Fuzzy Similarity-Based Hierarchical Clustering for Atmospheric Pollutants Prediction

F. Camastra[1], A. Ciaramella[1](✉), L. H. Son[2], A. Riccio[1], and A. Staiano[1]

[1] Department of Science and Technology, University of Naples "Parthenope",
Isola C4, Centro Direzionale, 80143 Naples (NA), Italy
{francesco.camastra,angelo.ciaramella,angelo.riccio,
antonino.staiano}@uniparthenope.it

[2] Vietnam National University, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam
sonlh@vnu.edu.vn

Abstract. This work focuses on models selection in a multi-model air quality ensemble system. The models are operational long-range transport and dispersion models used for the real-time simulation of pollutant dispersion or the accidental release of radioactive nuclides in the atmosphere. In this context, a methodology based on temporal hierarchical agglomeration is introduced. It uses fuzzy similarity relations combined by a transitive consensus matrix. The methodology is adopted for individuating a subset of models that best characterize the predicted atmospheric pollutants from the ETEX-1 experiment and discard redundant information.

Keywords: Fuzzy similarity · Hierarchical agglomeration · Ensemble models · Air pollutant dispersion

1 Introduction

The real-time simulation of pollutant dispersion or the accidental release of radioactive substances in the atmosphere is a challenging aspect of many national services and agencies. In particular, releases of harmful radionuclides (e.g. Fukushima, Chernobyl) could be simulated and monitored [1,10,13,20]. In this work we consider atmospheric compounds from the *ENSEMBLE* system [6–8]. ENSEMBLE is a web-based system aiming at assisting the analysis of multimodel data provided by many national meteorological services and environmental protection agencies worldwide. It is worth noting that in the case of multimodel ensemble for atmospheric dispersions, models are certainly more or less dependent from several intrinsic mechanisms (e.g., they often share features, initial/boundary data, numerical methods, parameterizations and emissions). For this reason, results obtained by ensemble analysis may lead to erroneous interpretations and in a multimodel approach the effective number of models may

R. Fullér et al. (Eds.): WILF 2018, LNAI 11291, pp. 123–133, 2019.
https://doi.org/10.1007/978-3-030-12544-8_10

be lower than the total number, since models could be linearly (or nonlinearly) dependent on each other.

To solve this problem, a number of techniques has been proposed in literature. In [15,17,18] the authors present a statistical analysis (i.e., *Bayesian Model Averaging*) for combining predictive distributions from different sources of a multi-model ensemble, and in [16] some basic properties of multi-model ensemble systems are investigated. Moreover, cluster-based approaches have also been proposed [2–4]. In this paper, we introduce a methodology that improves the forecasting by considering observations that may become available during the course of the event. The methodology is based on fuzzy similarity relations that allow to combine multiple hierarchical agglomerations, each for a different forecasting leading time. From the overall temporal agglomeration obtained by a consensus matrix it is possible to select a subset of models and discard redundant information.

The remainder of the paper is organized as follows. In Sect. 2 the proposed methodology is detailed. In particular, some fundamental concepts on t-norms and fuzzy similarity relations (Sect. 2.2) are given and the agglomerative based approach is described in Sect. 2.3. Finally, in Sect. 3 some experimental results, obtained by applying this methodology on an ensemble of prediction models, are described. Conclusions and future remarks are given in Sect. 4.

2 Fuzzy Similarity and Agglomerative Clustering

In general, when one deals with clustering tasks, *fuzzy logic* permits to obtain soft clustering, instead of hard (crisp or non-fuzzy) clustering of data. Hierarchical clustering is a methodology for cluster analysis which seeks to build a hierarchy of clusters and it can be agglomerative or divisive. In this work we consider an agglomerative clustering approach. One of the main aspects of this methodology is the use of a measure of dissimilarity between sets of observations, by using an appropriate metric. On the other hand, a dendrogram is a tree diagram used to illustrate the results produced by hierarchical clustering. In the following, we show that a dendrogram can be associated with a fuzzy equivalence relation based on Łukasiewicz valued fuzzy similarities. Successively, a consensus matrix, that is the representative information of all dendrograms, is obtained by combining multiple temporal hierarchical agglomerations of dispersion models. The main steps of the proposed approach are

1. Membership functions characterization;
2. Fuzzy Similarity Matrix calculation (or dendrogram) for all the models at a fixed time;
3. Consensus matrix construction for temporal hierarchical agglomerations.

