UNIVERSITY OF NOVA GORICA
GRADUATE SCHOOL

IMPROVEMENT OF THE PERFORMANCE OF AN AIR-POLLUTION DISPERSION MODEL FOR USE OVER COMPLEX TERRAIN

DISSERTATION

Boštjan Grašič

Mentors: Prof. dr. Juš Kocijan
Dr. Marija Zlata Božnar

Nova Gorica, 2008

Table of contents

List of tables

List of figures

Abbreviations

2D..... two- dimensional

3D..... three- dimensional

AP..... air pollution

BMU..... best matching unit

DBMS..... database management system

ELPD ..... enhanced Lagrangian particle dispersion

GIS..... geographical information system

LPD ..... Lagrangian particle dispersion

LPDM ..... Lagrangian particle dispersion model

LPM..... Lagrangian particle model

MPNN..... multilayer perceptron neural network

PDNC ..... particle density number coefficient

SOM..... self-organizing map

TPP..... thermal power plant

ABSTRACT

The aim of an air-pollution model is to simulate the dispersion of air pollutants in the ambient atmosphere. The impact of the air pollution from different well-known sources is studied on the basis of simulation results.

The improvements of the air-pollution modelling methodology based on the Lagrangian particle dispersion are proposed in this dissertation. The Lagrangian particle-dispersion modelling technique has recently become a well-known air-pollution reconstruction technique for use over complex terrain on the local, regional and global scales. It significantly evolved in the past ten years and it came from research use to use for regulatory purposes.

In this dissertation the limit capacities, properties and performance of the Lagrangian particle dispersion model are determined and evaluated on complex terrain. Three new methods are proposed to improve the computational performance based on these results. The new methods are developed in a manner such that the original methods in the air-pollution modelling methodology are not modified at all. The parameters, methods and structure of the original air-pollution model are preserved in their original form and no additional adjustments are performed. The methods that determine and optimize the reconstruction of the computationally expensive air-pollution situations are proposed and integrated into the existing air-pollution modelling methodology. The available computational capabilities are optimally exploited by the improved methodology.

The dissertation begins with a definition of the research problems, the goals of the research and the working hypotheses. A summarized presentation of the relevant studies is also given as an argument for using the air-pollution model based on Lagrangian particle dispersion. Two complex terrain field data sets used for experiments are described in detail. The Šaleška region field data set is selected for the presentation and validation of the new, proposed methods. Additionally, the Zasavje region field data set is used for the validation and the demonstration that the proposed methods can be generally used. The introduction ends with a presentation of the theoretical background where the basic air-pollution terminology is defined and the modelling techniques and tools used in the study are generally described.

The evaluation of the computational efficiency and the performance of the air-pollution computer model confirmed the hypothesis that the quality of the results depends on the number of virtual particles used in the air-pollution situation reconstruction. The quality of the results does not decrease significantly when the number of particles remains above a certain particle number threshold. This particle number threshold depends on the size of the domain and the size of the cells in the domain. When the number of particles falls below the threshold the quality of the results starts to decrease drastically.

The dependency between the performance of the air-pollution computer model and the number of virtual particles used in the simulations is determined by making an experimental simulation. In the experiment several simulation runs are performed on the Šaleška region field data set. The results obtained show that when the number of particles increases significantly, a strong linear dependency can be determined. This means that the time used for

the simulation run linearly depends only on the number of particles.

An evaluation is made to define the minimum acceptable number of particles that is necessary to achieve a good air-pollution reconstruction. In the evaluation process each result with a different number of used particles is compared to the reference. A comparison is performed by using the developed evaluation methods for ground concentration fields: the correlation coefficient, the root mean square error and the fractional bias.

Application of the clustering method for a reduction of the computational cost is the first contribution. It is made to decrease the computational cost by decreasing the number of active particles in the simulations. The hypothesis that the computational expenses would be reduced with the clustering method is confirmed, because the used computational complexity can be reduced by at least 50%.

The hypothesis that the use of the clustering method will have a minor influence on the quality of the results is only partially confirmed. The clustering method has a minor influence on the quality of the results only when the final number of particles after clustering in the domain remains above a certain particle number threshold.

The four basic parameters of the clustering method must be optimally set to achieve satisfactory results: N_{sub}, N_{size}, m_{max} and N_{max}. There are also some additional conclusions about the method:

- According to the finally acquired results it is concluded that the hierarchical clustering method with additional parameters can be used in practice only for the limitation of a very large number of particles. In such a case the number of particles exceeds the normal values because of the occurrence of extreme situations: like the failure of the desulphurization plant when the emissions increase by an order of magnitude or when a very stable meteorological situation with low winds occurs and the air pollution starts to accumulate in the domain.
- The strong limitation of the number of particles in the reconstructions during typical situations with the clustering method is not recommended because the quality of the results becomes very poor. To preserve the good quality of the results only a slight limitation is recommended. A comparison of the original results and the results obtained with the clustering method showed that the results obtained with the clustering method can become bubbled and less smooth than the original. This is the same effect that occurs when not enough particles are used in the reconstructions.

The cell concentration kernel density estimation method adaption is contributed to substitute the box counting concentration estimation method. It is used to improve the poor quality of the results when a smaller number of particles is used in the simulations. To evaluate the performance of the contributed method and the dependence on the input parameters, several simulations are performed.

In these experiments three cases of the optimal input parameters σ_x, σ_y and σ_z are determined

for the Šaleška region field data set according to the correlation with the reference: poor correlation, good correlation and very good correlation. The comparisons simulation results of these three different optimal points show that the final optimal parameters are set from the example where the correlation with the original result is very good.

The final comparisons prove that the correlation coefficient and the root mean square error are significantly improved. The fractional bias comparison showed that there is practically no underestimations in the results of the advanced simulation, which is crucial for long-term evaluations of air pollution. This final comparison proved the hypothesis that the poor quality of the simulation results in situations when a relatively small number of particles is used can be improved by using the kernel density concentration estimation method.

The presented final comparison also confirmed the hypothesis that the kernel density concentration estimation method can always be used to improve the quality of the results.

The Lagrangian particle-dispersion control method based on artificial neural networks also contributes to the control of the particle density parameter and the clustering parameter of the Lagrangian particle model. The rule for particle density control is the following:

- When high emissions occur the particle density should be decreased, and when low emissions occur the particle density can be increased.
- In special situations when the smallest possible particle density is used and it is still expected that the computational resources will be exceeded, the clustering must be activated to reduce the number of particles from a previous air-pollution episode reconstruction. This is very important for situations when extreme air pollution is expected. Such a common situation occurs in calm meteorological conditions when air pollution starts to accumulate in the domain for a longer time interval.

The proposed method consists of two main subsequent methods:

- in the first step the percentage of lost particles is predicted using the artificial neural network based on meteorology, emissions and the situation of the air pollution at the end of the previous episode reconstruction.
- in the second step the clustering parameters are determined using a decision-making method.

The hypotheses that the algorithm to determine the computationally complex air-pollution situations can be based on black-box modelling techniques is confirmed by using artificial neural networks successfully.

Abstract

The integration of the contributed methods into the enhanced Lagrangian particle dispersion model is finally proposed, where the mutual use of contributed methods is proposed to obtain the best possible results within the given computational resources.

The integration is realised in the enhanced Lagrangian particle dispersion (ELPD) computer model, which is validated for two field data sets:

- The Šaleška region - validation results show that the time used and the number of active particles are practically constant in the simulation run where the ELPD computer model is used. A significant improvement of the correlation coefficient and the root mean square error is achieved. The fractional bias comparison showed that the results are not over or under estimated.
- The Zasavje region - validation is performed to confirm that the proposed methods can be generally applied in various complex terrains. The results of the comparisons are practically identical to the results obtained in the Šaleška region field data set validation. The obtained results prove that the use of an enhanced air-pollution modelling methodology is recommended, not only in situations where computational resources are constrained, but also in general to optimally take advantage of the available computational resources. The computational complexity of all the air-pollution episode reconstructions is being balanced because approximately the same computational time is spent for each air-pollution episode reconstruction.

The hypothesis that controlling the number of particles in the simulation is actually preserving the quality of results at a constant level is again confirmed during the evaluation of the simulation results.

The simulation results from using both field data sets proved the hypothesis that an algorithm based on an artificial neural network is efficient and reliable and that the goal of reducing the computational complexity in simulations is successfully achieved.

Keywords: *air pollution modelling methodology, air-pollution model, Lagrangian-particle dispersion model, clustering, concentration estimation, kernel density, artificial neural network, air-pollution model evaluation*

POVZETEK

Izboljšanje okoljskega modela za rekonstrukcijo onesnaženja v ozračju nad razgibanim terenom

Metodologija modeliranja onesnaževanja ozračja zajema metode, ki se uporabljajo pri modeliranju onesnaževanja. Rezultat modeliranja je model razširjanja onesnaženja v zunanjem zraku. Model se uporablja v simulacijah za rekonstrukcijo onesnaženja zraka, katerih cilj je študija vpliva različnih virov onesnaženja na okolje.

V doktorski disertaciji je predstavljeno izboljšanje metodologije modeliranja onesnaževanja ozračja, ki temelji na Lagrange-vem modelu delcev. Postopek modeliranja na osnovi Lagrange-vega modela delcev predstavlja trenutno najbolj uporabno tehniko za rekonstrukcijo onesnaževanja ozračja nad razgibanim terenom. Izreden napredek v zadnjih desetih letih pa je omogočil tudi prenos njegove uporabe iz raziskovanega področja na področje zakonodaje, ki določa ukrepe in postopke za preprečevanje ali zmanjševanje onesnaženosti zraka iz različnih virov onesnaževanja.

V okviru disertacije so določene in ocenjene mejne zmogljivosti in lastnosti Lagrange-vega modela delcev nad razgibanim terenom. Na podlagi teh rezultatov so predstavljene tri nove metode za izboljšanje zmogljivosti. Nove metode so dodane k obstoječi metodologiji modeliranja na takšen način, da ostajajo ostale originalne metode povsem nespremenjene. Parametri, metode in struktura originalnega modela onesnaževanja ozračja so ohranjeni v njihovi prvotni obliki brez kakršnih koli sprememb. Dodatno so predstavljene tudi metode za ugotavljanje in optimizacijo rekonstrukcije računsko kompleksnih situacij onesnaženja, ki so vgrajene v obstoječo metodologijo modeliranja. Z uporabo izboljšane metodologije so računske zmogljivosti tudi bolj optimalno izkoriščene.

V uvodu disertacije so navedeni opisi raziskovalnih problemov, cilji raziskave in delovne hipoteze. Poleg tega pa je podan tudi zgoščen pregled pomembnih študij iz raziskovalnega področja razvoja in uporabe modeliranja onesnaženja na osnovi Lagrange-vega modela delcev. Zelo natančno so opisani rezultati dveh merilnih eksperimentov organiziranih v okolici večjih virov onesnaženja v okviru katerih so bili zbrani podatki o emisiji, meteorologiji in kvaliteti zunanjega zraka za določeno časovno obdobje. Rezultati merilnega eksperimenta zbrani na področju Zasavja so v disertaciji uporabljeni za predstavitev in vrednotenje predlaganih metod. Dodatno pa so rezultati merilnega eksperimenta zbrani na področju Šaleške doline uporabljeni za dodano vrednotenje metod in demonstracijo, da je možno predlagane metode uporabljati na kakršnem koli razgibanem terenu. Uvod se zaključi s predstavitvijo teoretičnih osnov, kjer je definirana terminologija, ki se uporablja na področju onesnaževanja zraka, in predstavitvijo modelirnih postopkov in orodij, ki so uporabljeni v okviru te raziskave.

Vrednotenje računske zmogljivosti računalniškega modela onesnaženja ozračja potrjuje hipotezo, da je kvaliteta rezultatov simulacij odvisna od števila uporabljenih navideznih delcev. Poleg tega pa je vrednotenje pokazalo tudi, da se kvaliteta rezultatov povečuje zelo počasi potem, ko je presežen določen prag števila navideznih delcev, ki so uporabljeni v

simulacijah. Ugotovljeno je, da je ta prag števila navideznih delcev odvisen od velikosti področja za katerega se izvaja rekonstrukcija in velikosti posameznih celic na katere je področje razdeljeno. V primeru padca števila navideznih delcev pod prag pa začenja kvaliteta rezultatov simulacij upadati zelo naglo.

Odvisnost med računsko zmogljivostjo računalniškega modela onesnaženja ozračja in številom navideznih delcev uporabljenih v simulacijah je ugotovljena eksperimentalno z izvajanjem simulacij. V simulacijah je bilo nekajkrat rekonstruirano onesnaženje zraka na področju Šaleške doline z uporabo rezultatov merilnega eksperimenta. Pridobljeni rezultati kažejo linearno odvisnost med številom uporabljenih navideznih delcev v simulaciji in časom izvajanja simulacije.

Za določitev minimalnega števila navideznih delcev v simulaciji, ki je potrebno za kvalitetno rekonstrukcijo onesnaženja zraka, je izveden dodaten eksperiment. V tem eksperimentu je vsak rezultat pridobljen z drugačnim številom navideznih delcev primerjan z referenčnih rezultatom. Primerjava je narejena z uporabo metod za vrednotenje rekonstruiranega polja talnih koncentracij in sicer s korelacijskim koeficientom, napako korena najmanjših kvadratov in metodo deleža odklona.

Uporaba metode rojenja (clustering) za zmanjšanje računske obremenitve predstavlja prvi prispevek, kjer se računska obremenitev zmanjšuje na račun zmanjšanja števila navideznih delcev v simulacijah. S tem je potrjena tudi hipoteza, da se računska obremenitev lahko zmanjša z uporabo metode rojenja. Na ta način je možno računsko kompleksnost zmanjšati vsaj za 50%.

Samo delno pa je potrjena hipoteza, da bo imela uporaba metode rojenja minimalen vpliv na kvaliteto rezultatov. Hipoteza drži samo dokler je zmanjšano število navideznih delcev večje od določenega praga števila navideznih delcev. Ta pomanjkljivost je bila odpravljena z uporabo druge metode za določitev prostorske koncentracije na podlagi jeder porazdelitvenih gostot.

Za zagotavljanje zadovoljivih rezultatov je potrebno metodi rojenja nastaviti štiri parametre: število rojev, velikost rojev, največjo dovoljeno maso navideznih delcev in največje dovoljeno število navideznih delcev. Pri uporabi metode pa je potrebno upoštevati še naslednje zaključke in priporočila:

- Metodo rojenja je priporočeno uporabljati za zmanjševanje samo v primerih relativno velikega števila navideznih delcev. To so primeri, ko nastopijo različne ekstremne situacije kot so na primer: izpad čistilne naprave na odvodnikih dimnih plinov, ko se emisije povišajo za razred velikosti ali pa v primerih, ko nastane zelo stabilna meteorološka situacija in se začne akumulacija onesnaženja na izbranem področju.
- Uporaba metode rojenja za močno zmanjšanje števila navideznih delcev je neprimerna za praktično uporabo, ker lahko povzroči drastično zmanjšanje kvalitete rezultatov. Za ohranitev kvalitetnih rezultatov je priporočeno samo rahlo zmanjšanje. Primerjava

rezultatov z zmanjšanim številom navideznih delcev z originalnimi je pokazala, da zmanjševanje lahko povzroča manj enakomerno porazdelitev koncentracij, kar je enak rezultat kot v primeru uporabe premajhnega števila navideznih delcev.

Prilagoditev metode za določitev prostorske koncentracije na podlagi jeder porazdelitvenih gostot je predlagana za nadomestitev metode za določitev prostorske koncentracije na podlagi štetja delcev v prostorskem elementu. Prilagojena metoda je namenjena izboljšanju rezultatov slabše kakovosti v primerih, ko je v simulacijah uporabljeno prenizko število navideznih delcev. Ocena zmogljivosti predlagane metode in odvisnost od vhodnih parametrov je narejena na podlagi rezultatov simulacijskih eksperimentov.

V simulacijskih eksperimentih opravljenih za področje Šaleške doline so določeni tudi trije optimalni vhodni parametri razpršitve navideznih delcev σ_x, σ_y in σ_z glede na vrednost korelacijskega koeficienta: odlična korelacija, srednje dobra korelacija in slaba korelacija. Primerjava rezultatov simulacij pridobljenih s temi tremi različnimi optimalni vhodi pokaže, da so najbolj optimalni vhodni parametri tisti, ki so bili pridobljeni v primeru odlične korelacije.

Končna primerjava rezultatov kaže na relativno močno izboljšanje korelacijskega koeficienta in napake korena najmanjših kvadratov. Vrednotenje z metodo deleža odklona pa kaže, da praktično ni prišlo do nobenega podcenjevanja ali precenjevanja koncentracij, kar je še posebej pomembno za rekonstrukcije onesnaževanja zraka za daljše časovno obdobje. Ta končna primerjava tudi potrjuje hipotezo, da je z uporabo metode na podlagi jeder porazdelitvenih gostot možno izboljšati rezultate tudi v primerih, ko je kvaliteta rezultatov slabša zaradi relativno nizkega števila uporabljenih navideznih delcev. Poleg tega pa je končna primerjava potrdila hipotezo, da je z uporabo metode jedra gostote porazdelitve vedno možno izboljšati kvaliteto rezultatov.

