

Sensitivity Study of a Large-scale Air Pollution Model on the Bulgarian Petascale Supercomputer Discoverer

Tzvetan Ostromsky, Ivan Dimov, Rayna Georgieva
Institute of Information and
Communication Technologies,
Bulgarian Academy of Sciences (ICT-BAS),
Acad. G. Bonchev Str., Block 25-A,
1113 Sofia, Bulgaria
Email: ceco@parallel.bas.bg,
ivdimov@bas.bg, rayna@parallel.bas.bg

Venelin Todorov
Department of Information Modelling,
Institute of Mathematics and Informatics,
Bulgarian Academy of Sciences (IMI-BAS),
Acad. G. Bonchev Str., Block 8,
1113 Sofia, Bulgaria
Email: vtodorov@math.bas.bg

Abstract—The focus of this study is on the optimal use of high performance computing in the area of environmental security (air pollution transport, in particular). Contemporary mathematical models of air pollution transport should include a fairly large set of chemical and photochemical reactions to be established as a reliable simulation tool. The investigations and the numerical results reported in this paper have been obtained by using a large-scale mathematical model called the Danish Eulerian Model (DEM).

For optimization of some applications of the Danish Eulerian Model in various important scientific, social and economic areas, it is of great importance to simplify the model as much as possible, preserving the high reliability of its output results. A careful sensitivity analysis is needed in order to decide how to do such simplifications. On the other hand, it is important to analyze the influence of variations of the initial conditions, the boundary conditions, the rates of some chemical reactions, etc. on the model results in order to make right assumptions about the possible simplifications, which could be done. The sensitivity analysis version of the Danish Eulerian Model was created for these purposes. Its complexity is of higher order, a real challenge for the top performance supercomputers nowadays. The sensitivity analysis version of DEM (SA-DEM) has been implemented on the new Bulgarian petascale supercomputer DISCOVERER. It is a part of the European High Performance Computing Joint Undertaking (EuroHPC), which is building a network of 8 powerful supercomputers across the European Union (3 pre-exascale and 5 petascale).

The results of some scalability experiments with SA-DEM on the new Bulgarian petascale supercomputer DISCOVERER are presented here. They are compared with similar experiments performed on the Mare Nostrum III supercomputer at Barcelona Supercomputing Centre – the most powerful supercomputer in Spain by that time, upgraded currently to the pre-exascale Mare Nostrum V, also part of the EuroHPC JU infrastructure.

Keywords: sensitivity analysis, air pollution, numerical model, supercomputer, parallel algorithm, scalability

I. INTRODUCTION

Environmental security is rapidly becoming a significant topic of present interest all over the world. It is necessary to carry out many comprehensive scientific studies and to

analyze carefully the most important physical and chemical processes during the transport, and transformations under the transport of air pollutants. An effective performance of such complicated procedures requires a joined research and collaboration between experts in the field of environmental modeling, numerical analysis and scientific computing.

The aim of the present work is to propose a new mechanism for investigation the sensitivity of the calculated concentration levels of important pollutants (like nitrogen dioxide NO_2 and especially ozone O_3) due to variation of rates of the involved chemical reactions in a real-life scenario of air pollution transport over Europe with the Unified Danish Eulerian Model (UNI-DEM).

In investigation of various highly complex engineering, physical, environmental, social, and economic systems it is important to measure relations that describe the effect on the output results when the conditions for the input change.

Sensitivity analysis (SA) is the study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input [20]. Two classes in sensitivity analysis are considered in the existing literature: local SA and global SA. Local SA studies how some small variations of inputs around a given value change the value of the output. Global SA takes into account all the variation range of the inputs, and apportions the output uncertainty to the uncertainty in the input factors.

In general, several sensitivity analysis techniques are available [20]. Most existing methods for providing SA rely heavily on special assumptions connected to the behavior of the model (such as linearity, monotonicity and additivity of the relationship between input factor and model output). Among quantitative methods, variance-based methods are the most often used [19]. The main idea of these methods is to evaluate how the variance of an input or a group of inputs contributes into the variance of model output.

Computational tasks arising in the treatment of large-scale

air pollution models are enormous. It is highly desirable to simplify as much as possible the model, keeping high level of reliability of models' results. Sensitivity analysis is rather helpful in order to decide where and how simplifications can be made. On the other hand, it is important to analyze the influence of variations of the initial conditions, the boundary conditions and/or the chemical rates on the model results in order to make right assumptions about the simplifications which have to be implemented. Such an analysis can give valuable information about the performance of reliable and reasonable simplifications or to identify parameters and mechanisms the accuracy of which should be improved, because the model results are very sensitive to variations of these parameters and mechanisms. Thus, the goal could be: (i) improving the model, (ii) increasing the reliability of the results, and (iii) identifying processes that must be studied more carefully.

The rest of the paper is organized as follows: In Section II the general concept of global sensitivity analysis is introduced in terms of ANOVA high-dimensional model representation. In Section III the Danish Eulerian Model is described, including its high-performance parallel code UNI-DEM and its special sensitivity analysis version SA-DEM. In Section IV numerical results from some scalability experiments with SA-DEM on two of the most powerful supercomputers in Europe are given. Finally, some conclusions are drawn.