2.1 Membership Functions

The effective of *fuzzy logic* is the transformation of linguistic variables in fuzzy sets. Fuzzification is the process of changing a real scalar value into a fuzzy value and it is achieved by using different types of membership functions. The membership function represents the degree of truth to which a given input belongs to a fuzzy set. In the proposed approach, *fuzzy sets* are described by the following *membership functions* [21]

$$\mu(\mathbf{x}_i) = \frac{\mathbf{x}_i - \min(\mathbf{x}_i)}{\max(\mathbf{x}_i) - \min(\mathbf{x}_i)}, \tag{1}$$

where $\mathbf{x}_i = [x_1^i, x_2^i, \ldots, x_L^i]$ is the i-th observation vector of the L considered models.

2.2 Fuzzy Similarity

We observe that fuzzy sets can be combined via the conjunction and disjunction operations and continuous triangle norms or co-norms are adopted, respectively. A *triangular norm* (*t-norm* for short), is a binary operation t on the unit interval $[0, 1]$. In particular, it is a function $t : [0, 1]^2 \to [0, 1]$, such that it satisfies the following four axioms for all $x, y, z \in [0, 1]$ [11]

$$\begin{aligned}
t(x,y) &= t(y,x) && (\textit{commutativity})\\
t(x,t(y,z)) &= t(t(x,y),z) && (\textit{associativity})\\
t(x,y) &\leq t(x,z) \quad \text{whenever } y \leq z && (\textit{monotonicity})\\
t(x,1) &= x && (\textit{boundary condition})
\end{aligned} \tag{2}$$

In practical situations the following four basic t-norms are considered

$$\begin{aligned}
t_{\mathbf{M}}(x,y) &= \min(x,y) && (\textit{minimum})\\
t_{\mathbf{P}}(x,y) &= x \cdot y && (\textit{product})\\
t_{\mathbf{L}}(x,y) &= \max(x+y-1,0) && (\textit{Łukasiewicz t-norm})\\
t_{\mathbf{D}}(x,y) &= \begin{cases} 0 & \text{if } (x,y) \in [0,1]^2 \\ \min(x,y) & \text{otherwise} \end{cases} && (\textit{drastic product})
\end{aligned} \tag{3}$$

However, in these years, several parametric and non-parametric t-norms have been introduced [11] and generalized versions have also been studied [5]. In the following, we focus on the properties of the Łukasiewicz t-norm ($t_{\mathbf{L}}$). One main operator adopted in fuzzy-based systems (e.g., fuzzy inference systems) is the *residuum* $\to_t$

$$x \to_t y = \bigvee \{z | t(z,x) \leq y\} \tag{4}$$

where $\bigvee$ is the *union* operator and, for the left-continuous basic t-norm $t_{\mathbf{L}}$, is given by

$$x \to_{\mathbf{L}} y = \min(1 - x + y, 1) \, (\textit{Łukasiewicz implication}) \tag{5}$$

Moreover, we also note that letting p be a fixed natural number in a *generalized Łukasiewicz structure*, we obtain

$$\begin{aligned} t_{\mathbf{L}}(x,y) &= \sqrt[p]{\max(x^p+y^p-1,0)} \\ x \rightarrow_{\mathbf{L}} y &= \min(\sqrt[p]{1-x^p+y^p},1) \end{aligned} \tag{6}$$

