Boundary-Layer Meteorology (2019) 171:537–555
https://doi.org/10.1007/s10546-018-0411-1

RESEARCH ARTICLE

A Concept for the Analysis and Presentation of the Ensemble Simulation Results in the UDINEE Exercise

Slawomir Potempski[1] · **Piotr Kopka**[1]

Received: 17 February 2018 / Accepted: 7 November 2018 / Published online: 4 December 2018

Abstract
We propose a general concept for the analysis of the results of urban dispersion simulations of high temporal resolution, taking into account multi-model ensembles. We are motivated by theoretical considerations related both to the characteristics of the measurements and to the representation of the multi-model ensemble. Based on typical mathematical notions, we propose and present several indices, and apply them to the results of the UDINEE dispersion-modelling exercise. We demonstrate that the median model is the proper representation of the ensemble results for the presented methodology.

Keywords Multi-model ensemble · Performance analysis · UDINEE exercise · Urban dispersion modelling

1 Introduction

The results of the UDINEE urban dispersion-modelling exercise are based on the Joint Urban 2003 (JU2003) experimental campaign conducted in Oklahoma City, U.S.A. (see Allwine and Flaherty 2006, 2007; Hernández-Ceballos et al. 2018a, b, and references therein). The observational data are characterized by a high-resolution timestep (0.5 s) and peaks in measured concentrations when the tracer traverses the station, resulting in measurement values sensitive to local fluctuations. However, as the peaks are often not long lasting, many models predict the highest tracer-concentration values at times either prior to or after the measured peaks, with some of the models underpredicting, while others overpredicting, the concentration magnitude.

Since the appearance of the ensemble technique applied to atmospheric dispersion modelling, many questions have been raised, such as how to present simulation results versus measurements, how to compare different models, and how to analyze the ensemble of the models (Straume et al. 1998; Bellasio et al. 1999; Dabberdt and Miller 2000; Delle Monache and Stull 2003; Galmarini et al. 2001, 2004). A number of indices can be applied, such as the factor-of-two index *FAC2*, which is defined as an index determining the number of model-predicted values in the range of 0.5 to 2 multiplied by the measured value, and data

✉ Slawomir Potempski
slawomir.potempski@ncbj.gov.pl

[1] National Centre for Nuclear Research, 05-400 Otwock-Swierk, Poland

can be compared using scatter diagrams and correlation coefficients. These indices have been used extensively, for example, while analyzing the results of the European Tracer Experiment (ETEX) (Girardi et al. 1998; Graziani et al. 1998; Van Dop and Nodop 1998; Mosca et al. 1998). Nevertheless, the optimal method for comparing model data and observations is not obvious. Specifically related to the results of urban dispersion simulations, this problem has been investigated, for example, in Zhou and Hanna (2007) and Hanna and Chang (2012), whose work is the basis for the analysis of the UDINEE dispersion-modelling-exercise results presented in Hernández-Ceballos et al. (2018a, b). Motivated by theoretical considerations, we propose a general approach here for comparing modelled and observed data specifically tailored to the analysis of the ensemble-dispersion models and measurements of high temporal resolution. However, the approach can be also adapted for boundary-layer processes, which can be investigated by using the ensemble-modelling technique.

Section 2 describes a general concept for model comparison, Sect. 3 deals with the proposed method for ensemble analysis, Sect. 4 presents a simple example of such an analysis related to the UDINEE exercise, and Sect. 5 contains conclusions.

2 Comparison of the Results and Presentation

Assume two datasets are $\{C_i\}_{i=0}^{n}$, $\{O_i\}_{i=0}^{n}$, where C_i and O_i are the simulated and observed concentrations, respectively, at time t_i=t$_0$ +$i\Delta t$, with i=0,…,n (the timestep Δt is fixed, but could also be variable in principle), and define interpolation functions for the interval $[t_0, t_n]$ for both the modelled data and observations in general as

$$C(t) = \text{Interpol}(\{t_i\}_{i=0}^{n}, \{C_i\}_{i=0}^{n}), \tag{1a}$$

$$O(t) = \text{Interpol}(\{t_i\}_{i=0}^{n}, \{O_i\}_{i=0}^{n}), \tag{1b}$$

where Interpol is an interpolation function (for example linear) over the whole interval $[t_0, t_n]$. The typical norm for integrable functions,

$$\|C\|_{t,\tau} = \int_{t-\tau}^{t} |C(t)|dt, \tag{2a}$$

$$\|O\|_{t,\tau} = \int_{t-\tau}^{t} |O(t)|dt, \tag{2b}$$

can be used to estimate the modelled and measured integrated concentrations, respectively, in the time interval $[t-\tau, t]$, where the parameter τ is the length of the interval over which the results are integrated (i.e. the time-integration interval).

We also define the positive and negative parts of the difference between the modelled and observed values as

$$(C-O)^{+}(t) = \max\{C(t) - O(t); 0\}, (C-O)^{-}(t) = -\min\{C(t) - O(t); 0\}, \tag{3}$$

with

$$\left\|(C-O)^{+}\right\|_{t,\tau} = \int_{t-\tau}^{t} (C-O)^{+}(t)dt, \left\|(C-O)^{-}\right\|_{t,\tau} = \int_{t-\tau}^{t} (C-O)^{-}(t)dt. \tag{4}$$

The positive (negative) part shows the integrated concentration resulting from model overprediction (underprediction) in the interval $[t-\tau, t]$. The total error between the measured and simulated concentrations in this interval is

$$\|C - O\|_{t,\tau} = \int_{t-\tau}^{t} |C(t) - O(t)|dt, \tag{5}$$

which is equal to the sum $\left\|(C-O)^{+}\right\|_{t,\tau} + \left\|(C-O)^{-}\right\|_{t,\tau}$.

Let us define the following indicators,

$$R_\tau(t) = \frac{\|C\|_{t,\tau}}{\|O\|_{t,\tau}}, \tag{6a}$$

$$N_\tau^{+}(t) = \frac{\left\|(C-O)^{+}\right\|_{t,\tau}}{\|(C-O)^{+}\|_{t,\tau} + \left\|(C-O)^{-}\right\|_{t,\tau}}, \tag{6b}$$

$$N_\tau^{-}(t) = \frac{\left\|(C-O)^{-}\right\|_{t,\tau}}{\|(C-O)^{+}\|_{t,\tau} + \left\|(C-O)^{-}\right\|_{t,\tau}}, \tag{6c}$$

where $R_\tau(t)$ is a general index determining whether the model overpredicts or underpredicts concentration values in the time interval $[t-\tau, t]$, and $N_\tau^{+}(t)$, $N_\tau^{-}(t)$ show the mutual relation between the overpredictions and underpredictions, respectively, with $N_\tau^{+}(t) + N_\tau^{-}(t) = 1$ for any t and τ.

Depending on the chosen integration interval, the index $R_\tau(t)$ can be treated either globally or locally, which shows global behaviour in the case when the interval τ is the length of the whole period considered, revealing in principle to what extent the total integrated concentration is in accordance with the observed value. It is probably more reasonable to select an integration time longer than the temporal resolution, but not too long, to still enable the capture of local effects such as peaks (depending on the duration of the peak). The index $R_\tau(t)$ can also be further analyzed statistically for all measurement stations, for example, by taking the mean, median, and standard deviation of the concentration. Note that the constant value $R_\tau(t) = 1$ demonstrates the best agreement between observations and simulated values. The comparison of simulated values with measurements can be presented as a function of time for the consecutive integration times. The same time series can be obviously used for comparison among models and for the analysis of the whole ensemble of models, which may be illustrated by presentation of hypothetical measurement and model data in Fig. 1. Figure 2 shows the values of the index $R_\tau(t)$ for an integration time five times longer than the temporal resolution of the data.

However, it is also possible to present on the same axes both the index $R_\tau(t)$ and additional information on the relation between the overpredictions and underpredictions via the indicators $N_\tau^{+}(t)$, $N_\tau^{-}(t)$ by defining an angle $\varphi(t)$ based on the ratio between these two indices $\tan\varphi(t) = \frac{N_\tau^{+}(t)}{N_\tau^{-}(t)}$. Here, the angle $\varphi(t) = \pi/4$ corresponds to an equal amount of overprediction and underprediction, with $\varphi = 0$ for only overprediction, and $\varphi = \pi/2$ for only underprediction (see Fig. 3). For clearer presentation, we transform this angle to $\psi = 2\varphi$, giving $\psi = \pi/2$ as the balance between over- and underprediction, where $\psi < \pi/2$ for overprediction, and $\psi > \pi/2$ for underprediction. Note that the value $\psi = \pi/2$ does not imply the absence of overprediction or underprediction, but only an equivalent degree of overprediction and underprediction.