II. SENSITIVITY ANALYSIS CONCEPT

A. Global Sensitivity Indices

When the sensitivity of the concentrations calculated by UNI-DEM (or any other deterministic mathematical model) is studied, it is convenient to introduce some stochastic variables and equations.

It is assumed that the mathematical model can be presented as a model function

$$u = f(\mathbf{x}), \quad \mathbf{x} = (x_1, x_2, \dots, x_d) \in U^d \equiv [0; 1]^d \quad (1)$$

is a vector of input parameters with a joint probability density function (p.d.f.) $p(\mathbf{x}) = p(x_1, \dots, x_d)$. In general, real problems are characterized by multiple outputs. Here it is assumed that a scalar output is given. It is also assumed that input variables are independent (non-correlated input variables) and the density function $p(\mathbf{x}) = p(x_1, x_2, \dots, x_d)$ is known, even if x_i are not actually random variables. This implies that the output u is also a random variable, as it is a function of the random vector \mathbf{x} , with its own p.d.f.

It is reasonable to introduce an indicator that measures the importance of the influence of a given input parameter onto the output. The main indicator referred to a given input parameter x_i , $i = 1, \dots, d$ (normalised between 0 and 1) is defined as

$$\frac{\mathbf{D}[\mathbf{E}[u|x_i]]}{\mathbf{D}_u}, \quad (2)$$

where $\mathbf{D}[\mathbf{E}[u|x_i]]$ is the variance of the conditional expectation of u with respect to x_i and \mathbf{D}_u is the total variance according to u . This indicator is named *first-order sensitivity index* by Sobol' [23] or *correlation ratio* by McKay [13]. A brief review

of measures of importance used in variance-based methods for sensitivity analysis is given in [3].

The *total sensitivity index* [10] provides a measure of the total effect of a given parameter, including all the possible joint terms between that parameter and all the others. The *total sensitivity index* (TSI) of input parameter x_i , $i \in \{1, \dots, d\}$ is defined in the following way [10], [23]:

$$S_{x_i}^{tot} = S_i + \sum_{l_1 \neq i} S_{il_1} + \sum_{l_1, l_2 \neq i, l_1 < l_2} S_{il_1 l_2} + \dots + S_{il_1 \dots l_{d-1}} \quad (3)$$

where S_i is called *the main effect (first-order sensitivity index)* of x_i and $S_{il_1 \dots l_{j-1}}$ is the j -th order sensitivity index (respectively *two-way interactions* for $j = 2$, *three-way interactions* for $j = 3$ and so on) for parameter x_i ($2 \leq j \leq d$). The higher-order terms describe the interaction effects between the unknown input parameters $x_{i_1}, \dots, x_{i_\nu}$, $\nu \in \{2, \dots, d\}$ on the output variance. Usually for practical computations the set of input parameters is classified according their TSI [3]: *very important* if $0.8 < S_{x_i}^{tot}$, *important* if $0.5 < S_{x_i}^{tot} < 0.8$, *unimportant* if $0.3 < S_{x_i}^{tot} < 0.5$, and *irrelevant* if $S_{x_i}^{tot} < 0.3$. In subsection II-B we will show how sensitivity indices $S_{i_1 \dots i_\nu}$ are defined via the variances of conditional expectations $\mathbf{D}_{l_1} = \mathbf{D}[f_{l_1}(x_{l_1})] = \mathbf{D}[\mathbf{E}(u|x_{l_1})]$, $\mathbf{D}_{l_1 \dots l_\nu}$, $2 \leq \nu \leq d$ (see, equation (8)). It is often reasonable to assume (see [12], [17]) that relatively small subsets of input variables in high-dimensional models have the main impact on the output. The high dimensional sums can be neglected when many practical problems are studied. This means that one can use low-order indices preferably, but should be able to control the contribution of higher-order terms.

B. The Sobol' Approach

The Sobol' method is one of the most often used variance-based methods. To our best knowledge the Sobol' sensitivity measure [23] was first published in [22]. An important advantage of this method is that it allows to compute not only the first-order indices, but also indices of a higher-order in a way similar to the computation of the main effects. The total sensitivity index can be calculated with just one Monte Carlo integral per factor.

The method for global SA applied here is based on a decomposition of an integrable model function f in the d -dimensional factor space into terms of increasing dimensionality:

$$f(\mathbf{x}) = f_0 + \sum_{\nu=1}^d \sum_{l_1 < \dots < l_\nu} f_{l_1 \dots l_\nu}(x_{l_1}, x_{l_2}, \dots, x_{l_\nu}) \quad (4)$$

where f_0 is a constant. The total number of summands in equation (4) is 2^d (see [25]) and, in general, this so called high dimensional model representation [23] is non-unique. But, if each term is chosen to satisfy the following condition

$$\int_0^1 f_{l_1 \dots l_\nu}(x_{l_1}, x_{l_2}, \dots, x_{l_\nu}) x_{l_k} = 0, \quad 1 \leq k \leq \nu \leq d \quad (5)$$

then (4) is unique. The representation (4) is called **ANOVA-representation** of the model function $f(\mathbf{x})$ [24]. Here and

hereafter the variables of integration are marked by a dot below in the integration formulae.