Another fundamental operation on a residuated lattice is the *bi-residuum* that will be used for our construction of the fuzzy similarities. It is defined as

$$x \leftrightarrow_t y = (x \rightarrow_t y) \wedge (y \rightarrow_t x), \tag{7}$$

where $\wedge$ is the *meet*. In the case of the left-continuous basic t-norm $t_{\mathbf{L}}$, we obtain the following *bi-residuum*

$$x \leftrightarrow_{\mathbf{L}} y = 1 - \max(x,y) + \min(x,y) \tag{8}$$

On the other hand, a binary *fuzzy relation* R is defined on $U \times V$ as a fuzzy set on $U \times V$ $(R \subseteq U \times V)$. A *similarity matrix* is a fuzzy relation $S \subseteq U \times U$ such that, for each $u, v, w \in U$, the following properties are satisfied

$$\begin{aligned} S\langle u,u\rangle &= 1 && (\textit{everthing is similar to itself}) \\ S\langle u,v\rangle &= S\langle v,u\rangle && (\textit{symmetric}) \\ t(S\langle u,v\rangle, S\langle v,w\rangle) &\leq S\langle u,w\rangle && (\textit{weakly transitive}) \end{aligned} \tag{9}$$

It is essential to observe that from fuzzy sets with membership functions $\mu : X \rightarrow [0,1]$, a fuzzy similarity matrix S can be generated as

$$S\langle a,b\rangle = \mu(a) \leftrightarrow_t \mu(b) \tag{10}$$

for all $a, b \in X$.

Moreover, to build the fuzzy similarity matrix a main result is considered [19,21]

Proposition 1. *Consider n Łukasiewicz valued fuzzy similarities S_i, $i = 1, \ldots, n$ on a set X. Then*

$$S\langle x,y\rangle = \frac{1}{n}\sum_{i=1}^{n} S_i\langle x,y\rangle \tag{11}$$

is a Łukasiewicz valued fuzzy similarity on X.

In this work, we consider for Eq. 11

$$S_i\langle x,y\rangle = x \leftrightarrow_{\mathbf{L}} y. \tag{12}$$

Now, let $t_{\mathbf{L}}$ be the Łukasiewicz product, it is worth noting that S is a fuzzy equivalence relation on X with respect to $t_{\mathbf{L}}$ iif $1 - S$ is a *pseudo-metric* on X.

Algorithm 1. Min-transitive closure

```
1: Input $R$ the input relation
2: Output $R^T$ the output transitive relation
3: Elaborate
   1. Calculate $R^* = R \cup (R \circ R)$
   2. if $R^* \neq R$ replace $R$ with $R^*$ and go to step 1
   else $R^T = R^*$ and the algorithm terminates.
```

2.3 Dendrogram and Consensus Matrix

We also have to observe that if a similarity relation is *min-transitive* ($t = \min$ in (9)) then it is a *fuzzy-equivalence relation* that can be graphically described by a *dendrogram* [12]. In other words, transitivity implies the existence of the dendrogram.

The min-transitive closure R^T of R can be obtained as follows [14]

$$R^T = \bigcup_{i=1}^{n-1} R^i \tag{13}$$

where R^{i+1} is defined as

$$R^{i+1} = R^i \circ R, \tag{14}$$

and n is the dimension of a relation matrix.

Considering two fuzzy relations R and S, we observe that the composition $R \circ S$ is a fuzzy relation defined by

$$R \circ S\langle x, y\rangle = \mathrm{Sup}_{z \in X}\{R\langle x, z\rangle \odot S\langle z, y\rangle\} \tag{15}$$

$\forall x, y \in X$, where $\odot$ stands for a t-norm (e.g., min operator) [14]. Then we can conclude that the min-transitive closure R^T of a matrix R can be easily computed and the overall process is described in Algorithm 1.

We also observe that to accomplish an agglomerative clustering a dissimilarity relation is needed. Here we considered the following result [14].

Lemma 1. *Letting R be a similarity relation with the elements $R\langle x, y\rangle \in [0, 1]$ and letting D be a dissimilarity relation, which is obtained from R by*

$$D(x, y) = 1 - R\langle x, y\rangle \tag{16}$$

then D is ultrametric iif R is min-transitive.

In other words, we have a one-to-one correspondence between min-transitive similarity matrices and dendrogram and between ultrametric dissimilarity matrices and dendrograms.

Finally, after the dendrograms have been obtained at each time, a consensus matrix, that is the representative information of all temporal dendrograms, is obtained by combining the transitive closures by using Eq. 15 (i.e., max-min) [14]. The overall approach is described in Algorithm 2.