Now we present values in the polar coordinate system $(R_\tau(t), \psi_\tau(t))$ in the upper half-space as $0 \le \psi_\tau(t) \le \pi$. The observational values are at the fixed point $(1, \pi/2)$ (i.e. (0, 1)

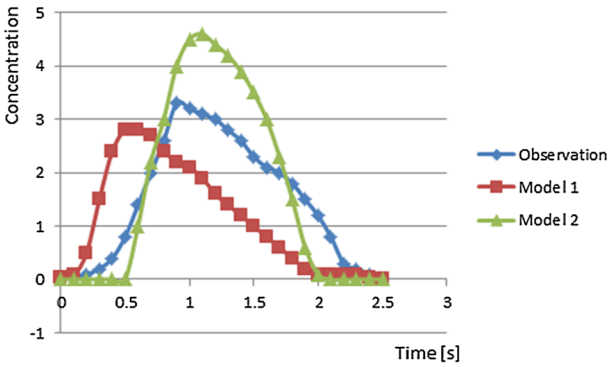

Fig. 1 Hypothetical observations and model results as a function of time with a temporal resolution of 0.1 s

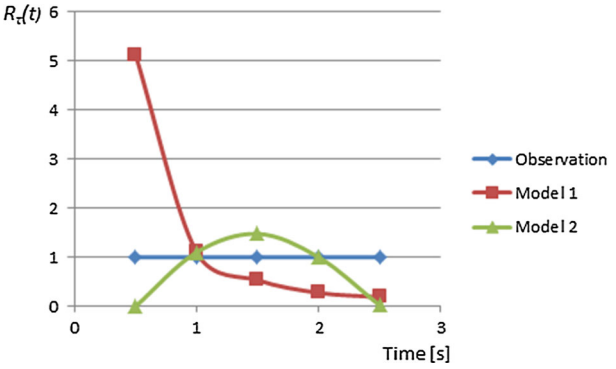

Fig. 2 The index $R_\tau(t)$ for the integration time $\tau = 0.5$ s (five times longer than the temporal resolution)

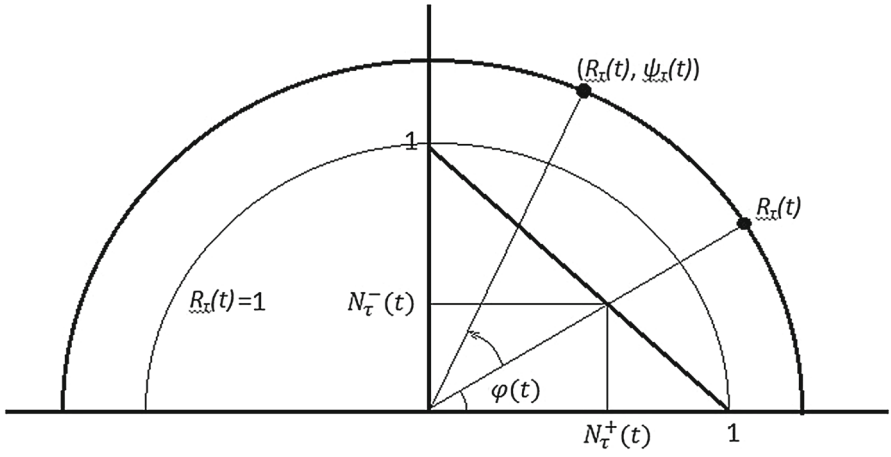

Fig. 3 Construction of the coordinates $(R_\tau(t), \psi_\tau(t))$

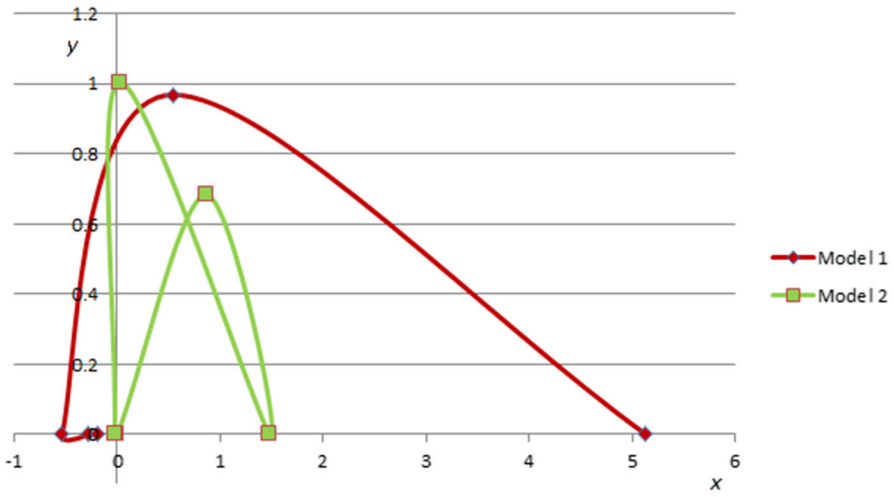

Fig. 4 The curve $t \rightarrow (R_\tau(t), \psi_\tau(t))$ for two hypothetical models (integration time $\tau = 0.5$ s), with $x = R_\tau(t)\cos\psi_\tau(t)$ and $y = R_\tau(t)\sin\psi_\tau(t)$

in the Cartesian coordinate system) for all times and the integration interval τ. The function $t \rightarrow (R_\tau(t), \psi_\tau(t))$ describes how far and on which side of the observational value the model-predicted concentrations are located: points located on the right-hand (left-hand) side imply a greater overprediction (underprediction). The arc value $R_\tau(t) = 1$ shows the threshold between the total overpredicted and underpredicted integrated concentrations. One can also add curves presenting factors of two or five—if necessary, a logarithmic scale can also be used. As an example, the curve $t \rightarrow (R_\tau(t), \psi_\tau(t))$ with an integration time of $\tau = 0.5$ s is presented in Fig. 4 (for the hypothetical data shown in Fig. 1), noting that the measurement point is always at a fixed point (at (0, 1) in the Cartesian coordinate system). In Fig. 4, two $t \rightarrow (R_\tau(t), \psi_\tau(t))$ curves are shown for two models: the scales of the axes are not preserved for better visualization. Such a graph enables observation of the behaviour of the models in consecutive timesteps for the assumed integration time.

A few general remarks on this presentation method:

1. In some cases, it can be relevant to consider only measurements and model data above some threshold. A possible reason can be the removal of low values considered as noise, or the interest in simply identifying peaks in the data (then a percentage of the peak value can be defined as the threshold), or the situation when the exceedance of some limit values is the main purpose of the analysis. When a threshold value is used, an appropriate choice of the analyzed time period should be made, for example, by taking the first and the last time points when the threshold is exceeded, for either observations or model data.

2. The time-integration interval τ can be any value starting from the timestep Δt up to the whole interval $[t_0, t_n]$. In any case, integration over an interval is related to taking an average for this interval. Obviously, a shorter interval time enables a more detailed analysis.

3. One of the general drawbacks of such an analysis is related to the fact that it is based on only point measurements. Since the model spatial resolution is usually only a few metres, and concentration peaks can appear in a short period of time, the peak concentration predicted by the model may also be shifted in space by a few metres compared with the measured peak

concentration. Therefore, it would be reasonable to consider an average concentration over some area, and formally this leads to the relation (instead of Eq. 5),

$$\|C - O\|_{t,\tau} = \int_{t-\tau}^{t} \left| \frac{1}{|A|} \int_A C(t,x)dx - O(t) \right| dt, \tag{5'}$$

where the model concentration is integrated over some area A and averaged, with $|A|$ the area measure. Actually, while models often use such a formulation, the choice of the value of A can be a delicate matter.