Regulacija Lagrange-vega disperzijskega modela delcev, ki temelji na osnovi umetnih nevronskih mrež, je predlagana za vodenje parametra gostote navideznih delcev in parametra rojenja. Pravilo vodenja je naslednje:

- V primeru velike emisije se gostota navideznih delcev zmanjša in v primeru nizke emisije se gostota navideznih delcev poveča do te mere, da skupno število navideznih delcev med simulacijo ne preseže vnaprej določene meje.
- V posebnih situacijah, ko je uporabljena najmanjša gostota navideznih delcev in se pričakuje, da bo skupno število navideznih delcev še vedno preseglo določeno mejo, pa se aktivira rojenje, ki dodatno zmanjša število navideznih delcev, ki se že nahajajo na področju rekonstrukcije. To je zelo pomembno za situacije, ko nastopijo izjemni primeri onesnaženja zraka, kot je na primer brezvetrje, ki traja daljše časovno obdobje in povzroča akumulacijo onesnaženja v domeni.

Predstavljeno metodo dejansko sestavljata dva osnovna koraka:

- V prvem koraku je izvedena predikcija odstotka izgubljenih navideznih delcev z umetno nevronsko mrežo. Vhode v umetno nevronsko mrežo predstavljajo meteorološki in emisijski parametri ter trenutna situacija onesnaženja zraka v področju.
- V drugem koraku pa se določita parametra gostote navideznih delcev in rojenja na osnovi preproste odločitvene metode.

Uporaba metode na osnovi umetnih nevronskih mrež potrjuje hipotezo, da lahko metoda za ugotavljanje računsko kompleksnih situacij temelji na osnovi modela črne skrinjice.

Integracija vseh treh predstavljenih metod v razširjen Lagrange-ev model delcev prestavlja zadnjo izboljšavo. Vzajemna uporaba predstavljenih metod omogoča pridobitev najbolj optimalnih rezultatov glede na razpoložljive računske zmogljivosti.

Integracija je izvedena tudi v obliki razširjenega računalniškega Lagrange-vega modela delcev, ki je vrednoten na dveh področjih za katera so na voljo podatki iz merilnih eksperimentov:

- Šaleška dolina – rezultati vrednotenja kažejo da sta porabljen računalniški čas in število navideznih delcev med simulacijskim tekom praktično konstantna pri uporabi razširjenega računalniškega Lagrange-vega modela delcev. Poleg tega pa se kaže tudi izboljšanje korelacijskega koeficienta in napake korena najmanjših kvadratov ter ni zaznati nobenih podcenjevanj ali precenjevanj glede na vrednotenje z metodo deleža odklona.
- Zasavje – vred je narejena za potrditev hipoteze, da je možna splošna uporaba predlaganih metod na različno razgibanih terenih. Rezultati vrednotenja so praktično enaki rezultatom vrednotenja na Šaleški dolini. Pridobljene izkušnje kažejo, da uporaba razširjene metodologije modeliranja onesnaževanja zraka ni priporočljiva samo za primere, kjer so računske zmogljivosti omejene, ampak tudi v splošnem za bolj optimalno izkoriščanje računskih zmogljivosti. V splošnem je z uporabo predstavljenih metod računska zmogljivost bolj uravnoteženo porazdeljena med rekonstrukcije posameznih epizod onesnaženja zraka

S tem je potrjena hipoteza, da nadzor števila navideznih delcev v simulacijah ohranja kvaliteto rezultatov na konstantnem nivoju. Poleg tega pa rezultati obeh vrednotenj potrjujejo tudi hipotezi, da je algoritem na osnovi umetnih nevronski mrež zanesljiv in učinkovit ter da je s tem cilj zmanjšanja računske kompleksnosti v simulacijah uspešno dosežen.

Ključne besede: *metodologija modeliranja onesnaževanja ozračja, model onesnaževanja ozračja, Lagrange-ev model delcev razširjanja onesnaženja, metoda rojenja, določitev prostorske koncentracije na podlagi jeder porazdelitvenih gostot, umetne nevronske mreže, vrednotenje modela onesnaževanja ozračja*

1. INTRODUCTION

1.1. General introduction

The aim of an air-pollution model is to simulate the dispersion of air pollutants in the ambient atmosphere, and the results of the simulations are used to study the impact of the air pollution from certain sources on the surrounding environment. The outcomes of the studies can be used:

- by governmental agencies to protect and manage the quality of the ambient air[1,2];
- in the engineering processes of new air-pollution facilities to reduce the impact on the environment by designing optimal stacks and determining the best location[3,4];
- by local communities or states to protect[5] and, in the worst case, also to evacuate the population most effectively during severe accidents involving air-borne releases[6].

The development of air-pollution modelling techniques began at the beginning of the 20th century. Perhaps the oldest and most commonly used model type is the Gaussian model[7], first described in 1936, and subsequently updated to its more generally known form by Sutton[8] in 1947. The possibility of reconstructing air pollution with the Gaussian model is limited to the situation of stable meteorological conditions over simple terrain. However, the rapid development of microcomputers since the 1980s has given rise to the development of other air-pollution modelling techniques. From among these models the air-pollution modelling technique based on the Lagrangian particle-dispersion model has become the best and most frequently used air-pollution reconstruction technique for use over complex terrain on the local, regional and global scales[9,10]. Its intense development started in the early 1980s and it is known to be a computationally very demanding method. At the beginning, in the early 1990s, the method was used only on small, complex domains, where only one or a few simple air-pollution sources were present. It was limited to research use, because of required computational power, which was enormous for those times.

Air-pollution modelling based on the Lagrangian particle dispersion has evolved rapidly over the past ten years, and it has developed from being just a research tool to being used for operational regulatory purposes[11,12,13]. All the improvements and adjustments were carefully developed and evaluated on a number of field-data sets from simple to complex terrains[11,14,15,16,17,18,19,20,21,22,23,24,25]. In the past ten years significant progress was also achieved in the computer industry, thus making possible the use of Lagrangian particle-modelling techniques on personal computers. Simple problems that could be simulated only for a short period of time several years ago on dedicated workstations (super computers), can these days be in a much more complex form and simulated for a longer period of time on personal computers. In recent years the general opinion is that the technique does not need any computational improvements because the time-consuming problems will be solved by the development of computers with greater processing power. This is definitely a solution for today's requirements. Unfortunately, however, future requirements are expanding rapidly: the areas of interest are becoming wider, the resolution demands are increasing, the accuracies must be

better, a lower sensitivity is demanded, longer periods of time must be reconstructed, and the complexity and the number of sources and species are increasing, etc. These increasing requirements bring us back to the initial problem of the Lagrangian particle-modelling technique, which is extremely expensive from the computational point of view.

In this thesis the air-pollution modelling methodology based on Lagrangian particle dispersion is presented. Its limit capacities, properties and performance are determined and evaluated for a complex terrain. Based on these results several methods are suggested to improve the computational performance. The new methods are developed in a manner such that the basic physical properties of the Lagrangian particle-dispersion model are not modified. All the parameters and methods of the model are preserved in their original form and there are no adjustments of the well tuned parameters. Methods to determine and simplify computationally expensive situations are developed and integrated into the existing Lagrangian particle-dispersion model to improve the computational performance in order to optimally exploit the available computational capabilities.

1.2. Definition of the research problems

Air pollution is a comprehensive term. It applies to the spreading of any chemical, physical or biological agent that modifies the natural properties of the atmosphere. Atmospheric dispersion models are used to present how pollutants in the ambient atmosphere disperse and, in some cases, how they react in the atmosphere with other atmospheric compounds. They are important to governmental agencies whose task is protecting and managing the quality of the ambient air. Several different air-pollution modelling approaches and techniques can be used to reconstruct the state of the atmosphere around the air-pollution sources.

The time evolution of the air-pollution dispersion of the emitted pollutant depends on the wind speed and direction, the turbulence, temperature, humidity, air pressure, solar radiation and precipitations as well as on the terrain's complexity. The meteorological parameters (variables) are usually four-dimensional, which means that they depend on space and time. The terrain's complexity, especially the biocoenosis[a], does not usually change so rapidly with time as the meteorological parameters, but it must also be considered as four-dimensional (4D) because it changes significantly during the seasons of year. Reconstruction of the air-pollution evolution is, in principal, understanding the meteorological conditions and the terrain's complexity. While some of these conditions can cause high concentrations of an air pollutant near the surface for an extended period of time, some other conditions can decrease the concentration levels much more rapidly.

Most commonly used atmospheric dispersion modelling techniques are based on the Gaussian plume model[26,27], the Lagrangian particle-dispersion model[9,12] and the Eulerian[28,29] transport model. Approaches differ in several parameters, like scale (local, regional, global), grid size (from 10 m to 200 km), resolution, source type (line, point, area, volume), pollutants, statistical elaboration time interval (1 hour, 1 day, etc.), topography (simple, complex) and others.

a A group of interacting animals and/or plants that form a particular ecosystem

Air-pollution models based on Lagrangian particle dispersion[9,10] are generally accepted as the most powerful tools to model the dispersion of atmospheric pollutants in the boundary layer over the local scale domain. Their quality has continually improved over the past 10 years. Most of the improvements gradually increased the complexity of the models. This increased complexity leads to their main disadvantage: air-pollution reconstruction based on the Lagrangian particle-dispersion model can become computationally very expensive. This disadvantage gets emphasized when complex terrain conditions are present. The time used for each air-pollution episode reconstruction is critical when used in on-line systems as well as for off-line systems. Both types of system, the on-line and the off-line, are of specific importance to our work. For an illustration of the problems of each type of system two examples are given:

- A domain is 20 km x 20 km wide and 2000 m high. It is split into vertical 20 layers and each layer consists of 100x100 cells. The size of each cell is 200 m x 200 m x 100m. The result of the reconstruction is concentration in each cell of the domain. There is one source of air pollution (i.e., the stack from the power plant) from which only one species (i.e., SO_2) is emitted. For the reconstruction of the air pollution over a period of one year at least 50 days of simulation time is spent on a high-performance personal computer (i.e., AMD Athlon 64 X2 Dual Core Processor 3800+, 2.01 GHz, 2 GB of RAM). Usually, this much time is not available in practice and results with a lower resolution are used, which still satisfy the requirements. To accomplish this task, sometimes clusters of computers are also used, but the requirements (i.e., higher resolution, larger domains, etc.) are increasing faster. There is an effort to use air-pollution models based on Lagrangian particle dispersion for long-term statistical elaborations when air pollution is simulated for a very long interval of time (i.e., the simulation of air pollution for a one-year time interval with a time step of half an hour).
- A lot of effort is also invested in the use of air-pollution models based on the Lagrangian particle dispersion in mobile systems. These mobile systems contain limited computational means. In addition to the hardware configuration there is also an energy limitation that limits the computational performance. Such mobile systems are used by mobile groups of administration for civil protection and disaster relief in the case of accidental air-pollution releases. As an illustration, the reconstruction of air pollution on a low-power mobile system can last for more than half an hour for the example that has been given in the previous paragraph. For the practice where the time between measurements is half an hour the air-pollution reconstruction is not fast enough to be declared as an on-line system.

The purpose of this thesis is to determine expensive procedures in an air-pollution model based on Lagrangian particle dispersion and to consider some suggestions for how to improve them. The need for improvements is important to accomplish the efforts that have been presented because the improvements of the computer performance cannot match the increasing requirements.

1.3. Purpose of the research and the working hypothesis

Several goals and hypotheses are defined in the following two sections to overcome the presented research problems.

1.3.1. Goals of the research

The main goals of the research, considering the improvement of the performance of the Lagrangian particle-dispersion model for use over a complex terrain are:

1. the computational complexity in simulations is going to be reduced,
2. efficient clustering algorithms to reduce the number of particles in the simulation is going to be applied,
3. an algorithm to determine when clustering of the particles is needed in the simulation cycle is going to be proposed,
4. an algorithm to control the number of particles used in the simulation to ensure a good result is going to be proposed,
5. a method to improve the results of the simulations when a relatively small number of particles are used is going to be proposed.

1.3.2. Hypotheses

The hypotheses for the improvement of the performance of the Lagrangian particle-dispersion model for use over complex terrain are:

1. the computational expenses will be reduced with use of clustering algorithms to decrease the number of particles in the simulation,
2. the use of clustering algorithms to decrease the number of particles in the simulation will have a minor influence on the quality of the results,
3. the algorithm to determine situations when the particle number should be reduced will be based on a black-box modelling technique,
4. the algorithm to determine situations when the particle number should be reduced will be efficient and reliable,
5. the computational expenses of the simulation with a clustering algorithm will be much lower than the computational expenses of the original simulation,
6. the quality of the results depends on the number of particles used in the simulations,
7. controlling the number of particles in the simulation preserves the quality of the results at a constant level,
8. the poor quality of simulation results in the situations when a relatively small number of particles are used can be improved by using the kernel density concentration estimation method
9. the kernel density concentration estimation method can always be used to improve the quality of the results.

1.4. Theoretical background

1.4.1. Air pollution

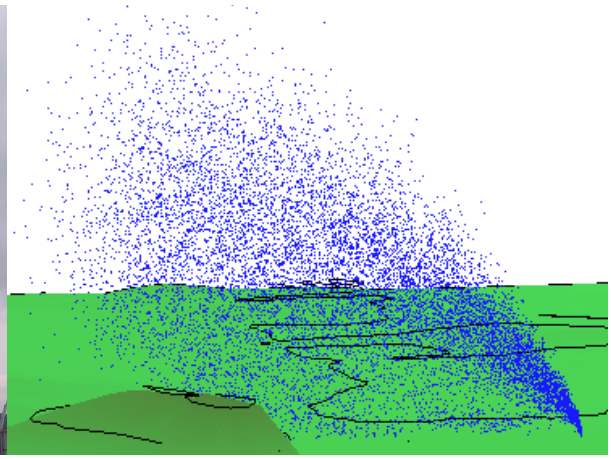
Air pollution is a comprehensive term. It applies to any chemical, physical or biological agent that modifies the natural composition of the atmosphere. It is the contamination of air by the discharging of harmful substances. Air pollution can cause health problems and it can also damage the environment and property[30]. Air pollutants can be classified into two major groups:

- *primary pollutants*: are directly released from a given source (such as carbon monoxide CO, sulphur dioxide SO_2 or nitrogen oxides NO_x, all of which are by-products of combustion),
- *secondary pollutants*: are formed in the atmosphere by subsequent chemical reactions involving direct release pollutants (a typical example is the formation of ozone in photochemical smog).

Many researchers and scientists are using different approaches to model air pollution. In general, air-pollution modelling can be classified in two major groups[31, 32]:

- *atmospheric dispersion models*: include those models that are attempting to reconstruct actual physical processes. This group will be the object of the research and it will be presented in detail in the following paragraphs. It consists mostly of plume-rise models, Gaussian models, Eulerian models, semi-empirical models, Lagrangian models and chemical models. Illustration of a reconstruction of an air pollution dispersion using Lagrangian-particle atmospheric dispersion model is presented in Figure 1.
- *forecasting air-pollution models*: include those models that are attempting to forecast air-pollution concentrations at some precisely defined locations on an observation area for some time in future. In general, these types of models are not usually necessarily based on a dispersion mechanism. The group of forecasting models consists mostly of receptor models and grey-box models. Illustration of an ozone forecasting air-pollution model based on artificial neural network is presented in Figure 2.

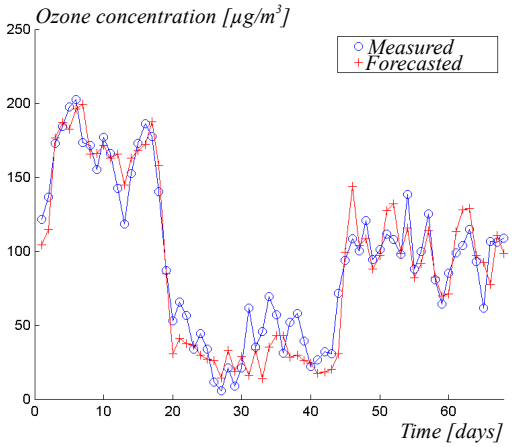
AIR POLLUTION DISPERSION



3D PRESENTATION OF RECONSTRUCTED AIR POLLUTION DISPERSION USING LAGRANGIAN-PARTICLE DISPERSION MODEL

Figure 1: Illustration of a reconstruction of an air pollution dispersion using Lagrangian-particle atmospheric dispersion model

AIR POLLUTION OF OZONE AT THE POSITION OF AN OZONE SENSOR



COMPARISON OF MEASURED AND FORECASTED OZONE CONCENTRATION AT THE POSITION OF AN OZONE SENSOR

Figure 2: Illustration of an ozone forecasting air-pollution model based on artificial neural network at the position of air pollution monitoring station located in Nova Gorica

1.4.2. Air-pollution dispersion

Air-pollution atmospheric dispersion (spreading) in the atmosphere is a complex process. A number of different approaches are used to model it. Usually, simulations are done to show how pollutants in the ambient atmosphere disperse and, in some cases, how they react in the atmosphere. The concentrations of air pollutants in the atmosphere over a modelling domain are reconstructed according to information about emissions and the meteorological situation.