Algorithm 2. Combination of dendrograms

```
1: Input $S^{(i)}$, $1 \leq i \leq L$ $L$ input similarity matrices (dendrograms)
2: Output $S$ the resulted similarity matrix (dendrogram)
   1. Aggregate the similarity matrices to a final similarity matrix $S = Aggregate(S^{(1)}, S^{(2)}, \ldots, S^{(L)})$
   a. Let $S^*$ be the identity matrix
   b. For each $S^{(i)}$ calculate e $S^* = S^* \cup (S^* \circ S^{(i)})$
   c. If $S^*$ is not changed $S = S^*$ and goto step 3 else goto step 1.b
3: Create the final dendrogram from $S$
```

3 Experimental Results

This Section aims to illustrate some results obtained by the proposed approach. In particular, we consider the multi-model ensemble simulated distributions of the ETEX-1 experiment [9]. The ETEX-1 experiment concerned the release of pseudo-radioactive material on 23 October 1994 at 16:00 UTC from Monterfil, southeast of Rennes (France). Briefly, a steady westerly flow of unstable air masses was present over central Europe. Such conditions persisted for the 90 h that followed the release with frequent precipitation events over the advection area and a slow movement toward the North Sea region. Just for an example, in Fig. 1 we show the integrated concentration after 78 h from release. In the experiment, the main objective of the several independent groups worldwide (25 members) was to forecast the observations with different atmospheric dispersion models. Moreover, each simulation was based on weather fields generated by (most of the time) different *Global Circulation Models* (GCM) and all the simulations relate to the same release conditions. For further information on the involved groups and the adopted models the reader can refer to [8] and [9].

Now we apply the proposed approach to analyze data of the ETEX-1 experiment. The preliminary step is the *fuzzification*. In particular, Eq. 1 is applied on the concentrations estimated by models at each time level. Successively, for each concentration at different times a dendrogram (similarity matrix) is produced (Eq. 11 with Łukasiwicz norm and $p = 1$). Finally, the consensus matrix that described the representative dendrogram is estimated by using the approach described in Algorithm 2. In Fig. 2 a particular of the representative dendrogram obtained after 78 h is visualized. We observe that different clusters of similar models are obtained.

To highlight the clustering outcomes, in Fig. 3, we show some representative distributions of the clustered models. For example, as confirmed by dendrogram, the distributions of the models 22 and 24 are very close. See Figs. 3a and b for a comparison. Instead, the model 21 has a very diffusive distribution, as highlighted by the dendrogram. This distribution is visualized in Fig. 3c. At this point, we can identify models that have similar behavior by analyzing the different clusters. In order to identify the group of models that more appropriately describe observations, we compare the distributions of the models by using a Kullback Leibler divergence.

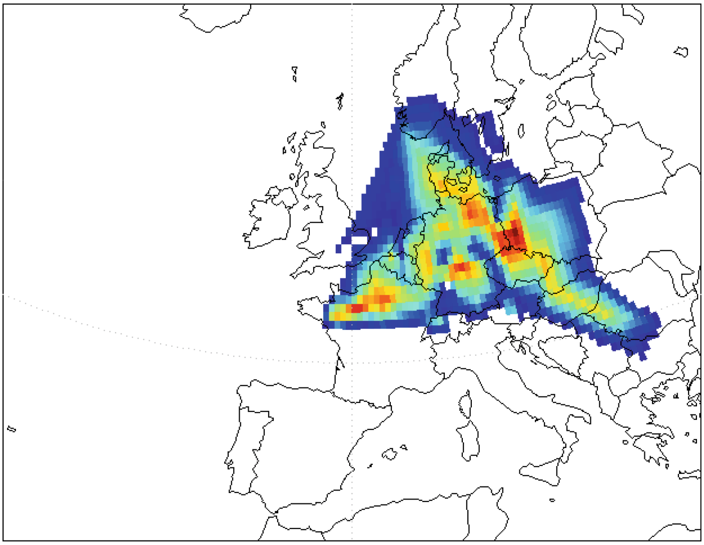


Fig. 1. ETEX-1 temporal integrated observations after 78 h.