3 How to Analyze the Ensemble of the Models?

While there are various ways to perform an analysis of the whole model ensemble, statistical analysis is usually the preferred method. However, the question arises as to which indicators should be applied, and what their meaning would be? Here, the error between the measurement O and the single model M is determined by the integral $\|M - O\|_{t,\tau} = \int_{t-\tau}^{t} |M(t) - O(t)|dt$. Suppose we are interested in finding the representative function for the ensemble of the models. Taking an analogous measurement, we seek a function that minimizes the error between this representative and the models—the ensemble members. In other words, we look for the function defined by the following optimization problem (see also Galmarini and Potempski 2012), i.e. find the function M^* such that

$$\sum_{j=1}^{m} \left\| M^* - M_j \right\|_{t,\tau} = \inf_M \sum_{j=1}^{m} \left\| M - M_j \right\|_{t,\tau},$$

where $\{M_j\}$ is the set of the (interpolation) functions representing the results from m models.

The quantity $\sum_{j=1}^{m} \left\| M^* - M_j \right\|_{t,\tau}$ can be considered as a total spread of the ensemble—hence logically the function M^*, which is supposed to reflect the behaviour of the whole ensemble, should be chosen to minimize this spread. It should be stressed that this does not necessarily imply that the value $\|M^* - O\|_{t,\tau}$ minimizes the error between the ensemble representative M^* and the measurement. The function M^* is that which best characterizes the full set of ensemble models based on this measure.

The relation between the ensemble spread and the error between the measurement and ensemble representative can be expressed using the following inequalities,

$$\left\| M^* - O \right\|_{t,\tau} \le \left\| M^* - M_j \right\|_{t,\tau} + \left\| M_j - O \right\|_{t,\tau}, \text{ for } j = 1, \ldots, m, \tag{7}$$

which are summed to yield

$$\left\| M^* - O \right\|_{t,\tau} \le \frac{1}{m} \sum_{j=1}^{m} \left\| M^* - M_j \right\|_{t,\tau} + \frac{1}{m} \sum_{j=1}^{m} \left\| M_j - O \right\|_{t,\tau}. \tag{8}$$

The above relation shows that the error between ensemble representative and the measurements can be estimated by the average spread and the average error between the ensemble models and observations. The meaning of this relation is analogous to the accuracy-diversity equation applied in many various contexts, such as in the fields of machine learning and neural networks (see Krogh and Vedelsby 1995) or Optiz and Shavlik 1996), where this type of expression is specifically used for the ensemble mean and mean squared error. The norms we use here are related rather to the time-integrated concentration (or doses) than to the concentration itself.

As the second term in (8) is independent of the chosen ensemble representative, it is justified that the function minimizing the first term of the right-hand side (i.e. the average spread) should be taken as M^*. To prove that the median function is the solution of the minimization problem posed above, let us first define the median function formally,

$$\mathrm{Med}(t) = \begin{cases} M_{\left[\frac{m}{2}\right]+1}(t) & \text{if } m \text{ is odd} \\ \dfrac{M_{\left[\frac{m}{2}\right]}(t) + M_{\left[\frac{m}{2}\right]+1}(t)}{2} & \text{if } m \text{ is even} \end{cases} \tag{9}$$

where $[x]$ indicates the highest integer number $<x$ (in fact, any value between $M_{\left[\frac{m}{2}\right]} + M_{\left[\frac{m}{2}\right]+1}$ can be chosen for an even number of models).

We need to show that for any function M, we have

$$\sum_{j=1}^{m} \left\| \mathrm{Med} - M_j \right\|_{t,\tau} \le \sum_{j=1}^{m} \left\| M - M_j \right\|_{t,\tau}, \tag{10}$$

and since

$$\sum_{j=1}^{m} \left\| M - M_j \right\|_{t,\tau} = \sum_{j=1}^{m} \int_{t-\tau}^{t} |M(t) - M_j(t)| dt = \int_{t-\tau}^{t} \sum_{j=1}^{m} \left| M(t) - M_j(t) \right| dt, \tag{11}$$

it is sufficient to show that the sum $\sum_{j=1}^{m} \left| M(t) - M_j(t) \right|$ for each point t is minimized by the median. Without loss of generality, we assume that the values $M_j(t)$ are in ascending order, i.e., $M_j(t) \le M_{j+1}(t)$ for any fixed point t. Obviously, for any value v outside the interval $[M_1(t), M_m(t)]$, we have $\sum_{j=1}^{m} \left| \mathrm{Med}(t) - M_j(t) \right| \le \sum_{j=1}^{m} \left| v - M_j(t) \right|$. Hence, suppose that $v \in [M_k, M_{k+1}]$ for some k, and assume that $k \le \left[\frac{m}{2}\right]$, then $\sum_{j=1}^{m} \left| \mathrm{Med}(t) - M_j(t) \right| = \sum_{j=1}^{[m/2]} \left(M_{m-j+1}(t) - M_j(t) \right)$, so that for the value $v \in [M_k, M_{k+1}]$, we have

$$\begin{aligned} \sum_{j=1}^{m} \left| v - M_j(t) \right| &= \sum_{j=1}^{k} \left(M_{m-j+1}(t) - M_j(t) \right) + \sum_{j=k+1}^{[m/2]} \left[\left(M_j(t) - v \right) + \left(M_{m-j+1}(t) - v \right) \right] \\ &= \sum_{j=1}^{k} \left(M_{m-j+1}(t) - M_j(t) \right) + \sum_{j=k+1}^{[m/2]} \left[\left(M_{m-j+1}(t) - M_j(t) \right) + 2\left(M_j(t) - v \right) \right] \end{aligned} \tag{12}$$

As the last term in the second sum is always positive, then $\sum_{j=1}^{m} \left| \mathrm{Med}(t) - M_j(t) \right| \le \sum_{j=1}^{m} \left| v - M_j(t) \right|$. The case $k > m/2$ can be treated analogously, and, as such, consideration is valid for any time t and v, which proves (10).

The main purpose of using an ensemble approach concerns the problem of model predictability. It is expected that the average of the ensemble represents the most probable realization of physical processes, while the spread is related to the inherent uncertainty, and shows the range of other possible realizations. Having this in mind, we assume that the ensemble spread is an indicator representing the uncertainty of the results. In order to examine this quantitatively, we use the following quantity,

$$\frac{\sum_{j=1}^{m} \left\| M^* - M_j \right\|_{t,\tau}}{\left\| M^* \right\|_{t,\tau}} \tag{13}$$

or the average spread, to observe the degree of discrepancy between models in comparison with the average concentration determined by the ensemble (according to the above consideration, the median is a good choice for M^*). These indicators depend on the problem under consideration—in principle, if the spread is a small percentage of the concentration values, then there is quite good agreement among models, which suggests low uncertainty. In contrast, several reasons can cause a higher spread, such as inherent uncertainties associated with the problem, limitations of the models, and various difficulties in modelling physical phenomena. In the case when measurements are additionally available, taking into account the inequality (8), one can check the relation between the ensemble spread $\sum_{j=1}^{m} \left\| M^* - M_j \right\|_{t,\tau}$ and the ensemble error $\sum_{j=1}^{m} \left\| M_j - O \right\|_{t,\tau}$ or $\|M^* - O\|_{t,\tau}$, representing the error of the whole ensemble, where the median function can be taken as M^*.

Three cases can be considered:

- the ensemble spread is small in comparison with the ensemble error: this shows the situation when there are probably more fundamental difficulties with modelling the problem;
- the ensemble spread and the ensemble error are comparable: this is the case when similar agreement is within the ensemble and with the measurements; hence, the uncertainty should not be too high;
- the ensemble spread is high in comparison with the ensemble error: this indicates high uncertainty, which could be caused by different factors, as already mentioned.