A developed methodology for analysing and predicting air pollution in the atmosphere is

known by the term *atmospheric dispersion modelling*. It is based on a variety of mathematical models that simulate how air pollutants disperse in the atmosphere.

1.4.3. Air-pollution dispersion models

Air-pollution dispersion models are also known as air-dispersion models, air-quality models, and atmospheric dispersion models. They are used for simulations of the process of air-pollutants dispersion in the ambient atmosphere. Simulations are performed with computer programs that solve the mathematical equations and algorithms that simulate the pollutant dispersion. The dispersion models are used to estimate or to predict the downwind concentration of the air pollutants emitted from sources such as industrial plants and vehicular traffic. Such models are important to governmental agencies whose task is to protect and manage the ambient air quality[33].

Air-pollution dispersion models require the input of data that includes:

- *Meteorological conditions*: such as wind speed and direction, the amount of atmospheric turbulence (as characterized by what is called the "stability"), the ambient air temperature and the height of any inversion aloft that may be present.
- *Emissions parameters*: such as source location and height, source exhaust stack diameter and exit velocity, exit temperature and mass flow rate.
- *Topography description* of the area of interest where, besides the three-dimensional topography data of the domain, also two-dimensional land-use data of the domain is required.
- *The location, height and width of any obstructions*: (such as buildings or other structures) in the path of the emitted gaseous plume.

1.4.4. The classification of air-pollution dispersion models

The European Topic Centre on Air and Climate Change (ETC/ACC) classifies air-pollution dispersion models by their properties[34], while the U.S. Environmental Protection Agency (EPA) classifies models into four categories[27]: preferred and recommended models, alternative models, screening tools and related programs. Because the ETC/ACC categorization is based on model properties, some selected model properties will be described according to the ETC/AAC categories:

- *air-pollution models*: Gaussian models, Eulerian models, Lagrangian models, plume-rise models, chemical modules, receptor models, stochastic models or semi-empirical models.
- *area of interest*: local (up to 30 km), local-to-regional (30-300 km), regional-to-continental (300-3000 km) or global (hemispheric to global scale).
- *release type*: a continuous or accidental release is defined as any spilling, leaking, pumping, pouring, emitting, emptying, discharging, injecting, escaping, leaching, dumping, or disposing into the environment of a hazardous or toxic chemical or

extremely hazardous substance or any component not naturally present in the atmosphere.

- *emission sources*: a source is defined as any place or object where pollutants are emitted from. A source can be a power plant, a factory, a dry-cleaning business, a gas station or a farm. Cars, trucks and other motor vehicles as well as consumer products and machines used in industry are also designated as sources. Sources fixed in space are referred to as stationary sources, whereas moving sources, such as cars or planes, are called mobile sources. Sources can also be classified more specifically: emissions from the stack of a plant (point source), traffic emissions (line source), area and volume sources (waste dump). Multiple sources can be treated individually.
- *terrain*: An ideal dispersion, as described by the Gaussian plume dispersion model, rarely occurs in real applications. The formation of extraordinary meteorological phenomena will usually influence the dispersion of pollutant emissions. In some cases of negligible deviations from the ideal there is no need to require the use of physical modelling approaches. But when the formation of extraordinary meteorological phenomena has a major influence on the dispersion of pollutants the use of physical modelling approaches should be invoked. For modelling in such environmental conditions the term *complex terrain* is used. Some attempts at defining a complex terrain have been made in the context of a regulatory framework[35,36] that will be described in the next subsection 1.4.5 Complex terrain.

1.4.5. Complex terrain

When taking a view of the world as a globe, the orography of the earth is revealed as almost imperceptible bulges and depressions. The most noticeable bulges that represent the highest mountain barriers only extend the radius of the earth by about 0.1% from its sea-level value,[36] while depressions are even less noticeable. In general it is the presence of mountains and hills with their endless varieties of slopes, passes and valleys that makes the conditions for the formation of countless extraordinary meteorological phenomena. For these kinds of environmental situations the term *complex terrain* is used to define the meteorological conditions that appear in the process of modelling the dispersion of pollutants.

The term is the opposite of *simple terrain*, which defines the flat or very slightly rough terrain with the condition of a strong, stable wind speed[37,38]. When these conditions are not met, the term *complex terrain* is used.

Some attempts to define a complex terrain have been made in the context of a regulatory framework. Using the definition of [35], a site is considered to be a complex terrain if:

1. The pollutant release height is less than two times the maximum terrain height. The maximum terrain height in this case is defined as the difference between the highest level (including tree tops) and the lowest level within the larger of twenty times the stack height or 1 km from the source complex.
2. In addition, the slope of the terrain represented by the gradient of the terrain height

with distance from the source complex must be greater than 1/5.

Similarly, there are criteria that distinguish when building turbulence may influence the plume dispersion[35]. Some examples are:

1. The dispersion of material released on or near a building or structure is assumed to be influenced by structure-induced turbulence if the release height is less than two times the building height.
2. If the release height is less than two times the building height, but is equal to or greater than ten times the building width, the pollutant emission will be considered to be outside the building-induced turbulence.
3. For a source upwind of a building, the influence of the structure on the plume dispersion will extend upwind a distance of 1.3 times the building height.

Using another definition from that EPA[39], that was made for the purposes of meteorological monitoring guidance for regulatory modelling applications, the term *complex terrain* is intended to mean any site where terrain effects on the meteorological conditions may be significant. Terrain effects include aerodynamic wakes, density-driven slope flows, channelling, flow accelerations over the crest of terrain features, etc. These flows primarily affect wind-speed and wind-direction measurements; however, temperature and humidity measurements may also be affected. The definition of significance depends on the application: for regulatory dispersion modelling applications, the significance is determined by comparing the stack-top height and/or the estimated plume height with the terrain height:

- the terrain that is below the stack-top is classified as simple terrain,
- the terrain between the stack-top height and the plume height is classified as intermediate terrain,
- the terrain that is above the plume height is classified as complex terrain.

The perturbations to meteorological variables caused by rugged terrain are discussed in detail in the meteorological monograph by Blumen et al.[36].There is also the description of the pollution problems that are usually thought to be more serious in complex terrain because of the decrease in the wind speed and the possibility of the plume's impact on high terrain. Despite some exceptions where air-pollution problems are mitigated (due to increased turbulence intensities, no decrease of the plume rise due to strong-wind conditions and the deflection of plumes around obstacles), there are some certain air-pollution situations that can lead to increased concentrations in complex terrain[36]:

- the plume impingement on high terrain, illustrated in Figure 3,
- the pooling in valleys, illustrated in Figure 4,
- the drainage toward population centres, illustrated in Figure 5
- the persistence due to channelling, illustrated in Figure 6.

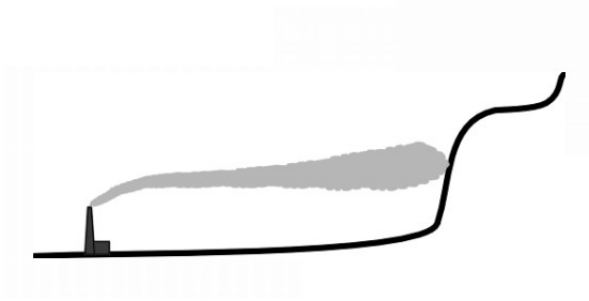


Figure 3: Plume impingement on high terrain[36]

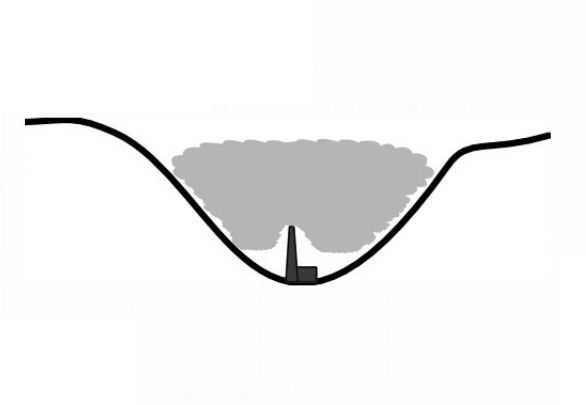


Figure 4: Pooling in valleys[36]

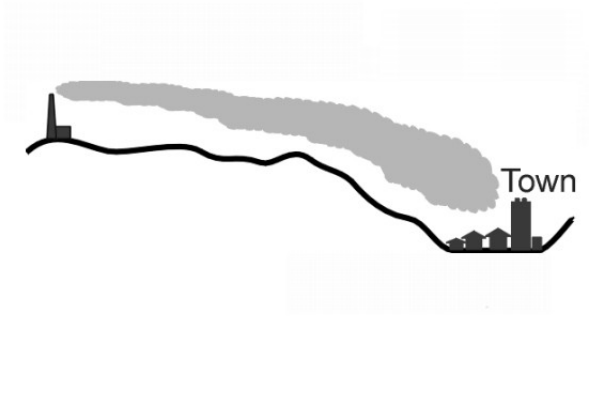


Figure 5: Drainage toward population centres[36]

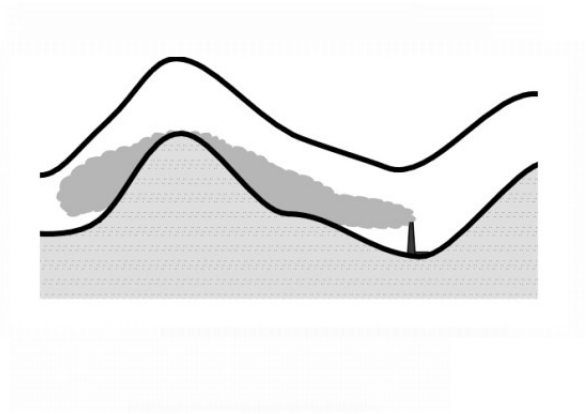


Figure 6: Persistence due to channelling[36]

1.4.6. The Lagrangian particle-dispersion model

The Lagrangian approach is based on a description of the flow path of the fluid elements. They include all the models in which plumes are split into elements, such as segments, puffs or particles. The point-like particles are representing a trace species. de Baas et al. illustrated[40] that most particle modelling studies of air-pollution dispersion are numerical solutions of the Lagevin stochastic differential equation[41] (1.1).

$$\frac{d\,\boldsymbol{v}_p(t)}{dt} = -\boldsymbol{T}(\boldsymbol{r}_p, t)\cdot\boldsymbol{v}_p(t) + \boldsymbol{u}_p(t) \tag{1.1}$$

$$\frac{d\,\boldsymbol{r}_p(t)}{dt} = \boldsymbol{v}_p(t) \tag{1.2}$$

$\boldsymbol{r}_p(t) \in R^n$ ...particle position vector
$\boldsymbol{v}_p(t) \in R^n$ ...particle velocity vector
$\boldsymbol{T}(\boldsymbol{r}_p, t) \in R^{n\times n}$ ...second-order tensor of Lagrangian time scales of speed components
$\boldsymbol{u}_p \in R^n$ ...vector of random-speed fluctuations, also called random forcing

The use of this equation for Lagrangian air-pollution dispersion[21] is described in detail by Thomson's theory developed in 1984 and the one derived from a subsequent work accomplished in 1987 in his papers[23,24]. According to this theory point-like particles in computer simulations are tracked on their path through the atmosphere with this type of model according to discrete equations (1.3) and (1.4) for each particle. Discretization of equation (1.1) that is necessary for computer simulation is described in detail in papers by Thomson[23,24] and Brusasca[21].

$$\boldsymbol{v}(t+\tau) = \mathbf{A}(\boldsymbol{r}, \tau)\cdot\boldsymbol{v}(t) + \boldsymbol{b}(\boldsymbol{r}, \tau) \tag{1.3}$$

$$\boldsymbol{r}(t+\tau) = \boldsymbol{r}(t) + \tau\left[\bar{V}(\boldsymbol{r}) + \boldsymbol{v}(t)\right] \tag{1.4}$$

$\boldsymbol{r}(t) \in R^n$ ...discrete particle position vector
$\boldsymbol{v}(t) \in R^n$ ...discrete particle velocity vector
τ ...time step

The principle is illustrated in Figure 7. The particles are moved according to a reconstructed mean wind field $\bar{V}(\boldsymbol{r})$ and are additionally subjected to the influence of turbulence, where $\mathbf{A}(\boldsymbol{r}, \tau)$ is a function determined by the atmospheric turbulence and $\boldsymbol{b}(\boldsymbol{r}, \tau)$ is a velocity randomly chosen with each time step. The turbulence effect is modelled by adding a random velocity to the mean motion for each particle. The random velocity derived by the Markov process is a function of the turbulence intensity. The concentrations of a pollutant in given sampling volumes are calculated by counting the particles.

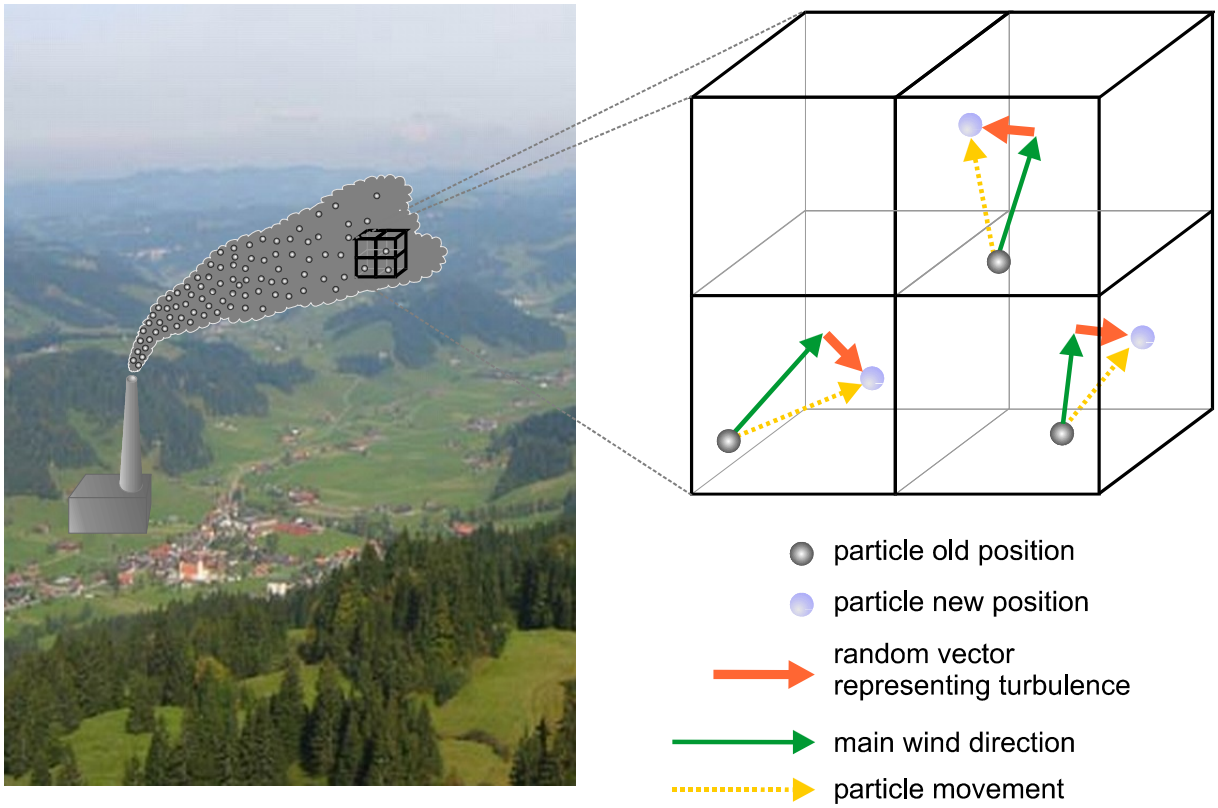


Figure 7: Lagrangian particle model principle

The main advantages of the model are[12,21]:

- the model's concept largely reflects the natural phenomena involved in turbulent diffusion,
- it eliminates numerical diffusion[b],
- it always yields non-negative mass densities and is mass-conserving,
- it can be applied to any source geometry desired for any temporal behaviour of a spatially variable source,
- it can accommodate the sedimentation of heavy dust and its re-suspension,
- it can accommodate physical and linear chemical transformation processes.

Lagrangian models perform especially very well on complex terrain on the local[37], regional and continental[42] scales. A three-dimensional (3D) presentation of the simulation result is shown in Figure 8, where the air-pollution situation is depicted using virtual particles.

b Numerical diffusion is an unwanted diffusion that occurs during computer simulations of continuous systems. Time and space in Eulerian equations is divided into a discrete grid. Continuous differential equations of motion are discretized into the finite-difference equation. In general the discrete equations are more diffusive than the original differential equations. The difference that occurs between the real system and the simulation depends on the system that is simulated and the type of discretization that is used[12, 21].