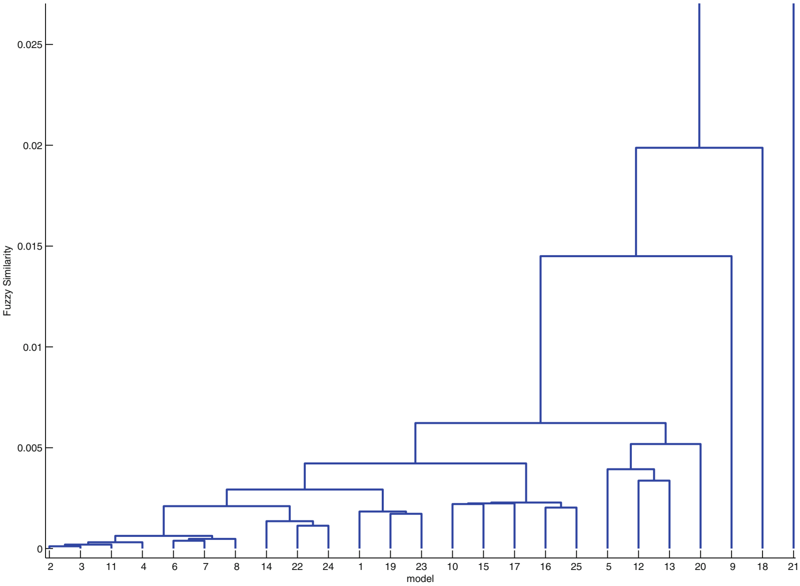
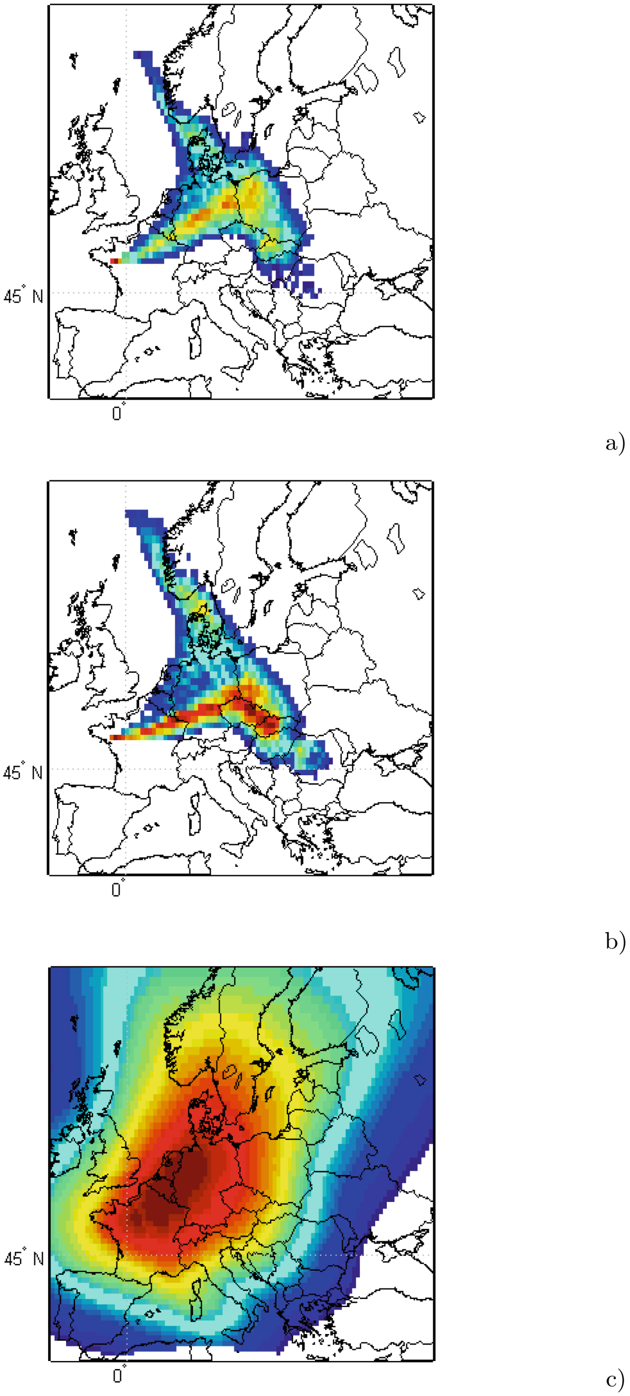


Fig. 2. Representative dendrogram obtained by consensus matrix: x-axis are related to the models and those on the y-axis are related to the model data similarities.