The functional decomposition of $[0; 1]^d$ ANOVA (meaning: **ANalysis Of Variance**) -representation has been studied by many authors [2], [9], [21], [26]. Sobol' has proven [22] that the decomposition (4) is unique on the assumption (5) and the functions of the right-hand side can be defined in a unique way by multidimensional integrals [24]:

- $f_0 = \int_{U^d} f(x) \bar{x}$;
- $f_{l_1}(x_{l_1}) = \int_{U^{d-1}} f(x) \prod_{k \neq l_1} \bar{x}_k - f_0, \quad l_1 \in \{1, 2, \dots, d\}$;
- $f_{l_1 l_2}(x_{l_1}, x_{l_2}) = \int_{U^{d-2}} f(x) \prod_{k \neq l_1, l_2} \bar{x}_k - f_0 - f_{l_1}(x_{l_1}) - f_{l_2}(x_{l_2}), \quad l_1, l_2 \in \{1, \dots, d\}$.

An additional essential property of the terms in the ANOVA-representation is their mutual orthogonality:

$$\int_{U^d} f_{i_1 \dots i_\mu}(x) f_{j_1 \dots j_\nu}(x) \bar{x} = 0, \\ (i_1, \dots, i_\mu) \neq (j_1, \dots, j_\nu), \quad \mu, \nu \in \{1, \dots, d\}.$$

It follows from the assumption that the above subsets of indices differ from one another at least one element and the corresponding integral vanishes for this index due to (5).

The quantities

$$\mathbf{D} = \int_{U^d} f^2(x) \bar{x} - f_0^2 \quad (6) \\ \mathbf{D}_{l_1 \dots l_\nu} = \int f_{l_1 \dots l_\nu}^2(x) \bar{x}_{l_1} \dots \bar{x}_{l_\nu}$$

are called variances (total and partial variances, respectively) and have been obtained after squaring and integrating over U^d the equality (4) on the assumption that $f(x)$ is a square integrable function (thus all terms in (4) are also square integrable functions). Therefore, the total variance of the model output is partitioned into partial variances [22] in the analogous way as the model function, that is the unique ANOVA-decomposition:

$$\mathbf{D} = \sum_{\nu=1}^d \sum_{l_1 < \dots < l_\nu} \mathbf{D}_{l_1 \dots l_\nu}. \quad (7)$$

It is obvious that the use of terms of probability theory is based on the following interpretation: in general, the input parameters are random variables distributed in U^d that defines $f_{l_1 \dots l_\nu}(x_{l_1}, x_{l_2}, \dots, x_{l_\nu})$ also as random variables with variances (6). For example f_{l_1} is presented by a conditional expectation:

$$f_{l_1}(x_{l_1}) = \mathbf{E}(u|x_{l_1}) - f_0$$

and respectively

$$\mathbf{D}_{l_1} = \mathbf{D}[f_{l_1}(x_{l_1})] = \mathbf{D}[\mathbf{E}(u|x_{l_1})].$$

Based on the above assumptions about the model function and the output variance, the following quantities

$$S_{l_1 \dots l_\nu} = \frac{\mathbf{D}_{l_1 \dots l_\nu}}{\mathbf{D}}, \quad \nu \in \{1, \dots, d\} \quad (8)$$

are called Sobol' global sensitivity indices [22], [24]. This formula coincides for $\nu = 1$ with (2) and the so defined measures correspond to the main effect of input parameters as well as the interactions effect. Using the definition of these measures as ratios of variances and dividing (7) by \mathbf{D} , it is easy to show that the following properties hold for the Sobol' global sensitivity indices: $S_{l_1 \dots l_\nu} \geq 0$, and

$$\sum_{\nu=1}^d \sum_{l_1 < \dots < l_\nu} S_{l_1 \dots l_\nu} = 1. \quad (9)$$

Based on the results discussed above it is clear that the mathematical treatment of the problem of providing global sensitivity analysis consists in evaluating total sensitivity indices (3) and in particular Sobol' global sensitivity indices (8) of corresponding order. And that leads to computing of multidimensional integrals: $I = \int_{\Omega} g(x) p(x) \bar{x}$, $\Omega \subset \mathbf{R}^d$, where $g(x)$ is a square integrable function in Ω and $p(x) \geq 0$ is a probability density function, such that $\int_{\Omega} p(x) \bar{x} = 1$. This means that in general case one needs to compute 2^d integrals of type (6) to obtain $S_{x_i}^{tot}$. As we discussed earlier the basic assumption underlying representation (4) is that the basic features of the model functions (1) describing typical real-life problems can be presented by low-order subsets of input variables [12], [17], that are constants, terms of first and second order. Thus, the high-dimensional sums (referred to higher-order interactions effects) in (4) can normally be neglected. Therefore, based on this assumption, one can assume that the dimension of the initial problem can be reduced.