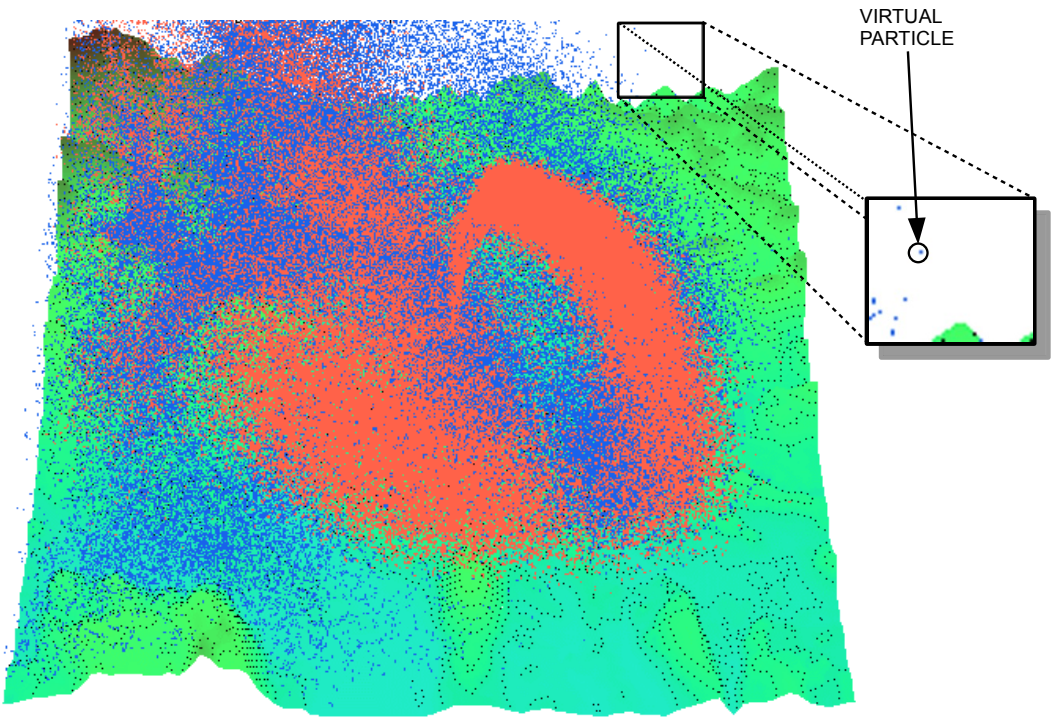


Figure 8: Lagrangian particle model simulation result

1.4.7. Black-box modelling techniques

Several improvements of the Lagrangian particle model are going to be proposed in this thesis. Because the improvements are also based on feedforward multilayer neural networks some basic theory of black-box modelling techniques is given in the following paragraphs.

System identification is used as a general term for describing the mathematical tools and methods that are used to make dynamic models from measured data. In this context a dynamic mathematical model is a mathematical description of the input-output behaviour caused by the dynamics of a system or process. The main issue in system identification[43,44,45] is finding a suitable model structure, within which a good model can be made. Prior knowledge and a physical insight about the system should be utilized when selecting the model structure. We distinguish between three levels of prior knowledge, which have been colour-coded as follows[43]:

- *White-box models,* also known as first-principles models, are used in the case when it has been possible to construct a model entirely from prior knowledge and physical insight; the model is perfectly known.
- *Grey-box models* are used when some physical insight is available, but several parameters remain to be determined from the observed data. Two sub-cases are considered:
  - *Physical modelling*: a model structure can be built on physical grounds, and has a

certain number of parameters to be estimated from data.

- *Semi-physical modelling*: a physical insight is used to suggest certain non-linear combinations of the measured data signal. These signals are then subjected to model structures of a black-box character.

- *Black-box models* are used when no physical insight is available or used, but the chosen model structure belongs to families that are known to have good approximation abilities.

A non-linear black-box structure for a dynamic system is a model structure that can describe virtually any non-linear dynamics[43,44,45]. The area is quite diverse and covers topics from mathematical approximation theory, via estimation theory and non-parametric regression, to algorithms and concepts like neural networks, wavelets and fuzzy models. The system identification problem consists of observed inputs *u(t)* and outputs *y(t)* from a dynamic system,as described in equations (1.5,1.6):

$$\boldsymbol{u}^{t-1}=[u(1),u(2),\ldots,u(t-1)]; \qquad u\in R^n \tag{1.5}$$

$$\boldsymbol{y}^{t-1}=[y(1),y(2),\ldots,y(t-1)]; \qquad y\in R^n \tag{1.6}$$

The goal of a system identification is to find a relationship between past observations $[\boldsymbol{u}^{t-1},\boldsymbol{y}^{t-1}]$ and the future output *p(t),* as described in equation (1.7):

$$p(t)=g(\boldsymbol{u}^{t-1},\boldsymbol{y}^{t-1})+v(t) \tag{1.7}$$

If $g(\boldsymbol{u}^{t-1},\boldsymbol{y}^{t-1})$ is considered as a good prediction of *p(t),* given past data, the goal is to reduce the additive term *v(t)* to as small a value as possible.

A very rich spectrum of possible model descriptions must be handled and nothing should be excluded when non-linear black-box identification is used. Some of the possibilities and limitations are discussed in the paper by Sjöberg et al.[43] This paper also has a companion paper by Juditsky et al.,[46] which complements the material with more theoretical aspects and provides a more mathematically comprehensive treatment.

Feedforward multilayer perceptron neural networks

Artificial neural networks are selected for the proposed improvements because it was proven by Hornik et al. that multilayer feedforward networks are universal approximators[47]. They have become a useful and efficient tool, in the past ten years, for establishing forecasting models in the field of air pollution. Many authors reported the successful forecasting of air pollution using artificial neural networks in recent years. An overview of applications is given by Gardner and Dorling[48]. Other examples of the use of black-box modelling techniques were presented in the paper by Grašič et al[49]. In the research a perceptron neural-network-based model and a Gaussian process prior model for ozone-concentration forecasting for the city of Nova Gorica was developed and evaluated. The methods of input determination and the selection of data for learning the model training process are the most crucial steps in the modeling techniques based on neural networks[31,32,50,49,51].

Artificial neural networks are very often used for system identification and control because of their ability to learn nonlinear relationships. An introduction to artificial neural networks and their use for system identification and control is presented in the book by Kocijan[52]. In the book a summarized systematic classification of neural-network-based control is given with the illustrative example of the predictive functional control of a simple mathematical model where the process is described by a nonlinear differential equation. The summarized classification is based on the paper by Agarwal[53] and on the book by Nørgaard et al.[54].

An artificial neural network is a mathematical model, or its computational implementation, developed from observations of biological neural networks. It consists of an interconnected group of nodes known as artificial neurons, where each node is a relatively simple (non)linear system. It is often created as an adaptive system that changes its structure according to the external or internal information that flows through the network during the learning phase[52,55,56].

The feedforward multilayer perceptron neural network was the first developed artificial neural network that was proved to be universal approximator by Hornik et al.[47]: the universal approximation theorem for neural networks states that every non-linear continuous function that transforms intervals of real numbers to some output intervals of real numbers can be approximated arbitrarily closely by a multi-layer perceptron with just one hidden layer. The theorem holds only for restricted classes of non-linear activation functions like, for example, the sigmoid functions. In practice the feedforward neural networks represent a non-linear statistical data modelling tool and can be used to model complex relationships between inputs and outputs[52,55,56].

Artificial neural networks have become a useful and efficient tool in the past ten years for establishing forecasting models in the field of air pollution. The successful forecasting of air pollution using artificial neural networks in recent years was reported by many authors[49,51,57,58,59,60,61,62,63]. An extensive overview of the applications is presented in the paper by Gardner and Dorling[48].

The feedforward multilayer perceptron neural network consists of an interconnected system of nodes (artificial neurons or perceptrons), as presented in Figure 9[52,55,56,64]. The nodes are distributed among several standard layers: one input layer, one or several hidden layers and one output layer. The number of nodes in the input layer is equal to the number of inputs, and the number of nodes in the output layer is equal to the number of outputs. The number of nodes in the hidden layers is arbitrary and is usually determined by the experiences gained in the learning process, or other methods and approaches. Nodes in a particular layer are not inter-connected. Only the outputs of a particular layer are connected over weights to the inputs of the next layer. The information flows in only one forward direction: from the input nodes, through the hidden nodes and finally to the output nodes. There are no loops or cycles in the feedforward neural network.

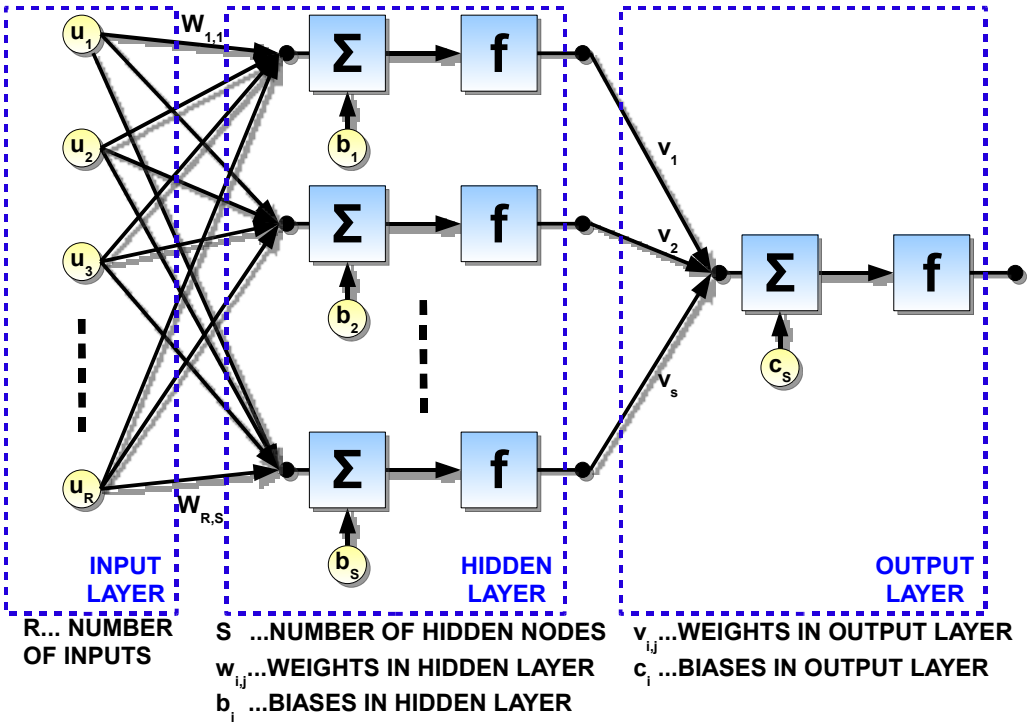


Figure 9: The structure of a feedforward multilayer perceptron neural network[64]

The node (artificial neuron or perceptron) is the basic element of the feedforward multilayer perceptron neural network[52,55,56]. It is presented in Figure 10. All the inputs into the node are multiplied by a particular weight $w_{i,R}$. The sum of all the weighted inputs and the bias b value is used as an input into activation function f, which must always be normalizable and differentiable. The use of several non-linear activation functions in feedforward networks guaranties the approximation of a non-linear function on the basis of the universal approximation theorem. In contrast with the use of linear activation functions only models of linear processes can be achieved[47]. Log-sigmoid and tan-sigmoid activation functions are the most common choice in practice, while the linear activation function is usually used in the output layer. The functions are presented in Figure 11.

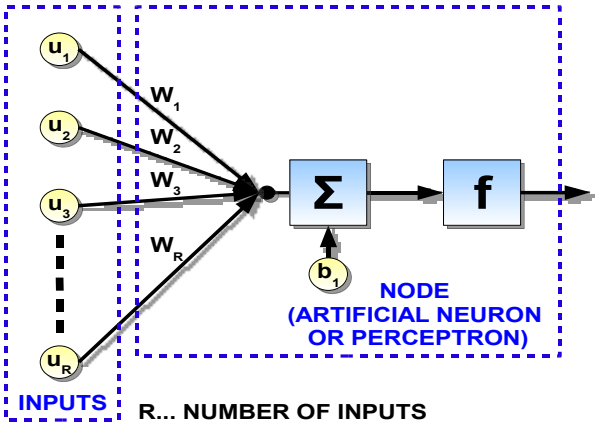


Figure 10: Node (artificial neuron or perceptron)[64]

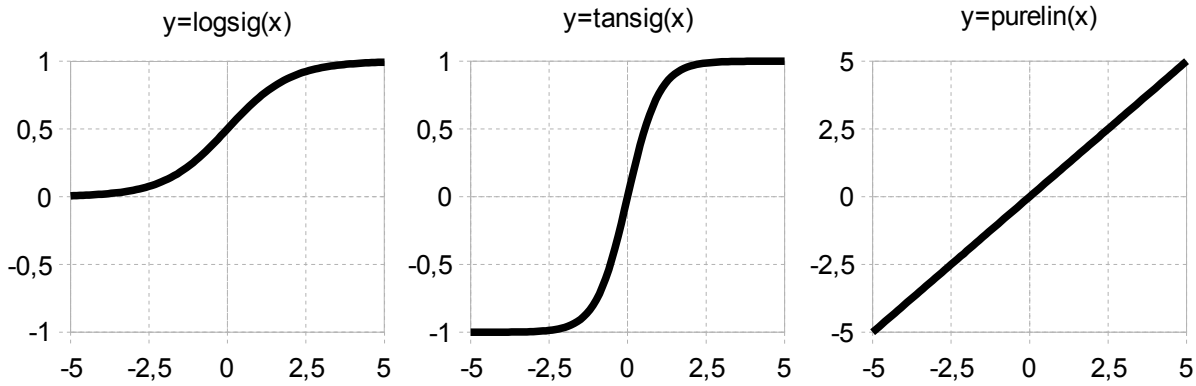


Figure 11: Most commonly used activation functions (right is log-sigmoid, middle is tan-sigmoid and left in linear)

A determination of the number of nodes in a hidden layer is one of the basic steps in creating a successful model based on a feedforward neural network. This issue has been addressed in the literature[65]. The number of nodes are usually determined with an iterative optimization procedure based on a selected cost function as a part of the neural network learning process. A possible alternative empirical approach to determining the number of hidden nodes is as follows. For the start, an equation (1.8) suggested by the *NeuroShell2*[66] neural networks modelling tool can be used for feedforward artificial neural networks with one hidden layer.

$$m = \frac{i+o}{2} + \sqrt{p} \tag{1.8}$$

m ...number of hidden nodes
i ...number of inputs
o ...number of outputs
p ...number of learning patterns

The number of hidden nodes determined by equation (1.8) depends on the number of data from the learning set. Usually, the learning set is very large, where a number of redundant data exist, and the result is an overestimated number of hidden nodes. To achieve satisfactory results the number of nodes must be decreased experimentally or learning data should be selected to avoid the redundant data. The presented equation is a relatively good and simple method to accomplish the goal, but during the modelling process we must be aware that the number of hidden nodes also depends on the complexity of the process and not only on the number of available data. In the case of too large a number of hidden nodes the feedforward neural network can overfit the learning data and fails to capture the true process that generates the data. Also with overfitting, the generalization ability of the feedforward neural network is smaller. In contrast, too strong a generalization combined with losing the ability to reproduce details can occur when not enough hidden nodes are used.

The feedforward neural network is ready for learning when its structure is determined; the

transfer function types, the number of hidden layers, the number of hidden nodes and the values of the weights and biases are initialized. In a learning process a set of learning data is used. Because each datum consists of an input-output pair the learning process is classified as a supervised learning paradigm. The learning data are introduced several times to the neural network during the learning process. At each iteration the weights and biases are changed until a certain input-output relation is accomplished. Learning the process is actually an iterative procedure, where the values of the weights and biases are set to minimize the cost function. A commonly used cost function is the mean-square error, which tries to minimize the average error between the network's output and the target value over all the example outputs of data.

A variety of learning techniques is available for a feedforward neural network where a back-propagation algorithm was the most commonly used[32,31,52,64,67] in the past. A back-propagation algorithm requires that the activation function used by the nodes (artificial neurons or perceptrons) is differentiable. There are two parameters, a learning rate and a momentum, that must be set before the back-propagation learning begins. The learning rate defines the rate of considering the calculated error during a current adjustment of the weight. And the momentum defines the measure of considering previous weight adjustments in the current weight adjustment. The consequence of the small learning rate is an unnecessary waste of computational time and strong increase in the possibility that the algorithm will be caught in local minimum. In contrast, a large learning rate can cause oscillations. Oscillations can also occur when too large or too small a value of momentum is used, where small values can also slow down the learning process.

A back-propagation algorithm is, for some practical purposes, often too slow. However, there are several high-performance algorithms that can converge from ten to one hundred times faster[64]. According to *Matlab's Neural Networks Toolbox* these faster algorithms can be classified into two categories. The first category is based on heuristic techniques, which were developed from an analysis of the performance of the standard steepest descent algorithm. It consist of two main heuristic techniques: *variable learning rate back propagation* and *resilient back propagation*. The second category of fast algorithms uses standard numerical optimization techniques[68]. It consists of three types of numerical optimization techniques for neural network training: *conjugate gradient algorithms*, *Quasi-Newton algorithms* and *Levenberg-Marquardt* algorithm.