We now present a few additional remarks concerning the representation of the ensemble results. If we are interested in the root-mean-square error expressed by the norm in L^2 space, i.e. $\|M - O\|_{t,\tau} = \left(\int_{t-\tau}^{t} |M(t) - O(t)|^2 dt \right)^{1/2}$ and, consequently, wish to find the representative function by looking for the solution of the optimization problem, $\sqrt{\sum_{j=1}^{m} \left\| M^* - M_j \right\|_{t,\tau}^2} = \inf_M \sqrt{\sum_{j=1}^{m} \left\| M - M_j \right\|_{t,\tau}^2}$, it can then be shown that the mean function $\text{Mean}(t) = \frac{1}{m} \sum_{j=1}^{m} M_j(t)$ should be the representative function for the whole ensemble. The simplest formal proof can be made by finding the derivative of the function (in some space function) $F(M) = \int_{t-\tau}^{t} \sum_{j=1}^{m} \big(M(t) - M_j(t)\big)^2 dt$ using the Gateaux derivative,

$$\frac{dF(M + hP)}{dh}|_{h=0} = \int_{t-\tau}^{t} \sum_{j=1}^{m} \frac{d}{dh} \big(M(t) + hP(t) - M_j(t)\big)^2 |_{h=0} dt =$$

$$\int_{t-\tau}^{t} 2 \sum_{j=1}^{m} \big(M(t) - M_j(t)\big) P(t) dt = 0. \tag{14}$$

As this equation should be valid for any function $P(t)$, we immediately obtain $\sum_{j=1}^{m} \big(M(t) - M_j(t)\big) = 0$, which produces the relation for the mean. A general framework for finding the optimal combination of ensembles can be found in Potempski and Galmarini (2009).

Similarly, one can use the supremum norm (Chebyshev norm) for the error, i.e. $\|M - O\|_{t,\tau} = \sup_{z \in [t-\tau, t]} |M(z) - O(z)|$ and, consequently, consider the optimization problem for this norm. It is easy to check that the midpoint function

$$\mathrm{Mid}(t) = \frac{\min\limits_j M_j(t) + \max\limits_j M_j(t)}{2} \tag{15}$$

is the solution of this minimization problem.

Summarizing all these points, different representations of the ensemble are related to a different metric applied for measuring the spread of the ensemble, which, due to the properties of these metrics, can be expressed in the following ways:

a. The midpoint with the corresponding spread (expressed as $\sup_{z\in[t-\tau,t]}\left|\mathrm{Mid}(z) - M_j(z)\right|$) defines the rectangular region containing all model values. This spread can be considered as the worst-case scenario as it shows the maximum discrepancy between the ensemble representative and the model, and, hence, is the most sensitive to the outliers.
b. The mean value with its spread $\sqrt{\sum_{j=1}^{m}\left\|\mathrm{Mean} - M_j\right\|_{t,\tau}^2}$ defines the circle containing the model values, which would correspond to the minimization of the variance if the model results were treated as random variables, and, hence, in some sense, gives the value loaded with the smallest uncertainty. This also corresponds to the fact that the mean squared error for the ensemble mean is less than that for any ensemble member (a more exhaustive explanation is given in Rougier 2016). In relation to climate models, see Christiansen (2018), where the detailed investigations are based on the assumptions that the models have a normal distribution with different variances. Concerning the problem of dimensionality described in Christiansen (2018), one can also mention Riccio et al. (2012), where the problem of the reduction of data complexity has been investigated to deal with this issue.
c. The median is the ensemble representative being possibly the least sensitive to the outliers. A general property of the median is such that it minimizes the mean absolute error associated with the random variable, i.e., it corresponds in some way to the bias.

However, there are also other possible less sensitive ensemble representatives than the mean, such as the winsorized mean or the trimmed mean. In both cases, the first step is to choose the limit percentile and either replace the outliers defined by this percentile by the nearest values within the percentile (for the winsorized mean) or simply remove them (for the trimmed mean) before calculating the mean. However, while this approach needs the definition of the limit percentile that, in general, is case dependent, the method can nonetheless be useful as a reasonable estimator if properly applied. The winsorized mean can also be presented in the form based on the notion of norms, but with a specific weighted norm that depends on the chosen percentile.

4 Example for the UDINEE Exercise

Here, a simple example of the application of the analysis method described above is demonstrated for the UDINEE exercise, using data from the second puff release of IOP4 (IOP – intensive operating period: time horizon of puff release) at sensor location L11, the fourth puff of IOP6 at station L15, and the first and second releases of IOP9 at the stations L05 and L17, respectively (see Fig. 5, and see Hernández-Ceballos et al. (2018a, b) for a full description of the UDINEE project). This selection was made because of the relatively good quality of the measurements. The main peaks are observed in the following periods: 160–180 s for IOP4, 110–180 s for IOP6, 330–370 s for the first puff of IOP9, and 210–320 s for the second puff of IOP9. Data are available from five models in the first two cases, and from six models in the last two cases.

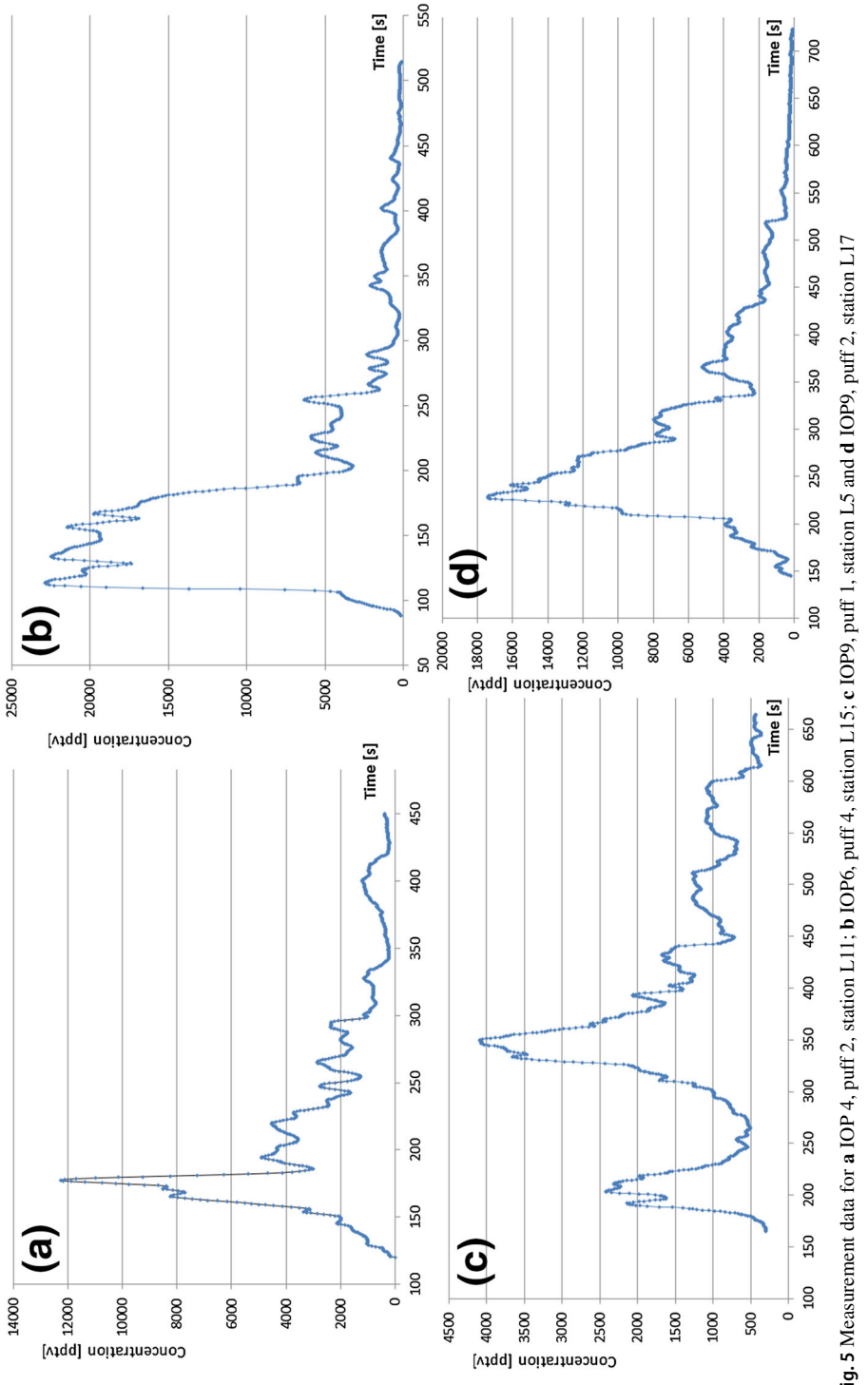

Fig. 5 Measurement data for **a** IOP 4, puff 2, station L11; **b** IOP6, puff 4, station L15; **c** IOP9, puff 1, station L5 and **d** IOP9, puff 2, station L17