Nevertheless, the calculating of the integrals defined by formulas (6) requires integration of different integrands that is not effective according to the computational cost. The procedure for computing global sensitivity indices measuring effect (main or otherwise) of the input parameters that is overcoming this disadvantage has been proposed by Sobol' [24]. Consider an arbitrary set of m variables ($1 \leq m \leq d-1$): $y = (x_{k_1}, \dots, x_{k_m})$, $1 \leq k_1 < \dots < k_m \leq d$, and let z be the set of $d-m$ complementary variables. Thus $x = (y, z)$. Let $K = (k_1, \dots, k_m)$.

The variances corresponding to the subsets y and z can be defined as

$$\mathbf{D}_y = \sum_{n=1}^m \sum_{(i_1 < \dots < i_n) \in K} \mathbf{D}_{i_1 \dots i_n}, \quad (10) \\ \mathbf{D}_z = \sum_{n=1}^{d-m} \sum_{(j_1 < \dots < j_n) \in \bar{K}} \mathbf{D}_{j_1 \dots j_n},$$

where the complement of the subset K in the set of all parameter indices is denoted by \bar{K} . The first sum in (10) is extended over all subsets (i_1, \dots, i_n) , where all indices i_1, \dots, i_n belong to K . Then the total variance corresponding to the subset y is $\mathbf{D}_y^{tot} = \mathbf{D} - \mathbf{D}_z$ and it is extended over all subsets (i_1, \dots, i_ν) , $1 \leq \nu \leq d$, where at least one $i_l \in K$, $1 \leq l \leq \nu$.

The procedure for computation of global sensitivity indices is based on the following representation of the variance $\mathbf{D}_y =$

$\int f(x) f(y, z') \bar{x} z' - f_0^2$ (see [24]). The last equality allows to construct a Monte Carlo algorithm for evaluating f_0 , \mathbf{D} and \mathbf{D}_y , where $\xi = (\eta, \zeta)$:

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^N f(\xi_j) &\xrightarrow{P} f_0, \\ \frac{1}{N} \sum_{j=1}^N f(\xi_j) f(\eta_j, \zeta_j) &\xrightarrow{P} \mathbf{D}_y + f_0^2, \\ \frac{1}{N} \sum_{j=1}^N f^2(\xi_j) &\xrightarrow{P} \mathbf{D} + f_0^2, \\ \frac{1}{N} \sum_{j=1}^N f(\xi_j) f(\eta'_j, \zeta_j) &\xrightarrow{P} \mathbf{D}_z + f_0^2. \end{aligned}$$

For example, for $m = 1, y = \{x_{l_1}\}, l_1 \in \{1, \dots, d\}$ and $z = \{1, \dots, d\} \setminus l_1$:

$$S_{l_1} = S_{(l_1)} = \mathbf{D}_{(l_1)} / \mathbf{D}, \quad S_{l_1}^{tot} = \mathbf{D}_{l_1}^{tot} / \mathbf{D} = 1 - S_z.$$

It is important to estimate the computational cost for computing the sensitivity indices in order to be able to compare this approach with other existing approaches. The computational cost of estimating all first-order ($m = 1$) and total sensitivity indices via the scheme proposed by Sobol' can be defined as $N(2d + 1)$ model function evaluations (N model runs for f_0 , dN model runs for the first-order terms, and dN model runs for the total effect terms), where N is the sample size and d is the number of input parameters. It should be noted that the most frequently used variance-based methods as Sobol' method and FAST (Fourier Amplitude Sensitivity Test) (and their improved versions) have a computational cost proportional to dN of estimating all main and total effects of input parameters (see [18]).

The computing of higher-order interactions effect can be performed by an iterative process. For example,

$$S_{(l_1 l_2)} = \mathbf{D}_{(l_1 l_2)} / \mathbf{D} = S_{l_1} + S_{l_2} + S_{l_1 l_2},$$

and $S_{l_1 l_2}$ can be obtained assuming that the corresponding first-order sensitivity indices have already been computed.

III. DESCRIPTION AND PARALLEL IMPLEMENTATIONS OF THE DANISH EULERIAN MODEL (DEM)

DEM is a powerful large scale air pollution model, with more than 30-year development history [28], [15], [16], [29]. Over the years it was successfully applied in different long-term environmental studies in various areas. processes in the atmosphere should be taken into account, which are mathematically represented by a complex PDE system. To simplify it a proper splitting procedure is applied. As a result the initial system is replaced by several simpler systems (submodels), connected with the main physical and chemical processes. These systems should be calculated in a large spatial domain, as the pollutants migrate quickly on long distances, driven by the atmosphere dynamics, especially on high altitude. Here they are exposed to temperature, light and other condition changes in extremely wide range, so does the speed of most chemical reactions. One of the major sources of difficulty is the

dynamics of the atmospheric processes, which require small time-step to be used (at least, for the chemistry submodel) in order to get a stable numerical solution of the corresponding system. All this makes the treatment of large-scale air pollution models a tuff and heavy computational task. It has always been a serious challenge, even for the fastest and most powerful state-of-the-art supercomputers. [7], [29].