For practical purposes a *Levenberg-Marquardt* optimization technique is used in this thesis. The optimization techniques are not an issue of this thesis, and therefore the detailed original description of the Levenberg-Marquardt algorithm is given in the paper by Marquardt[69]. The application of the Levenberg-Marquardt technique to neural network training is described in the paper by Hagan and Menhaj[70].

1.4.8. Clustering

Improvements of Lagrangian particle model proposed in this thesis are also based on clustering. Some basic theory of clustering is given in the following paragraphs for better illustration.

Clustering is defined as the classification of patterns into groups. The word patterns can be used to designate observations, data items or feature vectors, while the term groups denotes clusters. It has been used by researchers in many disciplines and addressed in many contexts because of its usefulness as one of the steps in exploratory data analysis[71]. A typical pattern-clustering activity involves following the steps[72] depicted in Figure 12.

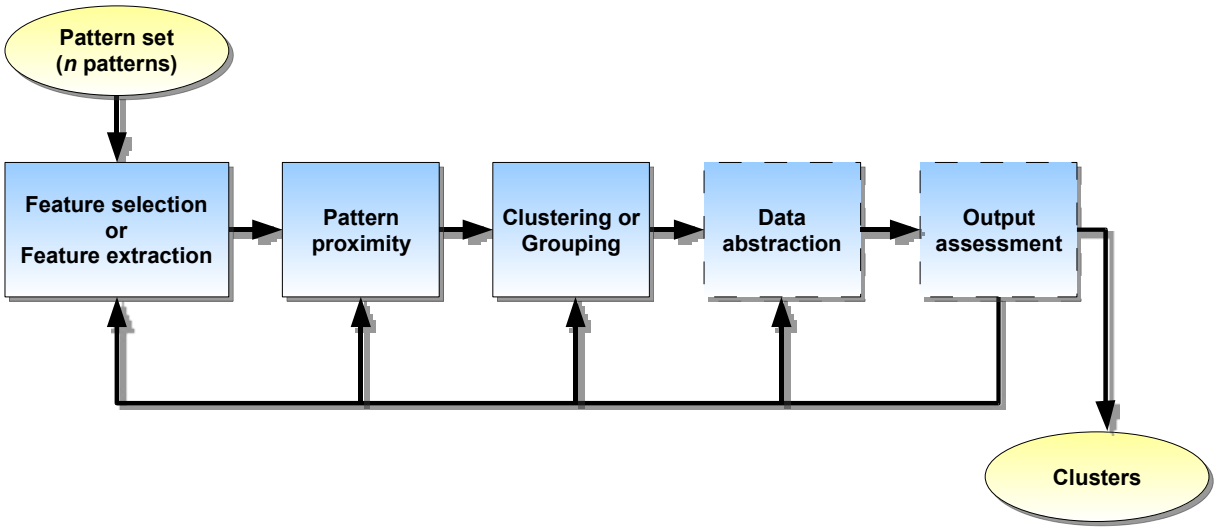


Figure 12: Stages in clustering[72]

Pattern representation refers to the number of classes, the number of available patterns, and the number, type and scale of the features available to the clustering algorithm. *Feature selection* is the process of identifying the most effective subset of the original features to use in clustering. *Feature extraction* is the use of one or more transformations of the input features to produce salient features. Feature extraction and/or selection are optionally included in pattern representation. Either or both of these techniques can be used to obtain an appropriate set of features in the process of clustering.

Pattern proximity or *inter-pattern similarity* is usually measured by a distance function defined on pairs of patterns. A variety of distance measures are in use. Distance measures are discussed in detail in the paper by Jain et al.[71].

Grouping or *clustering* is a stage that can be performed in a number of different ways. Different clustering techniques are described in the paper by Jain et al.[71]. According to their taxonomy there is, at the top level, a distinction between hierarchical (hierarchical algorithms find successive clusters using previously established clusters) and partitional (partitional algorithms determine all the clusters at once) approaches. Hierarchical approaches are additionally divided into two basic groups: *agglomerative* ("bottom-up") or *divisive* ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge

them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. The presented taxonomy is also supplemented by a discussion of cross-cutting issues that may affect all of the different approaches, regardless of their placement in the taxonomy as: agglomerative vs. divisive, monothetic vs. polythetic, hard vs. fuzzy, deterministic vs. stochastic and incremental vs. non-incremental.

Data abstraction is the process of creating a simple and compact data-set representation. Simplicity can be either human-oriented or defined from the perspective of an automatic analysis.

Output assessment refers to a cluster-validity analysis. Often this analysis uses a specific criterion of optimality but these criteria are usually arrived at subjectively. Usually a rule[71] is followed that validity assessments are objective and are performed to determine whether the output is meaningful. There are three types of validation studies:

- *an external assessment* of validity compares the recovered structure to an a priori known structure,
- *an internal validation* tries to determine if the structure is intrinsically appropriate for the data,
- *a relative test* compares two structures and measures their relative merit.

The vast collection of available clustering algorithms in the literature can be very confusing for users wanting to select an algorithm suitable for the problem at hand. According to the directions in the paper by Jain et al.[71] the following clustering methods are selected for the problem of the clustering of a large amount of patterns into a relatively large number of clusters: K-MEANS clustering algorithm, and self-organizing map (SOM - Kohonen artificial neural network). Both methods are described in detail in the review by Jain et al.[71]. The concept of both methods differs significantly: while the K-MEANS clustering is performed only in the input space, the SOM clustering is performed by transforming the data from high-dimensional input space into low-dimensional map space (usually two dimensional). Summarized descriptions of both methods are presented in the following paragraphs.

K-MEANS clustering algorithm

K-MEANS is one of the simplest unsupervised learning algorithms used for clustering n patterns into k clusters[73]. The algorithm follows a simple way to classify a given data set through a certain a priori fixed number of k clusters. It begins by arbitrarily setting up k centroids (cluster centres), one for each cluster, inside the hyper-volume containing the pattern set. When initial centroids are defined, each pattern is assigned to the closest centroid (cluster centre). For each obtained cluster new centroids are recalculated using the current cluster memberships. After that each pattern is again assigned to the newly obtained closest centroid (cluster centre). This procedure of assigning the patterns to centroids and recalculating new centroids is repeated until a convergence criterion is met. There are two

typical convergence criteria[71]: no (or minimal) reassignment of the patterns to new centroids or a minimal decrease in the squared error where the algorithm aims at minimizing the objective function defined in equation (1.9).

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \tag{1.9}$$

k ...number of centroids (cluster centres)
n ...number of patterns
$x_i^{(j)}$ ...pattern assigned to j cluster
c_j ...centroid (cluster centre)
$\| x_i^{(j)} - c_j \|$ ...chosen distance measure between a data point and the centroid

The algorithm is presented in Figure 13. Although the procedure always terminates, the *K-MEANS* clustering algorithm does not necessarily find the most optimal configuration that corresponds to the global objective function minimum. It is also significantly sensitive to the initial, randomly selected centroids (cluster centres). This effect can be reduced by performing it multiple times[71].

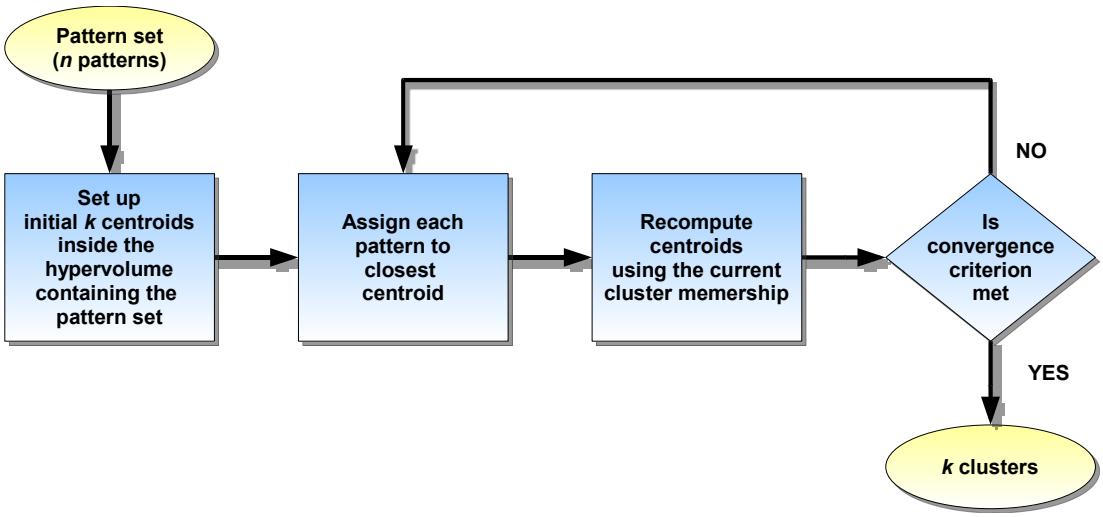


Figure 13: K-MEANS clustering algorithm[73]

An illustration of the clustering procedure is presented in Figure 14 using a very simple two-dimensional (2D) example. The positions of the centroids are presented using a rhombohedral mark and the positions of the patterns, using a dot mark.

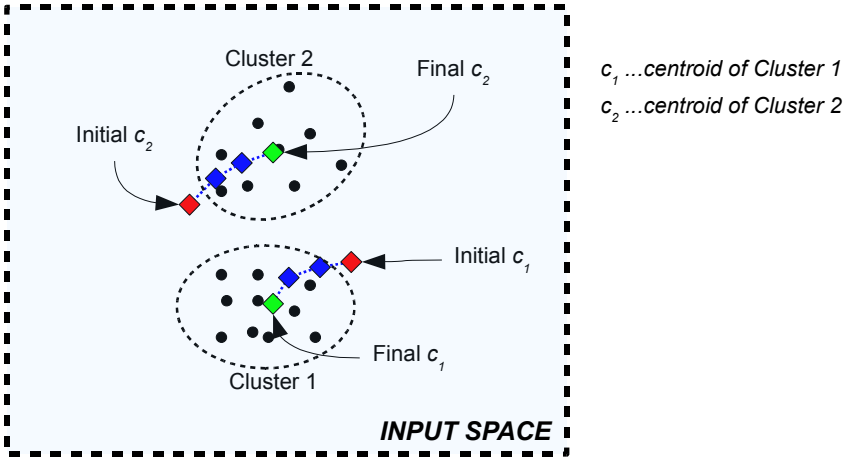


Figure 14: Simple illustration of the K-MEANS algorithm where random initial centroids are moving towards final centroids of clusters until the convergence criterion is met[73]

The SOM clustering algorithm

A self-organizing map (SOM) is a computational method for the visualization and analysis of high-dimensional data[71]. It is also known as a *Kohonen map* because it was first described as an artificial neural network by professor Teuvo Kohonen[74].

A SOM is a type of artificial neural network. It is trained using unsupervised learning to produce a low-dimensional (typically two dimensional), discretized representation of the input space of the training samples, called a map. The obtained map preserves the topological properties of the input space. The SOM is operating in two modes, like most artificial neural networks:

- *training* builds the map using input patterns in a competitive process, also called *vector quantization*;
- *mapping* automatically classifies new patterns that are introduced to the SOM according to the map created in the training process.

The basic components of a SOM are the nodes (or neurons). Each node is associated with a position in the map space and a weight vector. A weight vector is of the same dimension as the input data. Nodes are usually arranged in a regularly spaced hexagonal or rectangular grid map. The SOM describes the transformation from a higher-dimensional input space to a lower-dimensional map space: a vector from the input data set is placed on the map. It is a process of finding the node with the closest weight vector to the vector taken from the data space and to assigning the map coordinates of this node to our vector.

The main goal of learning in the SOM is to ensure that different parts of the network are responding similarly to certain input patterns. Before the training process begins the weight vectors of the nodes are usually initialized to small random values. During the training process a large number of input patterns that are as much as possible like the patterns that are

expected during mapping must be presented to the network.

The training is a form of competitive learning[74]. When a training example is presented to the network a distance measure to all the weight vectors is computed. The node that contains the weight vector closest to the input pattern is determined as the best matching unit (BMU). The weights of the best matching unit and the nodes that are close to it in the SOM grid are adjusted towards the vector of the input pattern. The magnitude of the adjustment decreases with time and distance from the best matching unit. A node weight vector is updated according to equation (1.10*)*.

$$\boldsymbol{W}_j(t+1)=\boldsymbol{W}_j(t)+\theta(j,t)\alpha(t)(\boldsymbol{x}_i-\boldsymbol{W}_j(t));\ j=1..k,\ i=1..n,\ t=1..m \tag{1.10}$$

k ...number of centroids (cluster centres)
n ...number of patterns
m ...number of iterations
$\boldsymbol{W}_j$ ...weight vector of node
$\boldsymbol{x}_i$ ...vector of input pattern
$\theta(j,t)$ ...neighbourhood function
$\alpha(t)$ ...monotonically decreasing learning coefficient

The neighbourhood function $\Theta(j, t)$ is dependent on the distance measure between the best matching unit and the node j. In the most simple form it is one for all the neurons close enough to the best matching unit and zero for the others. A very common choice for the neighbourhood function is the Gaussian function. Another important property of the the neighbourhood function is that it shrinks with time, regardless of the functional form. At the beginning when the neighbourhood is broad, the self-organizing is performed on the global scale. After a while when the neighbourhood shrinks to just a couple of neurons the weights converge to local estimates. This self-organizing process is repeated for all the input patterns n for the number of iterations m.

During the mapping process, there is only one *winning* node. The winning node represents the neuron whose weight vector lies closest to the input vector according to the selected distance measure. It is very simply determined by calculating the distance measure between the vector of the input pattern and the weight vector of each node.

Clustering based on the SOM is performed in two main steps, as presented in Figure 15: in the first *training* step a map (usually two-dimensional) of nodes (clusters) is determined and in the second *mapping* step each input pattern is assigned to one *winning* node (cluster).

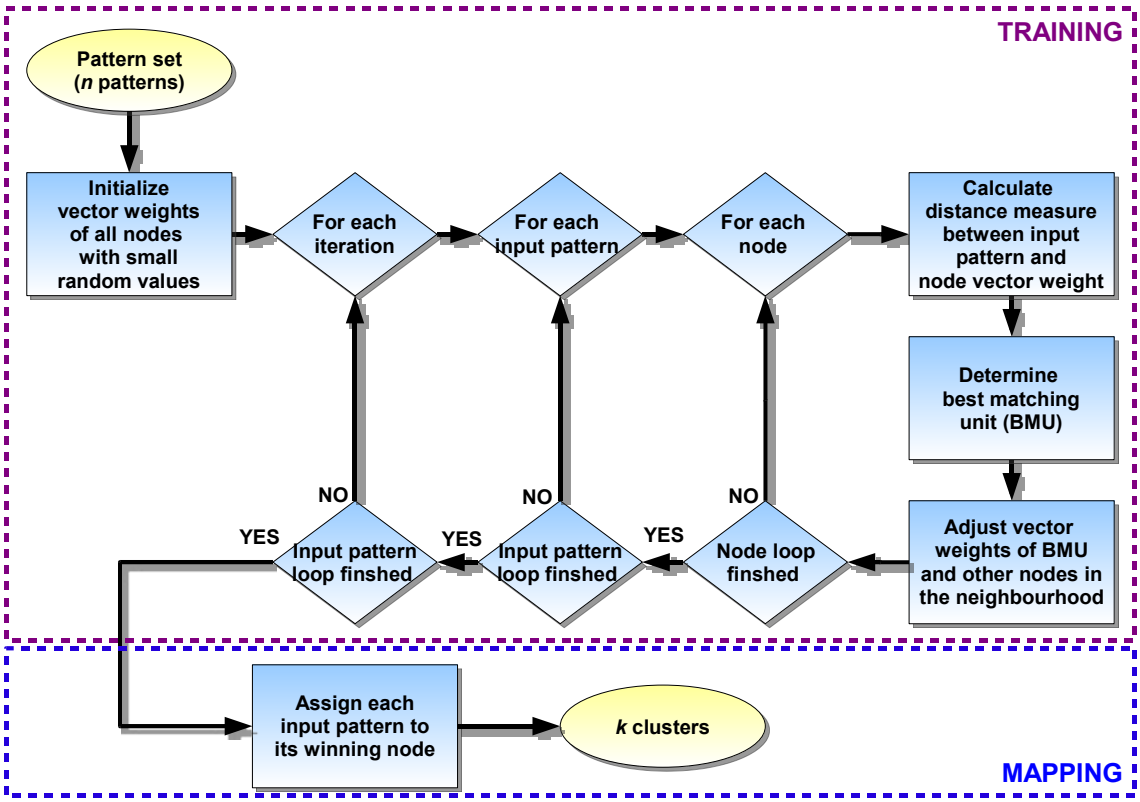


Figure 15: The SOM clustering algorithm[74]

An illustration of the SOM clustering algorithm is presented in Figure 16 using a very simple two-dimensional (2D) example. The thin-dotted lines are used to show how each point from the input space is assigned to its best matching unit (BMU) from the map space.