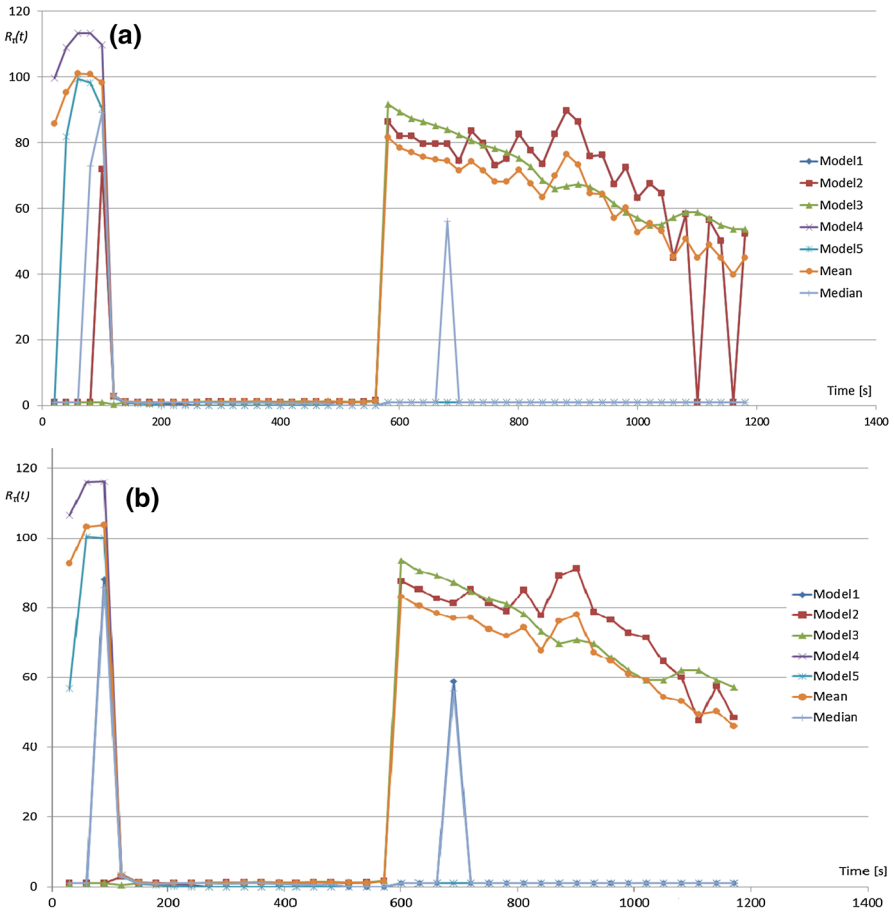

Fig. 6 Indicator values $R_\tau(t)$ for, **a** $\tau = 20$ s, and **b** $\tau = 30$ s for IOP 4, puff 2, station L11

First we present in Fig. 6 the values of the index $R_\tau(t)$ for $\tau = 20$ s and $\tau = 30$ s in consecutive timesteps for the five models, the mean and median for IOP4 (due to the high values, logarithmic values of concentration have been used in Fig. 6). Both values of τ give a decrease in the values of the index $R_\tau(t)$ starting at about 150 s, with higher values observed for several models starting from 600 s, illustrating that the choice of τ is important—a sharp peak for model 2 at time 100 s appears only in Fig. 6a for the shorter interval time $\tau = 20$ s (model 2 also produces two zero values from 1000–1200 s for $\tau = 20$ s).

Values of $R_\tau(t)$ for $\tau = 30$ s are presented in Fig. 7 for IOP6 and two puffs of IOP9, showing the common feature that, in the peak period, the values of $R_\tau(t)$ for all models improves (i.e. closer to one), but outside the peak periods, there are models when this value is reduced. In Fig. 8, $t \to (R_\tau(t), \psi_\tau(t))$ curves for IOP4 (corresponding to Fig. 6b) are shown for times up to 720 s, illustrating that most of the values are located on the right-hand side of the diagram, which corresponds in general to model overprediction. Restricting the time horizon to the interval from 130–230 s in Fig. 9, corresponding to the time of peak concentration, we demonstrate that the curves for $\tau = 10$ s have a better time resolution.

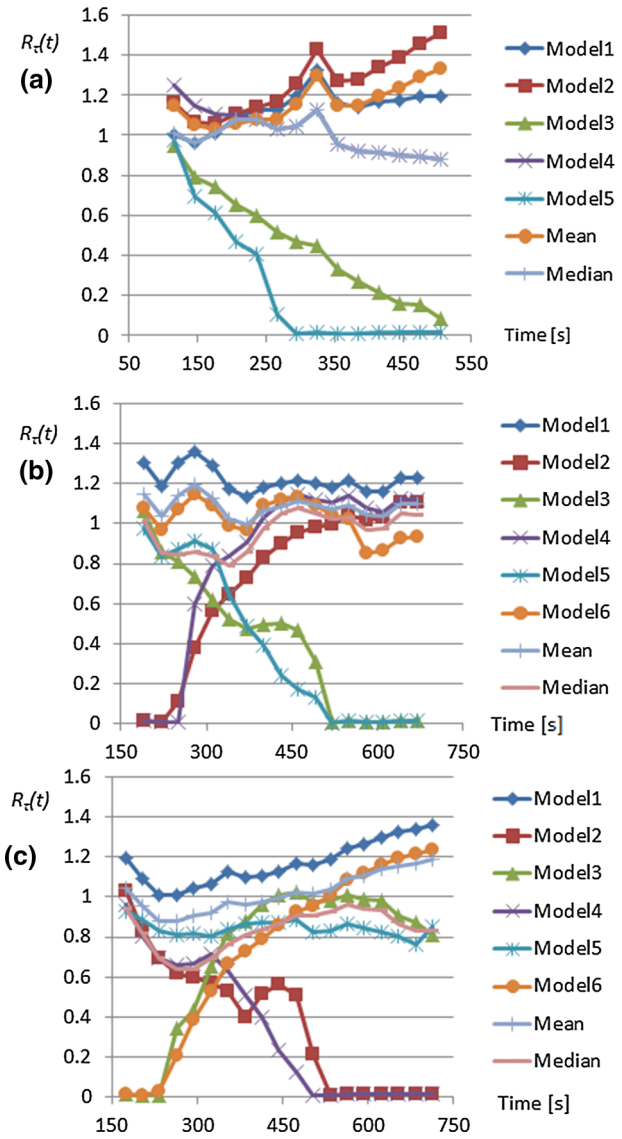

Fig. 7 Indicator values $R_\tau(t)$ for $\tau = 30$ s: **a** IOP6, puff 6, station L15; **b** IOP9, puff 1, station L5; **c** IOP9, puff 2, station L17

From these plots, one concludes that model results are closer to the measurements in the time frame corresponding to the peak time than in the other time periods. In general, the median model better characterizes the whole ensemble than the mean model, which is much more sensitive to the peculiar values of a single model. Similar diagrams are presented in Fig. 10 for IOP6, and two puff releases of IOP9 (for $\tau = 10$ s), illustrating the median as better representing the ensemble than the mean. An extreme case can be observed for IOP6 (Fig. 10a) when model 2 forces the mean to completely overpredict the concentration.