The Danish Eulerian Model (DEM) [27], [28] is mathematically represented by the following system of partial differential equations:

$$\begin{aligned} \frac{\partial c_s}{\partial t} = & -\frac{\partial(uc_s)}{\partial x} - \frac{\partial(vc_s)}{\partial y} - \frac{\partial(wc_s)}{\partial z} + \\ & + \frac{\partial}{\partial x} \left(K_x \frac{\partial c_s}{\partial x} \right) + \frac{\partial}{\partial y} \left(K_y \frac{\partial c_s}{\partial y} \right) + \\ & + \frac{\partial}{\partial z} \left(K_z \frac{\partial c_s}{\partial z} \right) + \\ & + E_s + Q_s(c_1, c_2, \dots, c_q) - (k_{1s} + k_{2s})c_s; \end{aligned} \quad (11)$$

$s = 1, 2, \dots, q;$

where the following notation is used:

- q - number of equations = number of chemical species,
- c_s - concentrations of the chemical species considered,
- u, v, w - components of the wind along the coordinate axes,
- K_x, K_y, K_z - diffusion coefficients,
- E_s - emissions in the space domain,
- k_{1s}, k_{2s} - coefficients of dry and wet deposition respectively ($s = 1, \dots, q$),
- $Q_s(c_1, c_2, \dots, c_q)$ - non-linear functions that describe the chemical reactions between the species.

A. Splitting into submodels and domain decomposition

The above rather complex system is split into three sub-systems (submodels), according to the major physical and chemical processes as well as the numerical methods applied in their solution (marked by different colors in the right-hand-side of the system). These are (i) the horizontal advection and diffusion, (ii) chemistry, emissions and deposition and (iii) vertical exchange submodels, respectively. The discretization of the spatial derivatives in the right-hand-sides of these submodels results in forming three large systems of ordinary differential equations.

Chemical reactions play a significant role in the model. Moreover, both non-linearity and stiffness of the equations are mainly introduced by the chemistry (see [30]). On the other hand, this is one of the models of atmospheric chemistry, where the chemical processes are described with great detail in a very accurate way. The chemical scheme used in the model is the well-known condensed CBM-IV (Carbon Bond Mechanism; the scheme was proposed in [8], but some enhancements have been obtained in [28] by adding several

reactions for handling the ammonia-ammonium transformations in the atmosphere). It includes 35 pollutants and 116 chemical reactions, where 69 are time dependent and the rest 47 are time-independent. The scheme is suitable and adequate to study cases of high concentrations of chemical species.

Another crucial point on the way towards efficient numerical solution of the sub-models is the domain decomposition technique. This is a natural way to achieve distributed memory parallelization of any numerical problem over a large spatial domain. In some cases, however, like the advection-diffusion equations in particular, there is always certain overhead due to the boundary conditions treatment. Minimizing this overhead is a key point towards efficient optimization. On the other hand, optimization should not restrict the portability of the parallel implementation, as the intensive development in the computer technology inevitably leads to regular updates or complete replacement of the outdated hardware. Standard parallel programming tools as MPI and OpenMP (for distributed / shared memory models) are used in order to preserve portability of the code. An important issue towards efficient parallel optimization is also the load-balance. Sometimes the MPI barriers, used to force synchronization between the processes in data transfer commands, do not allow good load-balance. This obstacle can be avoided to some extent by using non-blocking communication routines from the MPI standard library.

More details about the numerical methods, applied to solve these systems, can be found in [1], [11], [28].

B. Parallelization strategy

The MPI standard library is used as a main parallelization tool. The MPI (Message Passing Interface) was initially developed as a standard communication library for distributed memory computers. Later, proving to be efficient, portable and easy to use, it became one of the most popular parallelization tools for application programming. Now it can be used on much wider class of parallel systems, including shared-memory computers and clustered systems (each node of the cluster being a separate shared-memory machine). Thus it provides high level of portability of the code.

In the case of DEM, MPI parallelization is based on the space domain partitioning [15], [16]. The space domain is divided into sub-domains (the number of the sub-domains is equal to the number of MPI tasks). Each MPI task works on its own sub-domain. On each time step there is no data dependency between the MPI tasks on both the chemistry and the vertical exchange stages. This is not so with the advection-diffusion stage. Spatial grid partitioning between the MPI tasks requires overlapping of the inner boundaries and exchange of certain boundary values on the neighboring subgrids for proper treatment of the boundary conditions. The subdomains are usually too large to fit into the fastest cache memory of the corresponding CPU. In order to achieve good data locality, the smaller calculation tasks are grouped in chunks (if appropriate) for more efficient cache utilization. An input parameter CHUNKSIZE is provided, which controls

the amount of short-term reusable data in order to reduce the transfer between the cache and the main (slower access) memory. It should be tuned with respect to the cache size of the target machine.

More detailed description of the main computational stages of DEM and the parallelization techniques used in each of them can be found in [1], [4], [15], [16], [28], [29], [30].