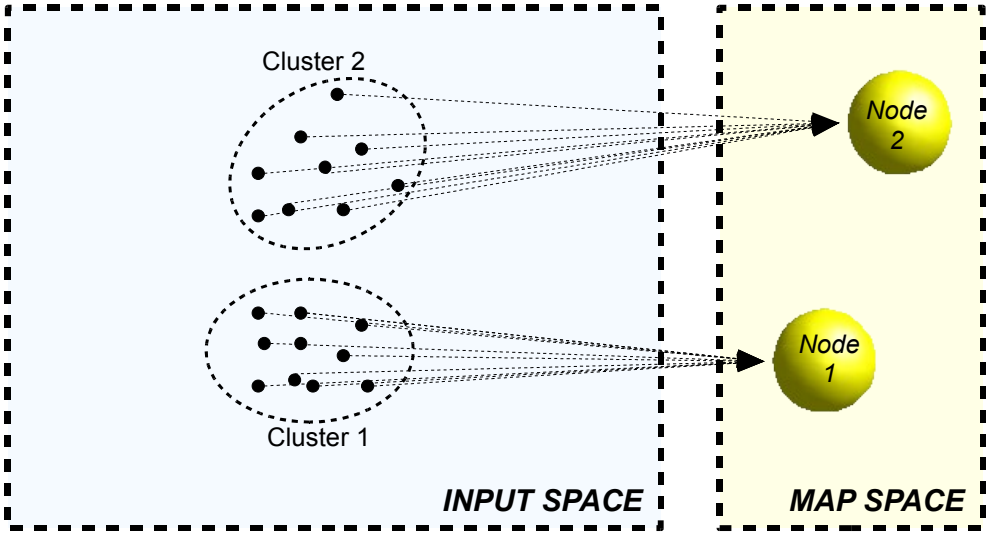


Figure 16: Simple illustration of the SOM algorithm where each pattern is assigned to its winning node in the map space[74]

1.5. Relevant studies

1.5.1. Reason for using the Lagrangian particle-dispersion model

The majority of Slovenian air-pollution facilities are located at the bottom of basins or river canyons (i.e., Thermal Power Plant Šoštanj in the Šoštanj basin, Thermal Power Plant Trbovlje and the Cement Factory in the Zasavje region, industrial facilities in the Ljubljana basin, etc.) surrounded by hills or mountains. On the basis of experiences from studies presented in the following paragraph air-pollution modelling based on Lagrangian particle dispersion is better than other dispersion-modelling techniques. It can provide enough complete, realistic and satisfactory results for complex terrain on a small-scale area of interest. The specific definition of a complex terrain is given in the literature[36,35,39].

Various review papers on air-pollution modelling and their approaches to dispersion in different environments are available[75,76,77]. Studies were done to compare different modelling techniques according to each other and to measurements. Comparisons were done only with some of the available modelling techniques, but none of them was done with all the techniques that were available at that moment. Only some of the studies will be highlighted according to the area of interest and their major conclusions will be presented:

- evaluations over simple terrain: Many evaluations[15,78,79,80] of different modelling techniques used for regulatory purposes over simple terrain were performed using the well-established and documented *Model Validation Kit* that has been used for a series of workshops and conferences on Harmonisation within Atmospheric Dispersion Modelling for Regulatory purposes[81]. The *Model Validation Kit* addresses the classic problem of a single stack emitting a non-reactive gas. It comprises data from the following four field experiments performed on simple terrain and controlled environment: Kincaid experiment (1980-81), Copenhagen experiment (1978-79), Lillestrøm experiment (1987) and the Indianapolis experiment (1985).
- qualitative evaluations over complex terrain: Some model evaluations for regulatory purposes were performed for complex terrain where the emphasis was on a qualitative evaluation rather than statistical analysis like: the evaluation of the RIMPUFF model over complex terrain in Northern Spain by Thykier-Nielsen et al.[16], the evaluation of the SCIPUFF model over complex terrain in New Mexico by Cox et al.[17] and the evaluation of the three models LASAT, ADMS and ONGAUSSplus over complex terrain of Zasavje region in Slovenia by Hirtl et al.[18].
- evaluation over complex terrain in Slovenia: In 1991 an experimental measuring campaign was performed around the largest Slovenian thermal power plant, Šoštanj[82]. Four air-pollution models were compared using the measuring campaign data in a paper presented by Božnar et al.[37]: Gaussian model, Gaussian-hybrid model for complex terrain, a Gaussian puff model and the Lagrangian particle model. The reconstructed concentrations from the Gaussian model for complex terrain in stable meteorological conditions were larger than the measured ones. Similar, but only slightly better results, were given by the Gaussian puff model because the orography

was also considered as an input to this type of model. The Gaussian puff model represents a link between the Gaussian models and the more sophisticated three-dimensional (3D) models. It also uses the same 3D wind field as the Lagrangian model. The obtained results were slightly better than with the Gaussian models and the authors suggested that it can be used for a quick analysis of simple pollution cases in a complex terrain. The best results were obtained with the Lagrangian 3D model, which proved to be a useful tool for the reconstruction of pollution episodes in very complex terrain and in situations with low winds or convective mixing.

- evaluation over complex terrain in north-western Spain: A comparison of the operational Lagrangian particle and adaptive puff models for plume dispersion forecasting was done by Souto et al. [83]. It was performed around thermal power plant As Pontes across the area characterized by steep hills and sea inlets bathed by the Atlantic Ocean. The transport and the dispersion of pollutants in the lower atmosphere was predicted by using both a Lagrangian particle model and an adaptive puff model coupled to the same mesoscale meteorological prediction model. The results of both models in forecasting the SO_2 ground-level concentration around a coal-fired power plant were compared under unstable meteorological conditions. In addition, meteorological and SO_2 ground-level concentration numerical results were compared to field measurements provided by 17 fully automated SO_2 ground-level concentration remote stations, nine meteorological towers and one SODAR, from a meteorological and air-quality monitoring network located in a circle 30km around the power plant. The main issue of this comparison was focused on the models' accuracy, from selected air-pollution episodes. With the same meteorological input, during the episodes considered, the Lagrangian particle model showed a better agreement with the ground-level concentration measurements than the adaptive puff model, especially when the plume impacts are sparse. Therefore, the Lagrangian particle model was shown to be more suitable to be applied for an operational air-pollution forecast in this environment. The authors concluded their work with an additional advantage of the continuous growth of computers' performance. This advantage will enable the consideration of more complex solutions in the Lagrangian particle model in the future.

- evaluation over complex terrain in Italy: Another comparison of the Gaussian model and the Lagrangian particle model for regulatory applications in flat and complex terrain in two Italian coastal sites had been made by Brusasca et al.[84]. The first site was located near to Venice and regarded two thermal power plants (TPP) Fusina and Porto Marghera. It is characterised by simple terrain, the land use nearby the emissions is mainly of industrial or urban type and agricultural inland. The second site was located around TPP in Vado Ligure, on the northern Mediterranean coast of Italy. It is characterised by very complex terrain: the climatology is dominated by the superposition of land/sea breezes and slope flows. The inter-comparison of both models on simple simple terrain characterised by weak space/time variations of wind

speed and direction gave congruous results. The areas of interest for pollutant impact are similar, but the relevant differences were detected in pollutant patterns for the different statistical indexes prescribed by air-quality standards. In the complex terrain and turbulent conditions the results obtained by both modelling approaches were completely dissimilar. According to the authors, steady-state models, like the Gaussian model, that implement simple algorithms for complex terrain appear to be unable to reproduce the features of pollutant dispersion in these kinds of conditions.

1.5.2. Reason for optimising the computational efficiency of the Lagrangian particle-dispersion model and simulation

The excessive computational demand of Lagrangian particle-dispersion models represents their major disadvantage. With the constant quality improvements of models, their complexity has gradually increased, which leads to enormous requirements for computational time. Not much work has been published so far on the aspect of saving computational time in comparison to the large amount of work that has been published on the quality of the model's results. In order to speed up simulations based on Lagrangian particle-dispersion models some work has already been done in the field of simplifying the underlying physics, for example, by Hurley and Physick[85], Ryall and Maryon[86], Ermak and Nasstrom[87]. Another approach to simplifying the computational demands was proposed by Melheim[88]. The clustering method for different purposes was presented in his paper[88] to reduce the number of equations to be solved in collisional particle dynamics in the Lagrangian framework. Its main idea is that only particles that interact or may interact during the next global time-step are integrated simultaneously, while particles that are far from other particles can be integrated alone. The clusters of particles were made of particles that interact or may interact during the next global time-step. The cluster-integration method was applied to the sedimentation of particles in a two-dimensional box.

In this thesis the focus will be on analysing the efficiency of the numerical procedures involved in simulation procedures rather than analysing the underlying physics. Similar goals had been set in the contribution by Schwere et al.,[10] where some specific tasks that waste computing time were outlined and three methods to reduce the computing time of these tasks were presented. In this thesis additional supplementary methods will be proposed. The authors focused on analysing and improving the efficiency of the numerical procedures involved in calculating the particle trajectories, which can be roughly split into three major parts: the calculation of the particle's acceleration, the generation of the random numbers, and the evaluation of a time-step criterion. It was shown that a speed-up factor of two can be reached for a more sophisticated model, which takes into account the skewed and inhomogeneous turbulence. The results were compared with the original model's results and it was shown that the speed-up procedures do not significantly alter the concentration patterns.

1.5.3. Reason for using black-box modelling techniques

In this thesis a proposal to use black-box modelling techniques to optimize the computational efficiency of the Lagrangian particle-dispersion models will be made. Similar work has already been done in the field of meteorological numerical models by Krasnopolsky and Chevallier[89]. The computational efficiency improvement of certain processes in environmental models was applied based on the application of generic neural networks. An approach was used to accelerate the calculations and improve the accuracy of the parametrization of several types of physical processes that generally require computations involving complex mathematical expressions, including differential and integral equations, rules, restrictions and highly non-linear empirical relations based on physical or statistical models. It has been shown that neural networks can be used to replace primary parametrization schemes because comparable input-output response was achieved as with different mathematical expressions that can be considered as an approximation. The research work was supported by the presentation of four, particular, real-life applications where the neural network approach was used.

Krasnopolsky et al.[90] also presented their new approach based on a synergistic combination of white-box and black-box modelling within atmospheric models. It was applied for the development of an accurate and fast approximation of an atmospheric long-wave radiation parametrization for the NCAR Community Atmospheric Model, which is the most computationally consuming component of the model's physics. The authors reported a computational expense improvement of the neural-network-based simulation by two orders of magnitude (50-80 times faster than the original parametrization). A development framework and practical validation criteria for the neural network emulations of the model physics components were outlined.

1.6. Outline of the dissertation

The dissertation begins with the introduction, where the research problems are defined and the purpose of the research and the working hypotheses are given. The introduction proceeds with the presentation of the theoretical background and ends with a summarized presentation of the relevant studies where the arguments for using the Lagrangian particle-dispersion model and for optimizing its computational efficiency based on black-box modelling techniques are given.

In section 2 two field data sets used for the experiments in this thesis are described. Both were obtained during two measuring campaigns performed in Slovenia over local scale areas of interest where complex terrain conditions are present. The Šaleška region field data set is selected to evaluate the computational efficiency of the computer an air-pollution model based on the Lagrangian particle-dispersion model and to evaluate new methods for the efficiency improvement. Improvements to the computer air-pollution model are additionally evaluated on the Zasavje region field data set to demonstrate that the proposed methods to improve the Lagrangian particle-dispersion model can be generally used over complex terrain.

The air-pollution modelling methodology based on the Lagrangian particle dispersion is described in section 3, where the computational efficiency and the performances of the computer air-pollution model based on Lagrangian particle dispersion are also evaluated. The main components of computer air-pollution model are presented and discussed. Several evaluation methods are developed and used in the evaluation of the performance of computer air-pollution model according to the particle number density coefficient (*PDNC*), which is also presented and defined in section 3. The presentation of the air-pollution modelling methodology ends with a demonstration of a computational problem that can occur during different operational purposes of the computer air-pollution model.

Three new methods to improve the computational efficiency of the computer air-pollution model based on Lagrangian particle dispersion are proposed in section 4, where the emphasis is on preserving the quality of the results.

The first presented is the clustering contributed to decrease the computational cost by decreasing the number of particles in the simulations. A concentration-estimation method based on kernel density is adapted to substitute the box-counting concentration estimation method and to improve the poor quality of the results when a smaller number of particles are used in the simulations. To keep the number of particles in the simulations at a constant value a third Lagrangian particle-dispersion control method is developed, which consists of the two main subsequent methods. In the first step the percentage of lost particles is predicted with the use of an artificial neural network based on meteorology, emission and the initial situation of the air pollution. In the second step the clustering parameters are determined by a decision-making method. The performances of all the developed methods are validated on the Šaleška region field data set.

In section 5 all the proposed methods are mutually integrated into a new, enhanced, Lagrangian particle model. After the presentation of the integration, the performance of the

enhanced Lagrangian particle computer model is validated on the Šaleška region field data set.

The performance of the enhanced Lagrangian particle-dispersion computer model with new, integrated methods is validated on a Zasavje region field data set in section 6. The validation is performed to confirm that the developed methods can be generally applied in various complex terrains. Before the final evaluation, the adjustments of the parameters of the developed methods according to the properties of the terrain are also presented.

The thesis ends with the final section 7 where the conclusions and recommendations are given. In the conclusions the results of the study and the development are summarized and evaluated. The goals and hypotheses are discussed and critically evaluated. Finally, the recommendations about the possibilities to additionally improve the air-pollution modelling methodology based on Lagrangian particle dispersion that occurred during the research and will be used in further developments and research.

2. FIELD DATA SETS

In the following sections of this thesis an air-pollution modelling methodology based on Lagrangian particle dispersion is presented, and later the methods to improve its computational efficiency are proposed and tested. All the experiments are performed on two experimental field data sets obtained during two measuring campaigns. Both measuring campaigns were performed in Slovenia over local scale areas of interest where complex terrain conditions are present. To evaluate the computational efficiency of the Lagrangian particle-dispersion (LPD) computer model and to validate new methods for an efficiency improvement, a field data set from the Šaleška region was selected. To demonstrate that the proposed methods can be generally used over complex terrain, an enhanced Lagrangian particle-dispersion (ELPD) computer model is validated on another field data set from the Zasavje region.

2.1. The Šaleška region field data set

In the following section *4.Proposed improvements in air pollution modelling methodology* the contribution of proposed methods to improving the computational efficiency of the LPD model is presented. During the development process the field data set is used for the development of new methods and to compare the results of the original LPD model and the results obtained with the newly proposed methods. To ensure that the methods can be generally used the selected field data set must spread over the area where all the complex terrain meteorological conditions occur.

The Šaleška region was selected as a field data set for several reasons:

- it spreads over complex orography (basin surrounded by high hills and a semi-mountainous continuation of the Karavanke Alps) where almost all possible complex terrain conditions occur,
- the database of ambient measurements is available from the measuring campaign organized in spring 1991
- the emissions from the three stacks of thermal power plant Šoštanj represent the main air-pollution source in the region because desulphurization facilities were not yet installed during the measuring campaign performed in spring 1991.

2.1.1. Terrain description

The area of the investigation extends across the Šaleška valley, which is situated in the north-eastern part of Slovenia, as presented on the left side of Figure 17[91]. The central part of the Šaleška valley is a plain north of the river Paka, with an average altitude of three hundred metres above sea level. The basin is surrounded by isolated hills on the south, and by the semi-mountainous continuation of the Karavanke Alps on the west, north and east, as presented on the right side of Figure 17. Two small towns are located in the basin: Šoštanj had approximately 3 000 inhabitants, Velenje about 24 000 and about 9 000 people lived in the villages around [82,92] during the measuring campaign organized in spring 1991.

Due to the complex orography, as presented on right side of Figure 17, and unfavourable climatic conditions (thermal inversion in the basin), very high concentrations of SO_2 occur, especially during the winter. The Šaleška valley represents a typical case of complex terrain[36] by all the criteria given in the subsection *1.4.5.Complex terrain*. There are certain air-pollution situations that lead to increased concentrations in the complex terrain: plume impingement on high terrain depicted in Figure 18, pooling in the valleys depicted in Figure 19, drainage towards the population centres depicted in Figure 20 and persistence due to channelling depicted in Figure 21. On the left side of the presented figures the complex terrain phenomenon is illustrated, in the middle the three-dimensional presentation of a phenomenon that occurred over the area is encircled and on the right side the impact of the phenomenon at the ground-level concentration is presented. The Veliki Vrh hill located to the south of the power plant is the nearest high-terrain obstacle. The plume impingement on the hill is encircled in the middle of Figure 18 and an increased ground-level concentration caused by the phenomenon is shown on the right side. In Figure 19 the pooling in the valleys between isolated hills on the south of the domain is presented, where on the right side of the figure the increased ground-level concentration in the valley far from the power plant is presented. Figure 20 represents the drainage of the air pollution towards the town of Velenje, where on the right side of the figure an increased ground-level concentration is present, downwind from the power plant towards the town. The persistence of the increased ground-level concentration due to channelling caused by the topography (a chain of hills) on the south-west of the Šaleška valley is presented on the right side of Figure 21.