The Kullback Leibler (KL) divergence between two discrete n-dimensional probability density functions $\mathbf{p} = [p_i \dots p_n]$ and $\mathbf{q} = [q, \dots q_n]$ is defined as

$$KL(\mathbf{p}||\mathbf{q}) = \sum_{i=1}^{n} p_i \log \left(\frac{p_i}{q_i} \right) . \tag{17}$$



a)

b)

c)

Fig. 3. Model distributions: (a) model 22; (b) model 24; model 21.

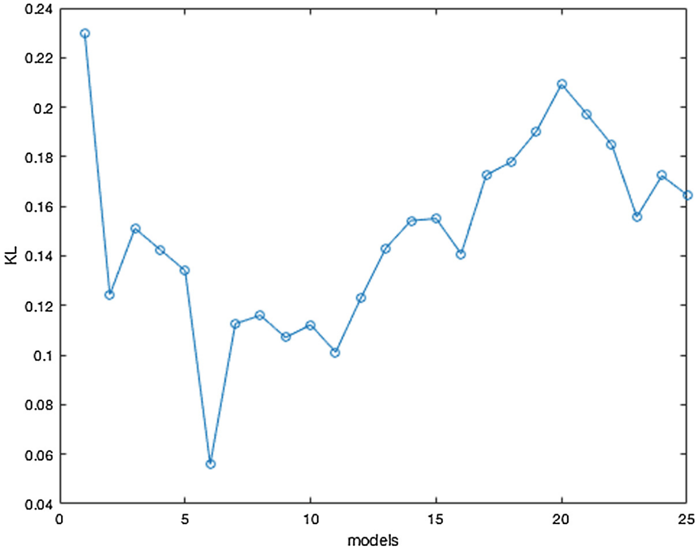


Fig. 4. KL divergence varying the clustering number.

This is known as the relative entropy. It satisfies the Gibbs' inequality

$$KL(\mathbf{p}||\mathbf{q}) \geq 0 \tag{18}$$

where equality holds only if $\mathbf{p} \equiv \mathbf{q}$. In general $KL(\mathbf{p}||\mathbf{q}) \neq KL(\mathbf{q}||\mathbf{p})$. In our experiments we use the symmetric version [2] that can be defined as

$$KL = \frac{KL(\mathbf{p}||\mathbf{q}) + KL(\mathbf{q}||\mathbf{p})}{2}. \tag{19}$$

First of all, we compute the KL divergence between each model and the median value of the overall cluster. Successively, for each cluster, the model with the minimum KL is selected. The *median model* of these considered models is compared with the real observations by KL. In Fig. 4 we show the KL obtained by varying the number of clusters.

We observe that varying the number of clusters this procedure permits to select the models that have the best approximation of the real observation (see [17] and [4] for more details). After our analysis, we conclude that the best approximation is obtained by using 6 clusters. Moreover, we stress that a lower KL does not necessarily correspond to the use of a large number of models. This suggest an approach for systematic reduction of ensemble data complexity and the use of the consensus matrix permits to obtain a more robust and realistic temporal analysis.

4 Conclusions

In this work we focused on models comparison in a multi-model air quality ensemble system. A methodology based on temporal hierarchical agglomeration is introduced for real-time simulation of pollutant dispersion or the accidental release of radioactive nuclides in the atmosphere. The proposed methodology is able to combine multiple temporal hierarchical agglomerations of dispersion models and it is based on fuzzy similarity relations combined by a transitive consensus matrix. The methodology is adopted for individuating models that characterize the predicted atmospheric pollutants from the ETEX-1 experiment. The results show that this methodology is able to discard redundant temporal information, reducing the data complexity. In the next future, further experimentations will be devoted to real pollutant dispersions (e.g., Fukushima) and different similarity relations also using ordinal sums.

Acknowledgments. This work was partially funded by the University of Naples Parthenope (*Sostegno alla ricerca individuale per il triennio 2016–2018* project).