IV. NUMERICAL EXPERIMENTS WITH SA-DEM ON THE PETASCALE SUPERCOMPUTER IBM MARENOSTRUM III IN BARCELONA, SPAIN AND DISCOVERER EUROHPC IN SOFIA, BULGARIA

Results of scalability experiments with the 2-D fine-resolution grid version of SA-DEM on two of the most powerful supercomputers in Europe are shown in Tables I and II in this section. Some values of the user-defined parameters of SA-DEM used in the experiments on both machines are as follows:

- Grid-version: $(480 \times 480 \times 1)$;
- Time period of modelling: 1 year;
- Time step: 90 sec. (both in advection and chemistry stages);
- Cache utilization parameter: $NSIZE = 32$.

A. Numerical experiments on the IBM MareNostrum III supercomputer at BSC - Barcelona, Spain

Characteristics of the system IBM MareNostrum III

- 3028 nodes IBM dx360 M4, 16-core, 32 GB RAM per node;
- 48488 cores in total (Intel SandyBridge-EP E5-2670, 2600 MHz);
- Total RAM > 94 TB; Disk storage 1,9 PB;
- Interconnection networks: Infiniband / Gigabit Ethernet;
- Theoretical peak performance ~ 1 PFLOPS.

B. Numerical experiments on the EuroHPC JU supercomputer DISCOVERER in Bulgaria

Below are described some of the most important technical characteristics of the DISCOVERER supercomputer, installed 2 years ago in Sofia Tech Park. by Atos company. The machine is part of a new network of 8 powerful supercomputers in the the European Union, build up and governed by the European High Performance Computing Joint Undertaking (EuroHPC JU).

System properties:

- System model type: ATOS BullSequana XH2000;
- The system consists of 12 racks, 376 blades, 1128 nodes (18 of them – Fat nodes), 2 login nodes (for public access to the system);
- There are 2256 processors and 144384 cores in total;
- Total RAM: 302592 GB (128 GB per node); total disk storage: ~ 12 PB;
- Interconnection: Dragonfly+ with 200 Gbps (IB HDR) bandwidth per link;
- Sustained max. performance: 4.518 PFLOPS (on Linpack standard benchmark tests);

TABLE I
TIME (T) IN SECONDS, SPEED-UP (Sp) AND PARALLEL EFFICIENCY (E) OF SA-DEM (FINEST GRID) ON THE SPANISH SUPERCOMPUTER IBM MARENOSTRUM III AT BSC, BARCELONA

Time and speed-up of SA-DEM (MPI + OpenMP) on IBM MareNostrum III (480 × 480 × 1) grid, 35 species, CHUNKSIZE=32										
# CPU /threads	# NODES	Advection			Chemistry			TOTAL		
		T [s]	(Sp)	E [%]	T [s]	(Sp)	E [%]	T [s]	(Sp)	E [%]
10	1	80606	(10)	100 %	73426	(10)	100 %	165006	(10)	100 %
40	3	18760	(43)	108 %	15938	(46)	115 %	38775	(43)	106 %
80	5	9837	(82)	103 %	8551	(86)	107 %	21728	(76)	95 %
160	10	5130	(157)	98 %	4332	(169)	106 %	12525	(132)	82 %
320	20	2870	(281)	88 %	2292	(320)	100 %	8097	(204)	64 %
640	40	1511	(534)	83 %	1192	(616)	96 %	5299	(311)	49 %
960	60	1206	(669)	70 %	790	(929)	97 %	4034	(409)	43 %
1600	100	869	(927)	58 %	486	(1510)	94 %	3269	(505)	32 %
2400 /2	150	728	(1107)	46 %	407	(1804)	75 %	2415	(683)	28 %
4800 /4	300	265	(3040)	63 %	156	(4712)	98 %	1482	(1113)	23 %
15360/16	960	105	(7698)	50 %	48	(15170)	99 %	509	(3239)	21 %

TABLE II
TIME (T) IN SECONDS, SPEED-UP (Sp) AND THE TOTAL EFFICIENCY (E) OF SA-DEM ON THE EUROHPC JU SUPERCOMPUTER DISCOVERER IN SOFIA, BULGARIA

Time (T) in seconds and speed-up (Sp) of SA-DEM on DISCOVERER (480 × 480 × 1) grid, 35 species, CHUNKSIZE=32								
NP (MPI)	# NODES	Advection		Chemistry		TOTAL		
		T [s]	(Sp)	T [s]	(Sp)	T [s]	(Sp)	E [%]
10	1	72142	(10.0)	64726	(10.0)	146335	(10)	100 %
20	2	36175	(19.9)	30027	(21.6)	71129	(21)	103 %
40	3	18297	(39.4)	14295	(45.3)	36619	(40)	100 %
80	5	9523	(75.8)	7839	(82.6)	20383	(72)	90 %
160	10	4781	(150.9)	3925	(164)	11769	(124)	78 %
320	20	2525	(285.7)	2037	(317)	6861	(213)	67 %
640	40	1332	(541.7)	1034	(626)	4852	(302)	47 %
960	60	1017	(709.7)	697	(929)	3472	(421)	44 %
1600	100	787	(916.7)	463	(1398)	2822	(519)	32 %

- Theoretical peak performance: 6 PFLOPS, ratio (max to peak): 0.753;
- TOP500 ranking: # 91 in the world, # 27 in EU by the time of instalation (June 2021).