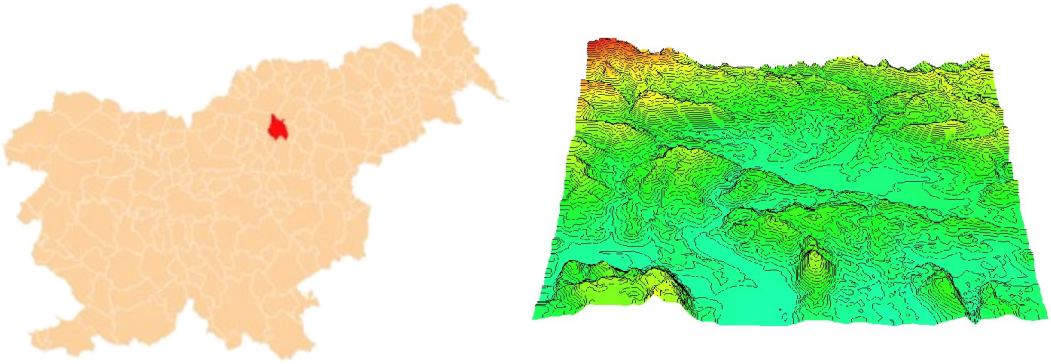


Figure 17: Location of the Šaleška region in Slovenia on the left[91] and the topography of the region on the right

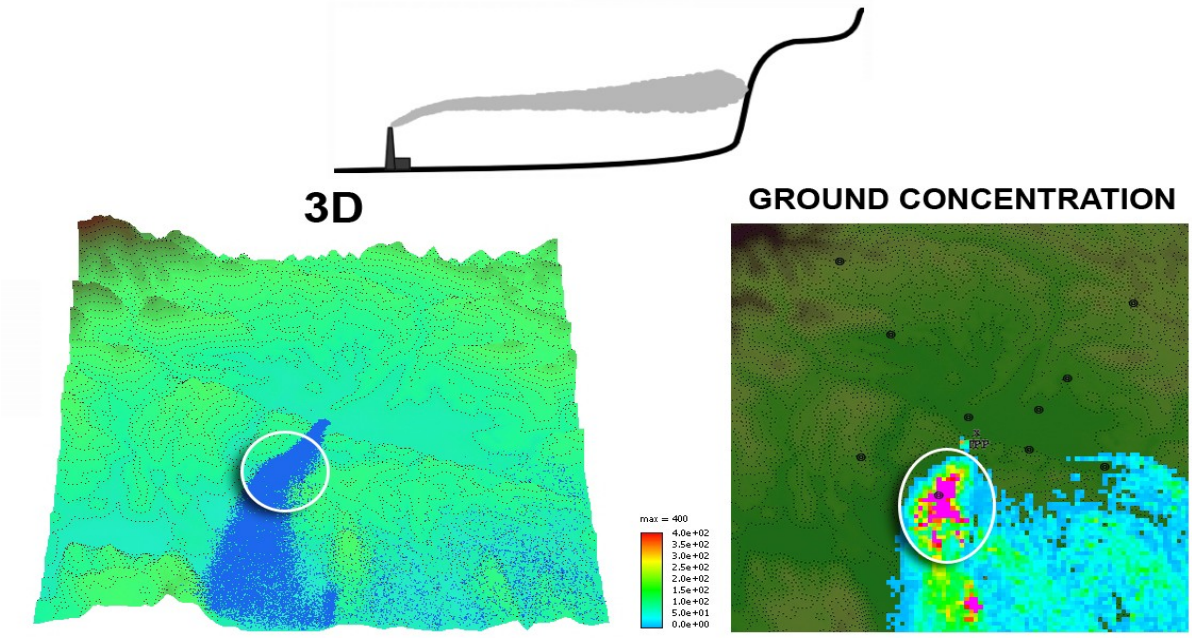


Figure 18: Plume impingement on high terrain – an example from the Šaleška region

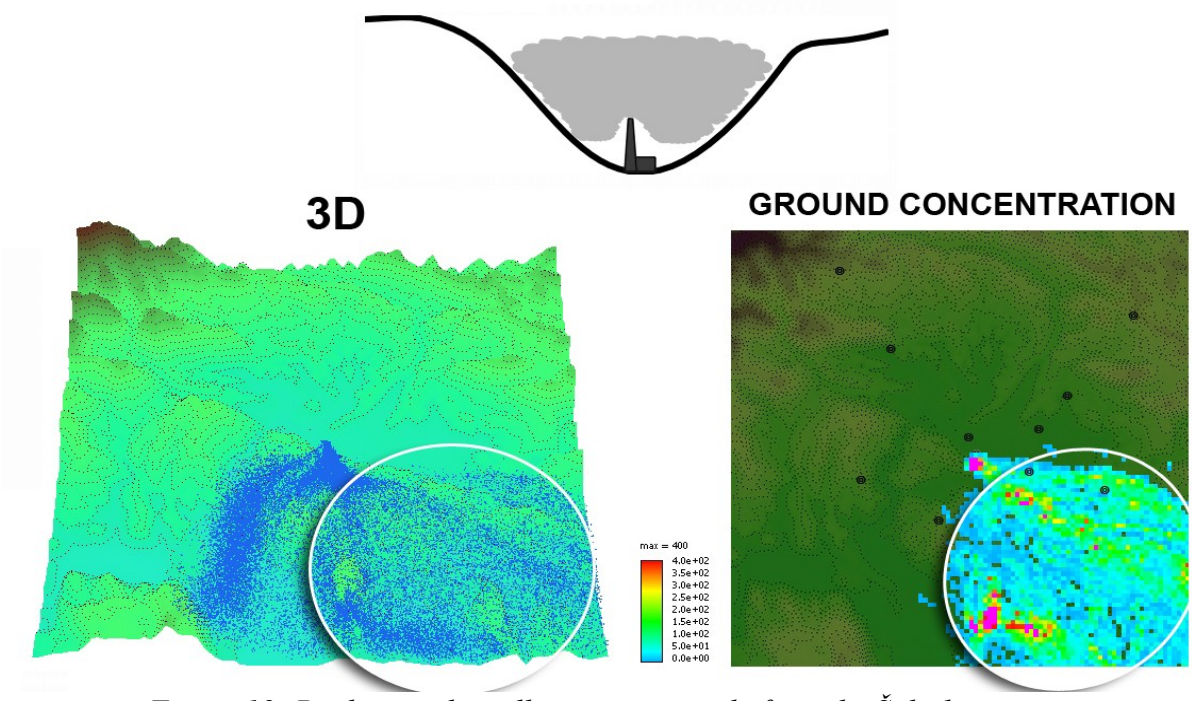


Figure 19: Pooling in the valleys – an example from the Šaleška region

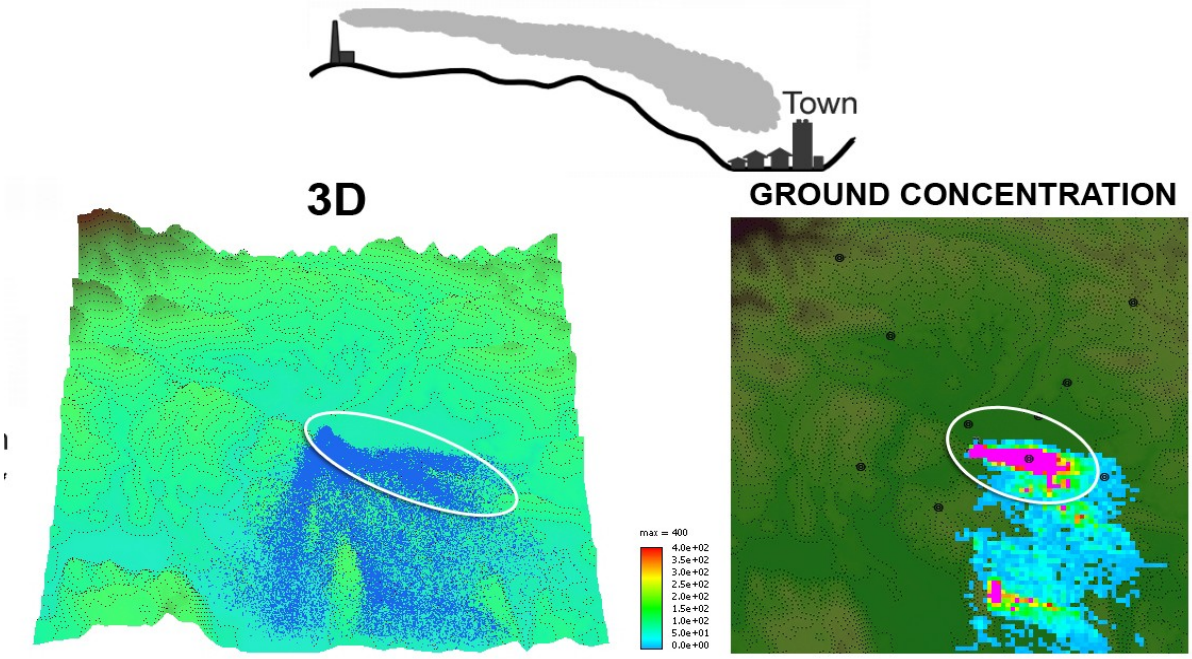


Figure 20: Drainage toward population centres – an example from the Šaleška region

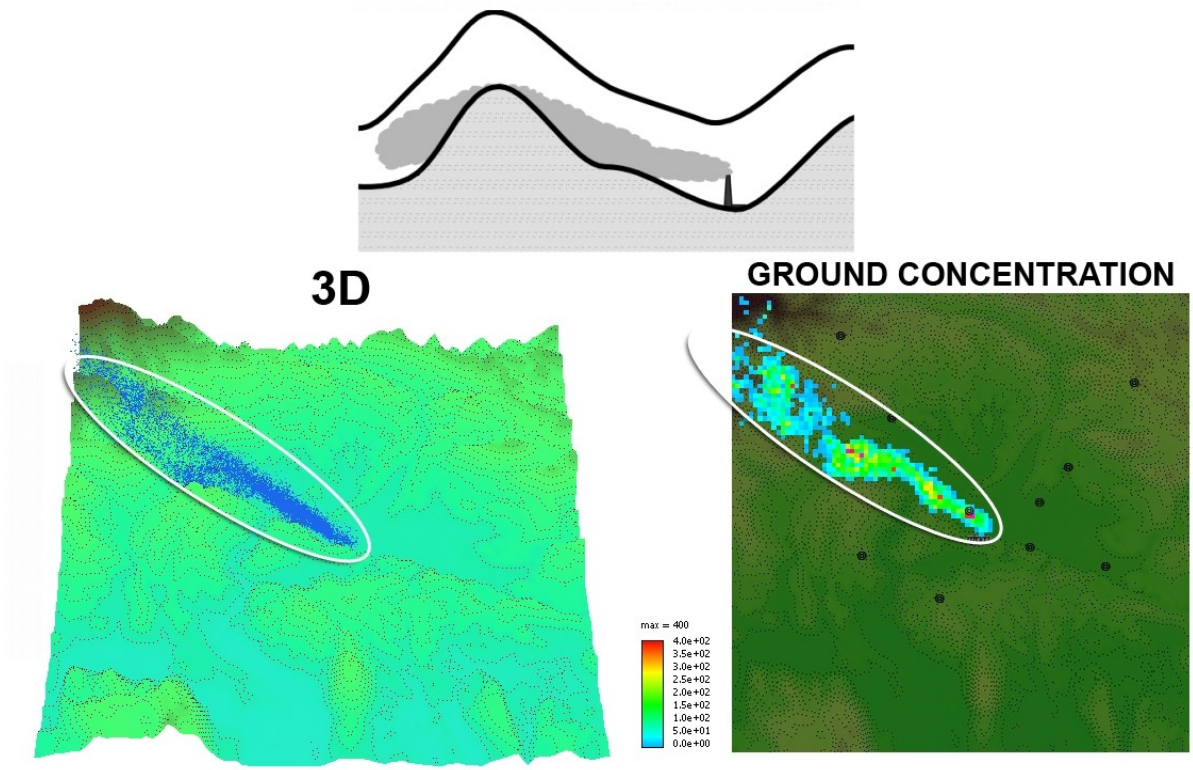


Figure 21: Persistence due to channelling – an example from the Šaleška region

2.1.2. Experimental measuring campaign

To evaluate the performance of the selected Lagrangian particle-dispersion model, data from a measuring campaign performed from 15th of March to 5th of April 1991 across the Šaleška region is used. The measuring campaign was performed as the joint effort of three institutions: ENEL-CRAM and CISE from Milano in Italy and the Jozef Stefan Institute from Ljubljana in Slovenia. The database from the measuring campaign was published and distributed on floppy disks in order to be available for further processing and research. The contents of the floppy disks are available as part of the final report[82,92]. In this campaign, organised during the spring of 1991[82,92], the concentrations of SO_2 higher than 1 mg/m^3 were measured at surrounding stations and all of them were caused by the high emissions from the three stacks of a thermal power plant that did not have desulphurization plants installed at that time. The data obtained during that measuring campaign can thus be used as a tracer experiment, because all other local sources of emission can be neglected. The database was constructed from different measurement sources, like the Environmental Informational System (EIS) of the Šoštanj TPP, one mobile Doppler SODAR, DIAL and an automatic mobile laboratory.

2.1.3. Sources of emission

The main source of air pollution in the Šaleška valley is the Šoštanj thermal power plant (TPP), presented in Figure 22. It is located in the centre of the domain and it has three stacks that are 100 m, 150 m and 230 m high. During the measuring campaign in 1991 the desulphurization facilities were not yet installed. It was estimated that emissions were about 100 000 tons of SO_2 and 12 400 tons of NO_X per year [38,82,92]. The power plant's pollutant emission concentration and the flux data, smoke temperature and exit velocity were measured by the emission station[82,92]. For evaluating the results of the simulation runs in following section only the emission of SO_2 from the 100-m stack (named *Stack123*) is used as the source of emissions.

Figure 22: Thermal power plant Šoštanj

2.1.4. Automatic environmental measuring system

The EIS (Environmental Information System) of Šoštanj TPP consisted of six stationary, automated, environmental measuring stations, located around the power plant, as presented in Figure 23. At all the stations the wind velocity and direction, the air temperature, the relative humidity and the SO_2 concentrations were measured. At some of the stations, other parameters were also measured, such as the global solar radiation, the precipitation, the air pressure and the other pollutant concentrations (NO_X and O_3).

The measuring parameters monitored at each automatic measuring stations are presented in Table 1. The real measured meteorological data from the measuring system will be used for the air-dispersion simulations.

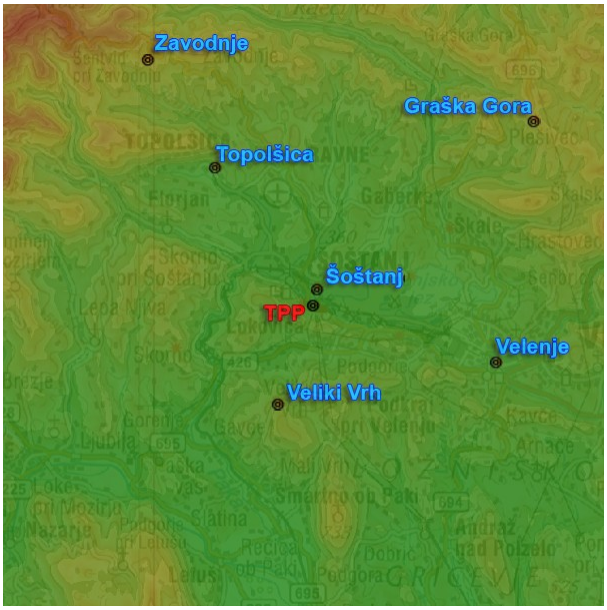


Figure 23: Locations of the automatic measuring stations across the Šaleška region

Table 1: Measuring parameters monitored across the Šaleška region

	Zavodnje	Graška Gora	Topolšica	Veliki Vrh
Air temperature	x	x	x	x
Relative humidity	x	x	x	x
Global solar radiation				
Precipitation				
Air pressure				
Wind	x	x	x	x
SO2	x	x	x	x
NO	x			
NOx	x			
O3	x			

	Šoštanj	Velenje
Air temperature	x	x
Relative humidity	x	x
Global solar radiation		x
Precipitation	x	
Air pressure	x	
Wind	x	x
SO2	x	x
NO		
NOx		
O3		

Figure 24: Ambient automatic measuring stations Šoštanj (upper-left) and Zavodje (upper-right), SODAR (lower-left) and DIAL (lower-right)[82]

2.1.5. Situation selection from the Šaleška region field data set

A database of measurements is available for the full duration of the measuring campaign, that is from 15th of March until 5th of April 1991. A situation that lasted from 1st of April 1991 at 20:00 until 2nd of April 1991 at 20:00 was selected, and it will be outlined because of its complexity. This complexity makes it very difficult to reconstruct and represents the greatest challenge to all air-pollution reconstruction modelling techniques. It will be used for the development and testing of new methods to improve the computational efficiency of the Lagrangian modelling methodology in the following section.