References

1. Ascione, I., Giunta, G., Mariani, P., Montella, R., Riccio, A.: A grid computing based virtual laboratory for environmental simulations. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1085–1094. Springer, Heidelberg (2006). https://doi.org/10.1007/11823285_114
2. Ciaramella, A., et al.: Interactive data analysis and clustering of genomic data. Neural Netw. **21**(2–3), 368–378 (2008)
3. Napolitano, F., Raiconi, G., Tagliaferri, R., Ciaramella, A., Staiano, A., Miele, G.: Clustering and visualization approaches for human cell cycle gene expression data analysis. Int. J. Approximate Reasoning **47**(1), 70–84 (2008)
4. Ciaramella, A., Giunta, G., Riccio, A., Galmarini, S.: Independent data model selection for ensemble dispersion forecasting. In: Okun, O., Valentini, G. (eds.) Applications of Supervised and Unsupervised Ensemble Methods. SCI, vol. 245, pp. 213–231. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-03999-7_12
5. Ciaramella, A., Pedrycz, W., Tagliaferri, R.: The genetic development of ordinal sums. Fuzzy Sets Syst. **151**, 303–325 (2005)
6. Galmarini, S., Bianconi, R., Bellasio, R., Graziani, G.: Forecasting consequences of accidental releases from ensemble dispersion modelling. J. Environ. Radioactiv. **57**, 203–219 (2001)
7. Galmarini, S., et al.: Ensemble dispersion forecasting, part I: concept, approach and indicators. Atmos. Environ. **38**, 4607–4617 (2004)
8. Galmarini, S., et al.: Ensemble dispersion forecasting? Part II: application and evaluation. Atmos. Environ. **38**, 4619–4632 (2004)
9. Girardi, F., et al.: The ETEX project. EUR Report 181–43 EN, 108 pp. Office for official publications of the European Communities, Luxembourg (1998)
10. Giunta, G., Montella, R., Mariani, P., Riccio, A.: Modeling and computational issues for air/water quality problems: a grid computing approach. Nuovo Cimento C Geophys. Space Phys. **28**, 215–224 (2005)

11. Klement, E.P., Mesiar, R., Pap, E.: Triangular Norms. Kluwer Academic Publishers, Dordrecht (2001)
12. Meyer, H.D., Naessens, H., Baets, B.D.: Algorithms for computing the min-transitive closure and associated partition tree of a symmetric fuzzy relation. Eur. J. Oper. Res. **155**(1), 226–238 (2004)
13. Montella, R., Giunta, G., Riccio, A.: Using grid computing based components in on demand environmental data delivery. In: Proceedings of the Second Workshop on Use of P2P, GRID and Agents for the Development of Content Networks, UPGRADE-CN 2007, pp. 81–86 (2007)
14. Mirzaei, A., Rahmati, M.: A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations. IEEE Trans. Fuzzy Syst. **18**(1), 27–39 (2010)
15. Potempski, S., Galmarini, S., Riccio, A., 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** (2010). https://doi.org/10.1029/2010JD014210
16. Potempski, S., Galmarini, S.: Est modus in rebus: analytical properties of multi-model ensembles. Atmos. Chem. Phys. **9**(24), 9471–9489 (2009)
17. Riccio, A., Giunta, G., 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 (2007)
18. Riccio, A., Ciaramella, A., Giunta, G., Galmarini, S., Solazzo, E., Potempski, S.: On the systematic reduction of data complexity in multimodel atmospheric dispersion ensemble modeling. J. Geophys. Res. **117**(D5), D05314 (2012)
19. Sessa, S., Tagliaferri, R., Longo, G., Ciaramella, A., Staiano, A.: Fuzzy similarities in stars/galaxies classification. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 494–4962 (2003)
20. Solazzo, E., Riccio, A., Van Dingenen, R., Valentini, L., Galmarini, S.: Evaluation and uncertainty estimation of the impact of air quality modelling on crop yields and premature deaths using a multi-model ensemble. Sci. Total Environ. **633**, 1437–1452 (2018)
21. Turunen, E.: Mathematics Behind Fuzzy Logic. Advances in Soft Computing. Springer, Heidelberg (1999)