Computing node design:

- CPU type: AMD EPYC 7H12 (code name Rome), 64-core, frequency 2.6 GHz, power consumption 280W;
- CPU sockets per node: 2, CPU Cores per node: 128;
- Main memory per node : 256GB (Each of the 18x Fat nodes has 1024GB Memory);
- Memory type and frequency : 16GB DDR4 RDIMM 3200MT/s DR, (The fat nodes are equipped with 64-GB DDR4 RDIMM 3200MT/s DR);
- Node DP TeraFlop/s peak performance: 5.325TFlops;
- Node sustained performance on Linpack tests: 3.940TFlops;
- DP ratio TeraFlop/s – peak vs Linpack: 0,74 ;
- Linpack node power consumption: 665.1 W per 256 GB compute node; 747.0 W per Fat compute node (Cooling subsystem power consumption excluded);
- Number and bandwidth of network interfaces : 1x 200Gbps HDR.

High performance network properties:

- Interconnection family: IB HDR;
- Interconnection bandwidth per link: 200 Gbps (IB HDR);
- Expected latency (worst case for a 1 kB message): 520ns;
- Interconnection topology: Dragonfly+ ;
- Number of compute nodes per isle (2 Racks): 192;
- Blocking factor within isle : 2:1;
- Number of links to I/O partition: 120;
- Performance: 40 X HDR 200Gb/s ports in a 1U switch, 80 X HDR100 100Gb/s ports in a 1U switch;
- Aggregate switch throughput: 16 Tb/s;
- Up to 15.8 billion messages-per-second;
- Switch latency: 130 ns.

Management network properties:

- Network family: Ethernet;
- Network bandwidth : 10GbE/1GbE.

Storage system and I/O capacity:

- Total net capacity of data: 2031.89 TB;
- Total net capacity for metadata storage + home/apps – 15.25 TB; User home folders and application binaries will be in Data drive; The useable capacity of the filesystem will be 1 to 2
- Aggregated performance: 20 GB/s;
- Number of data modules: 164 HDD;

- Number of metadata modules: 11 SSD;
- Data module details: Net capacity provided (PB): 2.03 PB, Performance provided: 20 GB/s, Number and type of storage elements: 164 HDD + 11 SSD, Size per storage element: 6 TB + 1.92 TB, CPU Cores per server: 10, Main memory per server: 150 GB, Memory type and frequency: DDR4 2666 MT/s, Number and bandwidth interfaces to control data network: 4 x GigE RJ45 for OS access and hardware management, Number and bandwidth interfaces to bulk data network (RDMA): 4x HDR100 IB / 100GbE ports (same ports as for metadata).

V. CONCLUSIONS

Sensitivity analysis and particularly the results, reported in this work, have an important twofold role: for mathematical models verification and/or improvement, and/or on the other hand, for a reliable interpretation of experts of main effect, interaction and higher-order interaction effect of input parameters on model output. Variance-based analysis is an useful tool for an advanced investigation of relationships between model parameters, output results and internal mechanisms regulating the system under consideration. Specifying the most important chemical reactions for the model output the specialists from various applied fields (chemistry, physics) may obtain valuable information for an improvement of the model and thus it will lead to an increase of reliability and robustness of predictions.

The results of numerical experiments performed show that:

- The parallel MPI implementation of SA-DEM is well balanced, portable and runs efficiently on some of the most powerful supercomputers in Europe, including the Bulgarian Petascale supercomputer Discoverer, part of the EuroHPC JU network.
- The efficiency and speed-up is higher in the computationally-intensive stages. In particular, the chemistry stage (which does not need any communication between the tasks) has almost linear overall speed-up. The advection stage scales pretty well too, taking into account that there is some unavoidable computational overhead due to overlapping boundaries of the partitioning.
- The time for the computationally-intensive stages is additionally reduced in relation with the number of threads in the hybrid MPI-OpenMP code with the OpenMP lower level of parallelism switched on, which can be exploited on core level within a node.
- Further attention should be payed on the optimization of the I/O processes in order to reduce the slowdown of the execution on large number of nodes.

ACKNOWLEDGMENTS

This research was supported in part by the Bulgarian NSF project "Efficient methods for modelling, optimization and decision making" (contract # KP-06-N52/5), and by the European High-Performance Computing Joint Undertaking via the Discoverer Fast Track & Benchmark Access project "Installing, benchmarking, performance and scalability analysis of the UNI-DEM package for air pollution modelling".