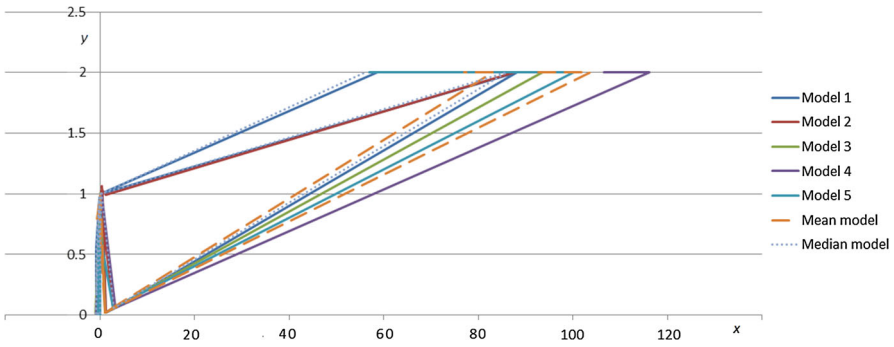

Fig. 8 The $t \rightarrow (R_\tau(t), \psi_\tau(t))$ curves for $\tau = 30$ s for IOP 4, puff 2, station L11

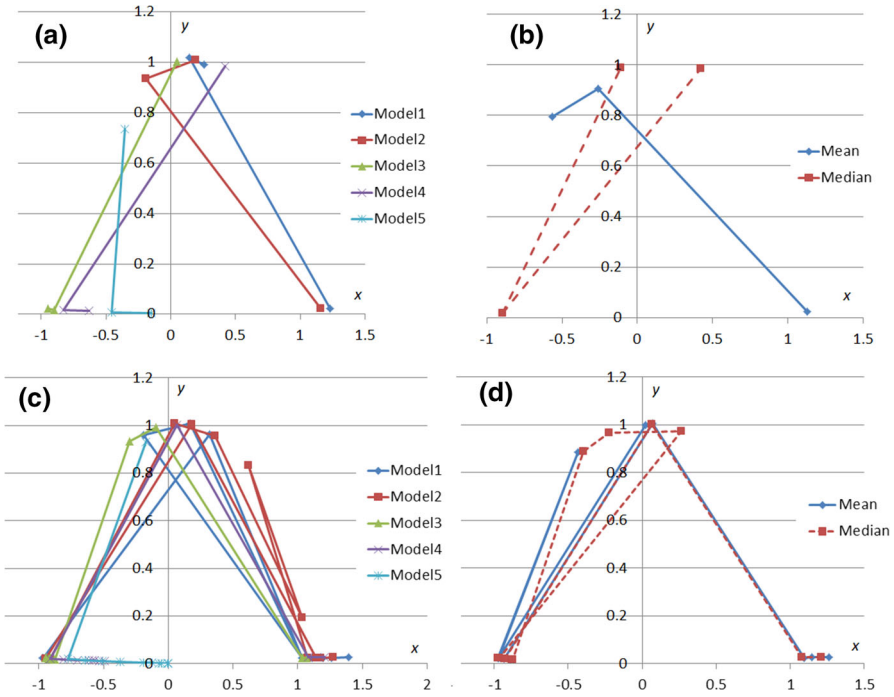

Fig. 9 The $t \rightarrow (R_\tau(t), \psi_\tau(t))$ curves for IOP 4, puff 2, station L11 from 130–230 s showing the five models in the left column: (**a**) and (**c**), and the mean and median models in the right column: (**b**) and (**d**); $\tau = 30$ s – graphs (**a**) and (**b**), $\tau = 10$ s –plots (**c**) and (**d**)

A cumulative column diagram enables a detailed analysis of the behaviour of the models, for which Fig. 11 shows overprediction and underprediction in consecutive timesteps for the integration time $\tau = 20$ s for five models (see the diagram for the precise values of $N_\tau^+(t)$, $N_\tau^-(t)$). The mean and median of the ensemble for IOP4, IOP6, and two puffs of IOP9 are presented in Fig. 12.

The mean model generally gives higher values than the median model, which is particularly distinct for IOP6, and is explained by the deterioration of the mean by model 2 (see Fig. 10). Interestingly, at the times of peak concentration, more underprediction is observed. As models

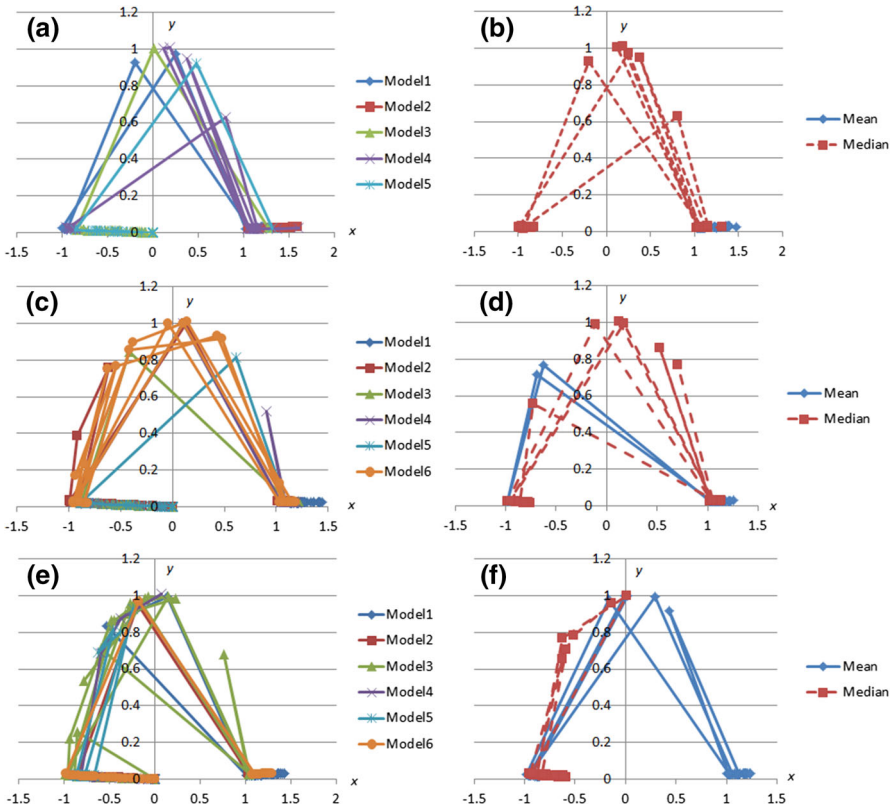

Fig. 10 The $t \rightarrow (R_\tau(t), \psi_\tau(t))$ curves for IOP6 4, puff 4, station L11 (**a**) and (**b**); IOP9, puff 1, station L5 (**c**) and (**d**); IOP9, puff 2, station L17 (**e**) and (**f**). Presented are the ensemble members in the left column (**a**), (**c**), (**e**), and the mean and median in the right column (**b**), (**d**), (**f**) for $\tau = 10$ s (IOP time periods as for Fig. 7)

have difficulties in timing the exact peak, if the integrated values are considered (doses), the models better predict the higher values (which, in principle, is generally true), and the results are more dispersed in time. The reason for such behaviour may be because of difficulties in modelling very local phenomena, and the usage of parametrizations based on averaged values in both time and space. As already mentioned, another problem is related to the point measurements; more adequate would be to compare the model results with measurements from devices able to perform area scans.

Finally for IOP4 and for IOP6, Fig. 13 compares the spread with the median and the error of the ensemble (expressed as the error between the observation and median) in terms of the ratios between the respective quantities, illustrating that the spread is generally high, and usually the better agreement between the models is in the time interval related to the peak period; the spread is also higher in comparison with the ensemble error (again the smallest spread is in the peak period). In contrast, the spread may also depend on the value of the concentration—if the peak values are not high and sharp enough, the impact of noise can usually be observed. In general, this confirms the previous finding that the ensemble behaves better in the peak interval than in other periods.

While this simple analysis cannot be treated as representative of the results of the whole UDINEE modelling exercise, it does give indications relating to the behaviour of the ensem-

