferreira, J., Miranda, A. I., and Borrego, C.: Improvement of ensemble technique using spectral analysis and decomposition of air pollution data, 32st NATO/SPS International Technical Meeting on Air Pollution Modelling and its Application, 7–11 May, Utrecht, the Netherlands, 2012.

Tebaldi, C. and Knutti, R.: The use of multi-model ensemble in probabilistic climate projections, Philos. Trans. Roy. Soc., 365A, 2053–2075, 2007.

Van Loon, M., Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P. J. H., Christensen, J. H., Cuvelier, C., Graff, A., Jonson, J. E., Krol, M., Langner, J., Roberts, P., Rouil, L., Stern, R., Tarrasón, L., Thunis, P., Vignati, E., White, L., and Wind, P.: Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble, Atmos. Environ. 41, 2083–2097, 2007.

Vautard, R., Moran, M. D., Solazzo, E., Gilliam, R. C., Matthias, V., Bianconi, R., Chemel, C., Ferreira, J., Geyer, B., Hansen, A. B., Jericevic, A., Prank, M., Segers, A., Silver, J. D., Werhahn, J., Wolke, R., Rao, S. T., and Galmarini, S.: Evaluation of the meteorological forcing used for the Air Quality Model Evaluation International Initiative (AQMEII) air quality simulations, Atmos. Environ., 53, 15–37, 2012.

Wise, E. K. and Comrie, A. C.: Extending the Kolmogorov-Zurbenko filter: application to ozone, particulate matter, and meteorological trends, J. Air Waste Manag. Assoc., 55, 1208–1216, 2005.

Zurbenko, I. G.: The Spectral Analysis of Time Series, 236 pp., Amsterdam, the Netherlands, 1986.