REFERENCES

- [1] V. Alexandrov, A. Sameh, Y. Siddique and Z. Zlatev, Numerical integration of chemical ODE problems arising in air pollution models, *Env. Modeling and Assessment*, 2 (1997) 365–377.
- [2] G. Archer, A. Saltelli, I. Sobol', Sensitivity measures, ANOVA-like techniques and the use of bootstrap, *J. Stat. Comput. Simul.* **58** (1997), 99–120.
- [3] K. Chan, A. Saltelli, S. Tarantola, Sensitivity analysis of model output: variance-based methods make the difference, In: S. Andradottir, K.J. Healy, D.H. Withers, and B.L. Nelson, eds., *Proceedings of the 1997 Winter Simulation Conference* (1997), 261–268.
- [4] I. T. Dimov, *Monte Carlo Methods For Applied Scientists*, World Scientific (2007).
- [5] I. Dimov, Z. Zlatev, Testing the sensitivity of air pollution levels to variations of some chemical rate constants, *Notes on Numerical Fluid Mechanics* **62** (1997), 167–175.
- [6] I. Dimov, R. Georgieva, S. Ivanovska, Tz. Ostromsky, Z. Zlatev, Studying the sensitivity of pollutants' concentrations caused by variations of chemical rates, *J. Comput. Appl. Math.*, Vol. 235 (2010), pp. 391–402.
- [7] I. Dimov, K. Georgiev, Tz. Ostromsky, Z. Zlatev, *Computational challenges in the numerical treatment of large air pollution models*, *Ecological Modelling* **179** (2004), pp. 187–203.
- [8] M. Gery, G. Whitten, J. Killus, M. Dodge. *A photochemical kinetics mechanism for urban and regional scale computer modelling*, *J. Geophys. Res.* **94** (D10) (1989), 12925–12956.
- [9] W. Hoeffding, *A class of statistics with asymptotically normal distribution*, *Ann. Math. Statist.* **19** (3) (1948), 293–325.
- [10] T. Homma, A. Saltelli, *Importance measures in global sensitivity analysis of nonlinear models*, *Reliability Engineering and System Safety* **52** (1996), 1–17.
- [11] Ø. Hov, Z. Zlatev, R. Berkowicz, A. Eliassen and L. P. Prahm, *Comparison of numerical techniques for use in air pollution models with nonlinear chemical reactions*, *Atmospheric Environment*, Vol. 23 (1988), pp. 967–983.
- [12] J. Jacques, C. Lavergne, N. Devictor, *Sensitivity analysis in presence of model uncertainty and correlated inputs*, *Reliability Engineering and System Safety* **91** (2006), 1126–1134.
- [13] M. McKay, *Evaluating prediction uncertainty*, Technical Report NUREG/CR-6311, US Nuclear Regulatory Commission and Los Alamos National Laboratory (1995).
- [14] Tz. Ostromsky, I. Dimov, R. Georgieva, Z. Zlatev, *Air pollution modelling, sensitivity analysis and parallel implementation*, *Int. Journal of Environment and Pollution*, Vol. 46 (1-2), (2011), pp. 83–96.
- [15] Tz. Ostromsky, Z. Zlatev, *Parallel Implementation of a Large-scale 3-D Air Pollution Model*, in: *Large Scale Scientific Computing* (S. Margenov, J. Wasniewski, P. Yalamov, Eds.), LNCS-2179, Springer, 2001, pp. 309–316.
- [16] Tz. Ostromsky, Z. Zlatev, *Flexible Two-level Parallel Implementations of a Large Air Pollution Model*, in: *Numerical Methods and Applications* (I. Dimov, I. Lirkov, S. Margenov, Z. Zlatev - eds.), LNCS-2542, Springer (2002), pp. 545–554.
- [17] H. Rabitz, O. Alis, J. Shorter, K. Shim, *Efficient input-output model representations*, *Computer Phys. Comm.* **117** (1999), pp. 11–20.
- [18] A. Saltelli, *Making best use of model valuations to compute sensitivity indices*, *Computer Physics Communications* **145** (2002), pp. 280–297.
- [19] A. Saltelli, K. Chan, M. Scott, *Sensitivity Analysis*, John Wiley & Sons publishers, *Probability and Statistics series* (2000).
- [20] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, Halsted Press (New York, 2004).
- [21] I. Sobol', *Multidimensional quadrature formulas and Haar functions*, *Nauka, Moscow, 1969* (in Russian).
- [22] I. M. Sobol', *Sensitivity estimates for nonlinear mathematical models*, *Matem. Modelirovanie* **2** (1) (1990), pp. 112–118.
- [23] I. M. Sobol', *Sensitivity estimates for nonlinear mathematical models*, *Mathematical Modeling and Computational Experiment* **1** (1993), pp. 407–414.
- [24] I. M. Sobol', *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*, *Mathematics and Computers in Simulation*, **55** (1-3) (2001), pp. 271–280.
- [25] I. M. Sobol', *Theorem and examples on high dimensional model representation*, *Reliability Engineering and System Safety* **79** (2003), pp. 187–193.

- [26] A. Takemura, *Tensor analysis of ANOVA decomposition*, J. Amer. Statist. Assoc. **78** (1983), pp. 894–900.
- [27] WEB-site of the Danish Eulerian Model, available at: <http://www.dmu.dk/AtmosphericEnvironment/DEM>
- [28] Z. Zlatev, *Computer treatment of large air pollution models*, Kluwer (1995).
- [29] Z. Zlatev, I. Dimov, *Computational and Numerical Challenges in Environmental Modelling*, Elsevier, Amsterdam (2006).
- [30] Z. Zlatev, I. Dimov, K. Georgiev, Modeling the long-range transport of air pollutants, IEEE Computational Science & Engineering, **1** (3) (1994), pp. 45–52.