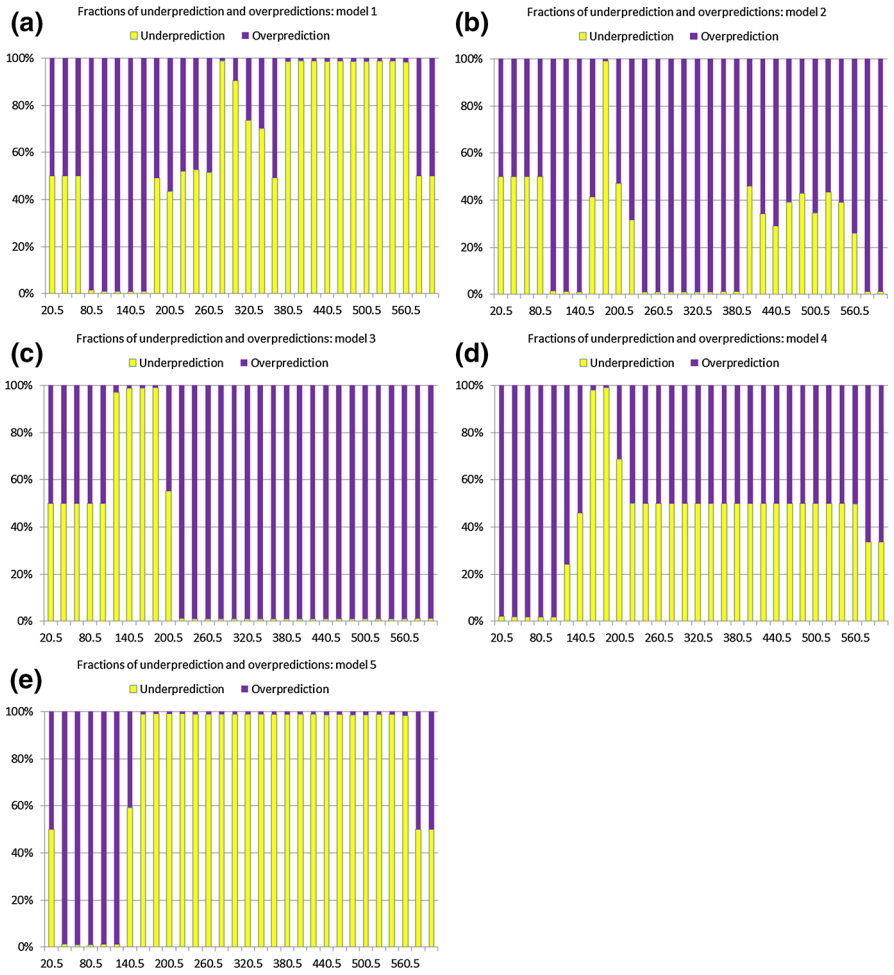

Fig. 11 Cumulative column diagram showing the overprediction and underprediction for five models for the case IOP4

ble, and illustrates one of the possible ways for performing such an analysis, with a valuable feature being the possibility of using various integration intervals in a unified way, while observing the variability in time. In most presentations, global indicators are used when sometimes it is difficult to catch some nuances.

5 Conclusions

Many different indicators can be used for the analysis and presentation of high-resolution results of dispersion simulations, including the model ensemble. A general concept has been elaborated here based on typical mathematical notions. The methodology is unified in the sense that it can be applied for different degrees of accuracy, which can be defined by changing simple parameters, such as the integration time and the time horizon of the

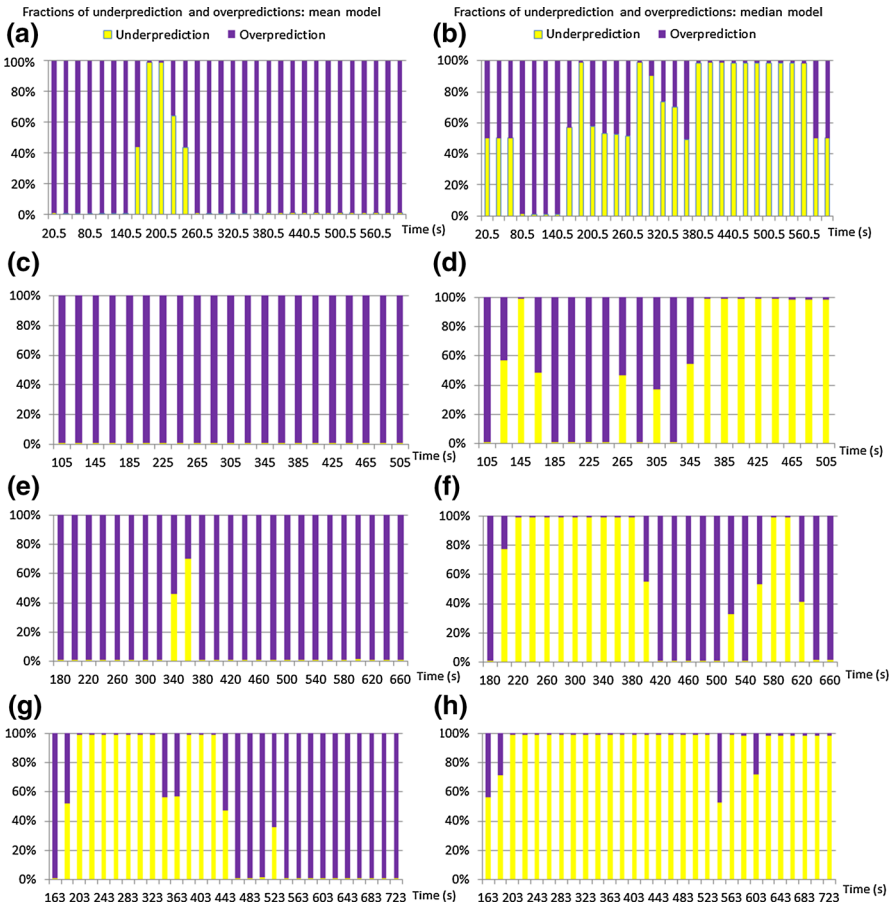

Fig. 12 Cumulative column diagram showing the overprediction and underprediction for the mean (left column) and median (right column) for cases IOP4 station L11 (**a**), (**b**), IOP6 station L16 (**c**), (**d**) and two puffs of IOP9 stations L05 and L17, respectively (**e**), (**f**) and (**g**), (**h**)

analysis, with the proper selection of the period of the integrated concentration playing an important role. However, in general, the method enables analysis of the ensemble results by taking into account the variability in time of the required precision. Simple theoretical considerations have shown that the median model can be treated as representative of the ensemble for this type of analysis. The multi-model ensemble approach gives additional information related to the uncertainty of the simulation results, while finding areas requiring further model improvement. Application of the presented method to selected cases from the UDINEE dispersion-modelling exercise reveals that, although the analysis cannot be treated as representative for all cases, the method describes the typical behaviour of the ensemble fairly well, with a quite large spread of model predicted concentrations in general, and better agreement in the time window related to the measured peak concentration. The proposed method may also be applied to other types of ensembles, such as those built by perturbing the initial parameters, and can be used in sensitivity studies. The sensitivity analysis for each

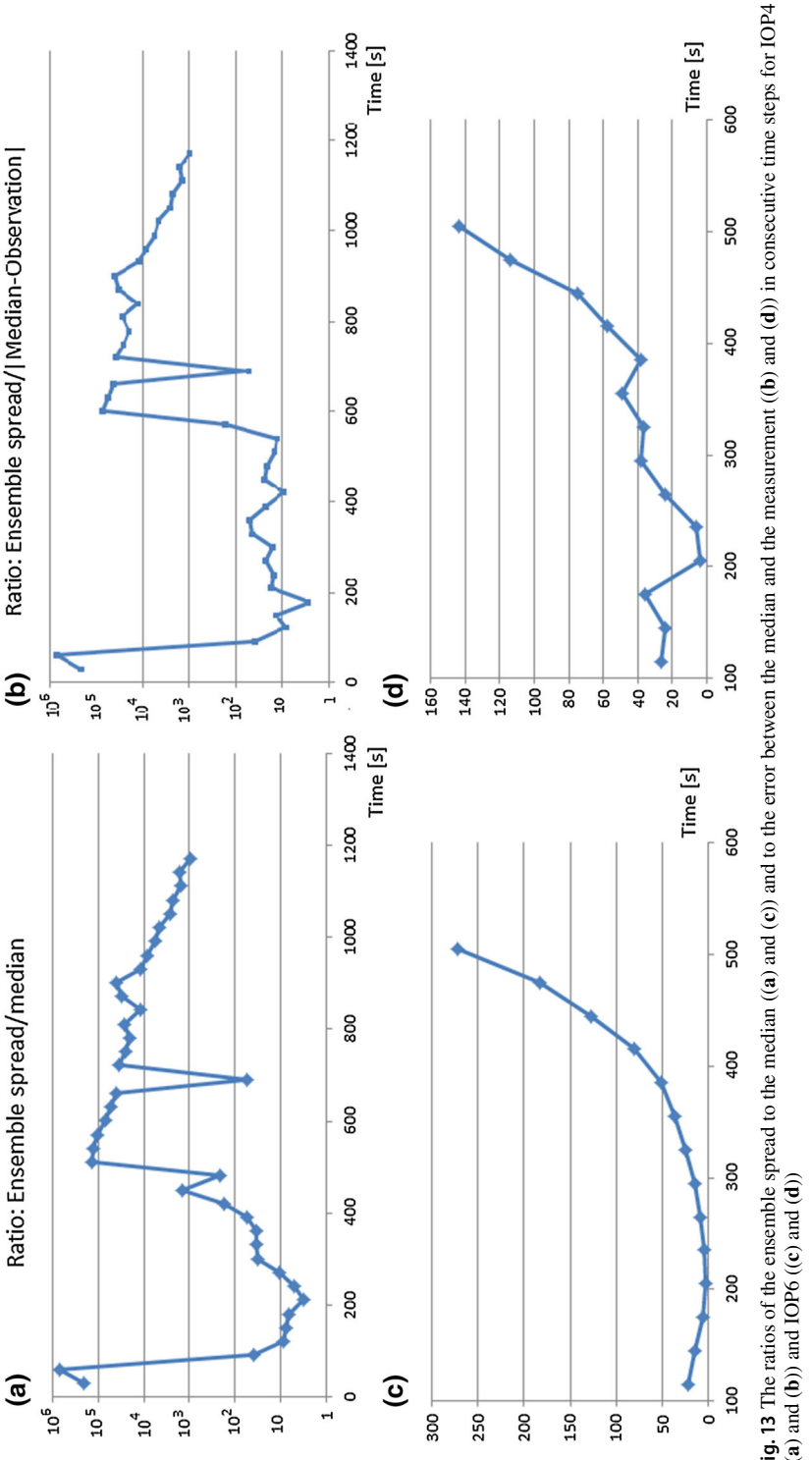

Fig. 13 The ratios of the ensemble spread to the median ((**a**) and (**c**)) and to the error between the median and the measurement ((**b**) and (**d**)) in consecutive time steps for IOP4 ((**a**) and (**b**)) and IOP6 ((**c**) and (**d**))

model can be very useful in the construction of an optimal ensemble (see Potempski and Galmarini 2009).

For this type of experiment, it would generally be better if the observations were a continuous function in time and space, resulting from, for example, a very dense network of sensors combined with interpolation based on geospatial techniques, which can also give quite a good estimation of the measurement uncertainties, or devices scanning the concentration over some area.

Acknowledgements We gratefully acknowledge the European Commission Directorate General for Migration and Home Affairs (DG HOME) for their support for the Urban Dispersion International Evaluation Exercise (UDINEE) activity. The authors wish to acknowledge the contribution of various groups to the UDINEE exercise. The following agencies have prepared the datasets used in this study: U.S. Army Dugway Proving Ground as manager of the JU2003 database; data from the tracer monitoring stations were provided by the National Oceanic and Atmospheric Administration Air Resources Laboratory Field Research Division; data from meteorological monitoring stations were provided by the Dugway Proving Ground. The Joint Research Center Ispra/Institute for Environment and Sustainability provided its ENSEMBLE system for the model output harmonization, analyses and evaluation.

References

Allwine KJ, Flaherty J (2006) Joint urban 2003: study overview and instrument locations. Technical Report, Pacific Northwest National Lab., Richland, WA. PNNL-15967

Allwine KJ, Flaherty J (2007) Urban dispersion program overview and MID05 field study summary. Technical Report, Pacific Northwest National Lab., Richland, WA. PNNL-16696

Bellasio R, Bianconi R, Graziani G, Mosca S (1999) RTMOD: an Internet based system to analyse the predictions of long-range atmospheric dispersion models. Comput Geoseci 25:819–833

Christiansen B (2018) Ensemble averaging and the curse of dimensionality. J Clim 31(4):1587–1596

Dabberdt WF, Miller E (2000) Uncertainty, ensembles and air quality dispersion modelling: applications and challenges. Atmos Environ 34:4667–4673

Delle Monache L, Stull RB (2003) An ensemble air-quality forecast over Europe during an ozone episode. Atmos Environ 37:3469–3474

Galmarini S, Potempski S (2012) Multi-model ensembles: metrics, indexes, data assimilation and all that jazz. In: Steyn DG, Trini Castelli S (eds) Air pollution modeling and its application XXI, vol 4(4). NATO science for peace and security series C: environmental security. Springer, Dordrecht, pp 419–426

Galmarini S, Bianconi R, Bellasio R, Graziani G (2001) Forecasting the consequences of accidental releases of radionuclides in the atmosphere from ensemble dispersion modelling. J Environ Radioact 57:203–219

Galmarini S, Bianconi R, Klug W, Mikkelsen T, Addis R, Andronopoulos S, Astrup P, Baklanov A, Bartnicki J, Bartzis JC, Bellasio R, Bompay F, Buckley R, Bouzom M, Champion H, D'Amours R, Davakis E, Eleveld H, Geertsema GT, Glaab H, Kollax M, Ilvonen M, Manning A, Pechinger U, Persson C, Polreich E, Potemski S, Prodanova M, Saltbones J, Slaper H, Sofiev MA, Syrakov D, Sørensen JH, Van der Auwera L, Valkama I, Zelazny R (2004) Ensemble dispersion forecasting, part 1: concept, approach and indicators. Atmos Environ 38(28):4607–4617

Girardi F, Graziani G, van Veltzen D, Galmarini S, Mosca S, Bianconi R, Bellasio R, Klug W (1998) The ETEX project. EUR Report 181-43 EN. Office for official publication of the European Communities, Luxembourg

Graziani G, Galmarini S, Grippa G, Klug W (1998), Real-time long-range dispersion model evaluation of the ETEX second release. EUR 17755 EN, Office for Official Publications of the European Communities, Luxembourg, ISBN 92-828-3656-8, 252 pp

Hanna SR, Chang J (2012) Acceptance criteria for urban dispersion model evaluation. Meteorol Atmos Phys 116:133–146

Hernández-Ceballos M, Hanna S, Bianconi R, Bellasio R, Mazzola T, Chang J, Andronopoulos S, Armand P, Benbouta N, Bourgouin P, Carný P, Ek N., Fojcíková E, Fry R, Huggett L, Kopka P, Korycki M, Lipták L, Millington S, Miner S, Oldrini O, Potempski S, Tinarelli G, Trini-Castelli S, Venetsanos A, Galmarini S (2018a) UDINEE: evaluation of multiple models with data from the JU2003 puff releases in Oklahoma City. Part I: comparison of observed and predicted concentrations. Boundary-Layer Meteorol (**submitted for publication**)

Hernández-Ceballos M, Hanna S, Bianconi R, Bellasio R, Mazzola T, Chang J, Andronopoulos S, Armand P, Benbouta N, Bourgouin P, Carný P, Ek, Fojcíková E, Fry R, Huggett L, Kopka P, Korycki M, Lipták L, Millington S, Miner S, Oldrini O, Potempski S, Tinarelli G, Trini-Castelli S, Venetsanos A, Galmarini S (2018b) UDINEE: evaluation of models with data from the JU2003 instantaneous releases in Oklahoma City. Part II: simulation of puff parameters. Boundary-Layer Meteorol (**submitted for publication**)

Krogh A, Vedelsby J (1995) Neural network ensembles, cross validation, and active learning. In: Tesauro G, Touretzky D, Leen T (eds) Advances in neural information, processing systems, vol 7. MIT Press, Cambridge

Mosca S, Graziani G, Klug W, Bellasio R, Bianconi R (1998) A statistical methodology for the evaluation of long-range dispersion models: an application to the ETEX exercise. Atmos Environ 32(24):4307–4324

Opitz D, Shavlik J (1996) Generating accurate and diverse members of a neural-network ensemble. In: Touretzky DS, Mozer MC, Hasselmo MC (eds) Advances in neural information, processing systems 8. MIT Press, Denver, pp 535–543

Potempski S, Galmarini S (2009) Est modus in rebus: analytical properties of multi-model ensembles. Atmos Chem Phys 9:9471–9489

Riccio A, Ciaramella A, Giunta G, Galmarini S, Potempski S (2012) On the systematic reduction of data complexity in multimodel atmospheric dispersion ensemble modeling. J Geophys Res Atmos 117(d5):D05314

Rougier J (2016) Ensemble averaging and the mean squared error. J Clim 29(24):8865–8870

Straume AG, N'dri Koffi E, Nodop K (1998) Dispersion modeling using ensemble forecasts compared to ETEX measurements. J Appl Meteorol 37(11):1444–1456

Van Dop H, Nodop K (1998) ETEX: a European tracer experiment. Atmos Environ 24:4089–4378

Zhou Y, Hanna SR (2007) Along-wind dispersion of puffs released in a built-up urban area. Boundary-Layer Meteorol 125:469–486