SODAR measurements performed during the selected period and presented in Figure 25 reveal that the wind speeds were very low and the wind changed course in all directions very rapidly. The arrows on the Figure 25 represent horizontal wind component at different heights above the SODAR. The three dimensional behaviour of the reconstructed plume presented in Figure 26 illustrates that the plume spread in all directions over a short period of time. This could also be supported by the SODAR measurements, which are at each presented reconstructed interval in Figure 26 consistent with the measurements in Figure 25. The reconstructed intervals were chosen to present the spreading of the air pollution in all possible directions. At the beginning of the selected interval, short calm meteorological conditions

occurred, which caused the accumulation of air pollution in the domain. The air pollution was accumulating during the whole night and morning. The accumulated air pollution was diluted in the afternoon, when the wind from the east strengthened. For the reconstruction, powerful computational resources were required. This complexity will be used for the development of new methods to improve the computational efficiency of the LPD model presented in the following sections.

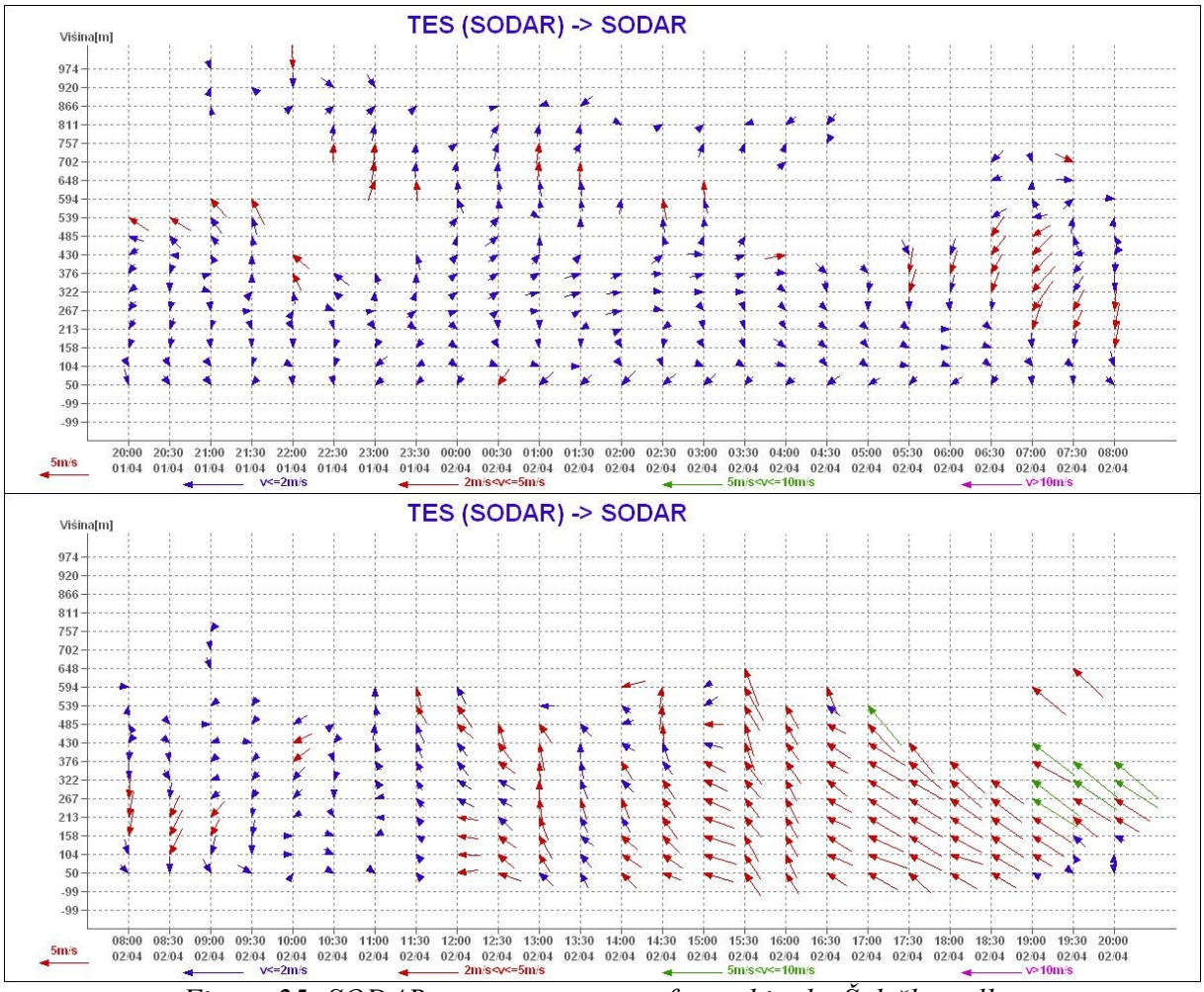


Figure 25: SODAR measurements performed in the Šaleška valley

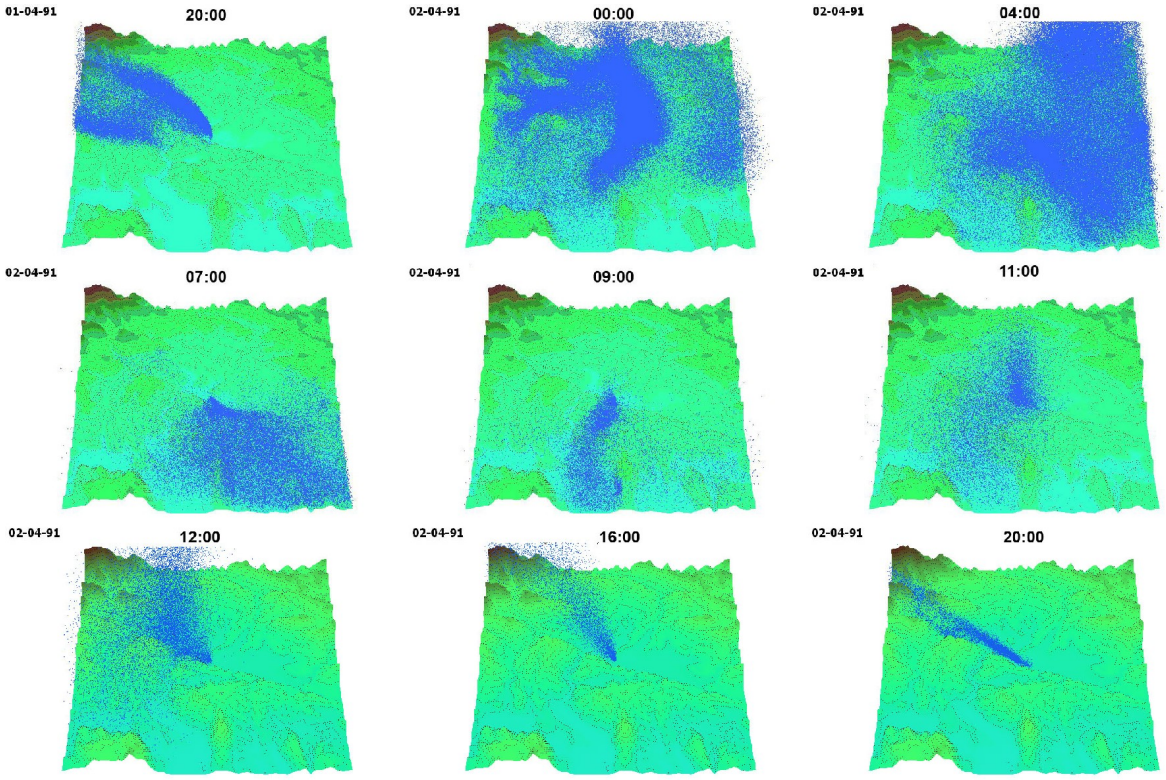


Figure 26: Plume spreading in all directions across the Šaleška valley

2.2. Zasavje region field data set

For a demonstration of the universality, namely the validation, of the developed methods presented in the following sections of this dissertation and for a demonstration of the efficiency of the ELPD computer model described in the following section *5.Integration of the proposed improvements in the computer model* a field data set is used that is available from another measuring campaign performed in 2005-2006 over the Zasavje region. The selected field data set's domain spreads over the area where all the complex terrain meteorological conditions are even more expressed than in the Šaleška region field data set presented in the previous section. The Zasavje region was selected as a field data set for several reasons:

- it spreads over a complex topography (deep valley surrounded with high hills),
- it has two major air-pollution sources located there: a cement factory and a thermal power plant,
- a database of ambient measurements is available from the measuring campaign performed from September 2005 till September 2006.

2.2.1. Terrain description

The Zasavje region is located in the central part of Slovenia. It extends over hilly terrain nearby the Sava river canyon in an area between the basin of Ljubljana and the Panonian lowland. The region's location is shown in Figure 27[93], where the regional townships from west to east direction are: Litija (approximately 6 500 inhabitants), Zagorje ob Savi (approximately 6 600 inhabitants), Trbovlje(approximately 18 000 inhabitants), Hrastnik (approximately 7 000 inhabitants) and Radeče (approximately 2 000 inhabitants).

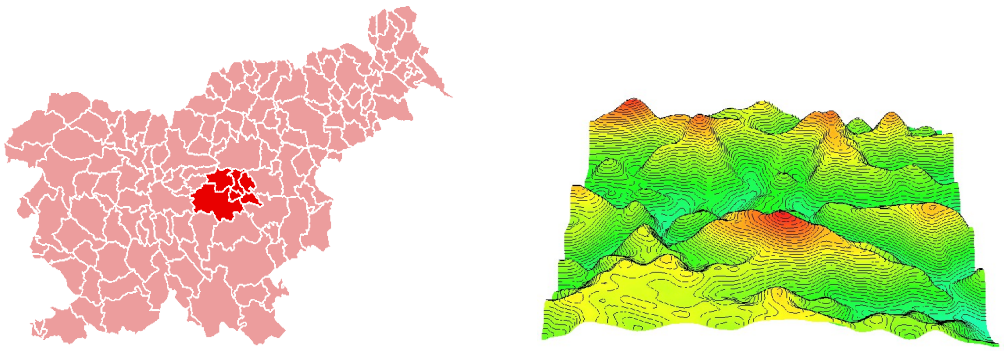


Figure 27: Location of the Zasavje region in Slovenia[93] and the topography of the region

The high pollution of the area is caused by high emissions, but it is also emphasized by the local micro-climatological conditions (low wind speeds, calm situations and strong thermal inversions) as the area is a highly complex terrain (canyon with steep slopes up to approximately 45 deg., several valleys perpendicular to the main canyon, hills with a relative height over 1000m). There are certain air-pollution situations that lead to increased concentrations in the complex terrain, as described in subsection 1.4.5.Complex terrain: plume impingement on high terrain depicted in Figure 28, pooling in valleys depicted in

Figure 29, drainage towards the population centres depicted in Figure 30 and persistence due to the channelling depicted in Figure 31. The plume impingement on Kum hill on the south from the centre of the domain is presented in the middle of Figure 28 and the increased ground-level concentration caused by the phenomenon is shown on the right side. In Figure 29 the pooling in the valleys transverse to the Sava river canyon is presented, where on the right side of the figure the increased ground-level concentration in all the regional towns is presented. Figure 30 represents the drainage of the air pollution towards the town of Hrastnik, where on the right side of the figure an increased ground-level concentration is present. The persistence of the increased ground-level concentration due to the channelling caused by the topography (upwards of the Sava river canyon) towards the town of Zagorje is presented on the right side of Figure 31.

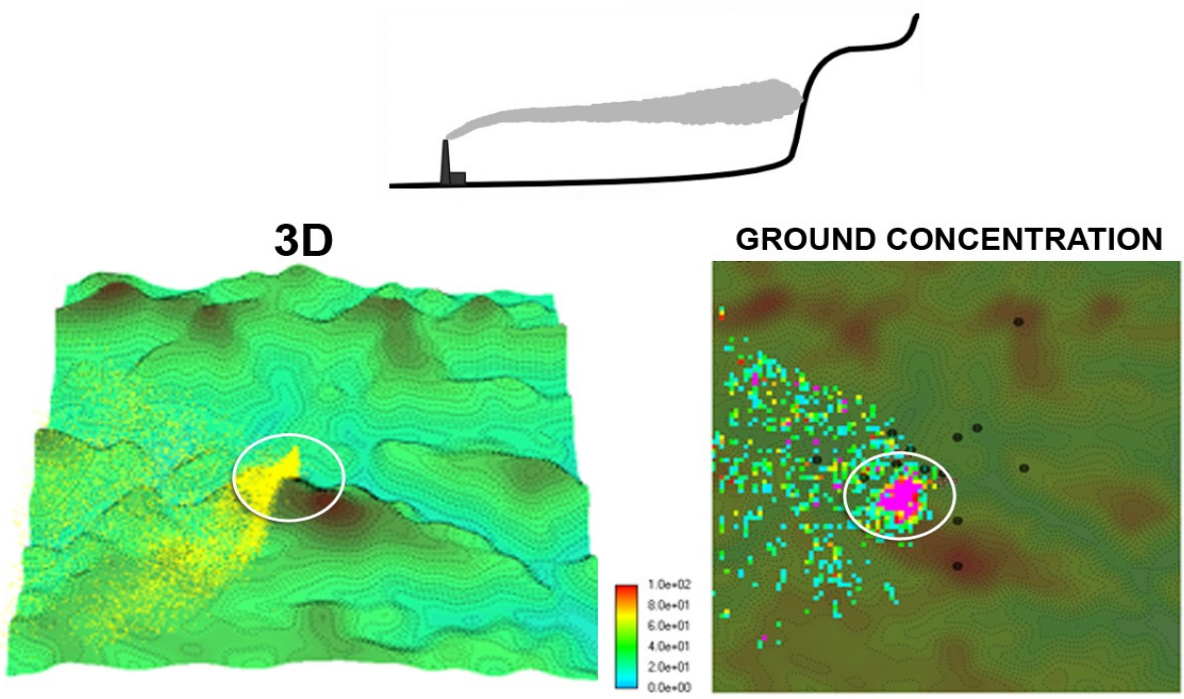


Figure 28: Plume impingement on high terrain – an example from the Zasavje region

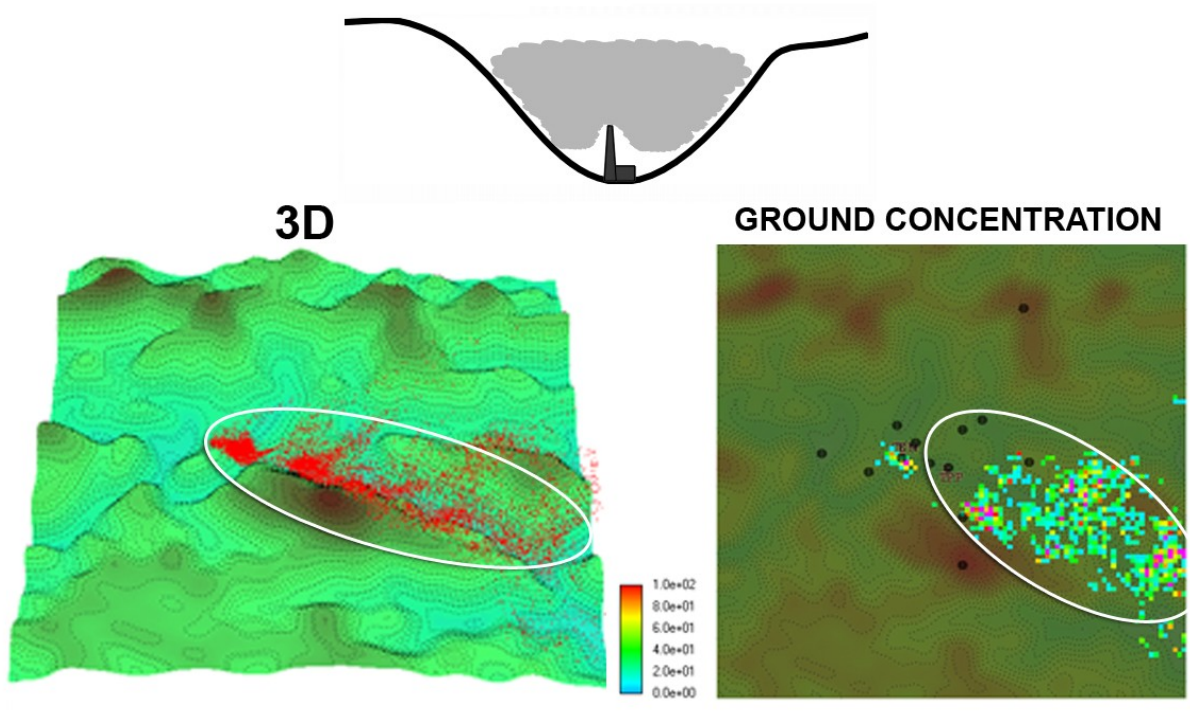


Figure 29: Pooling in the valleys – an example from the Zasavje region

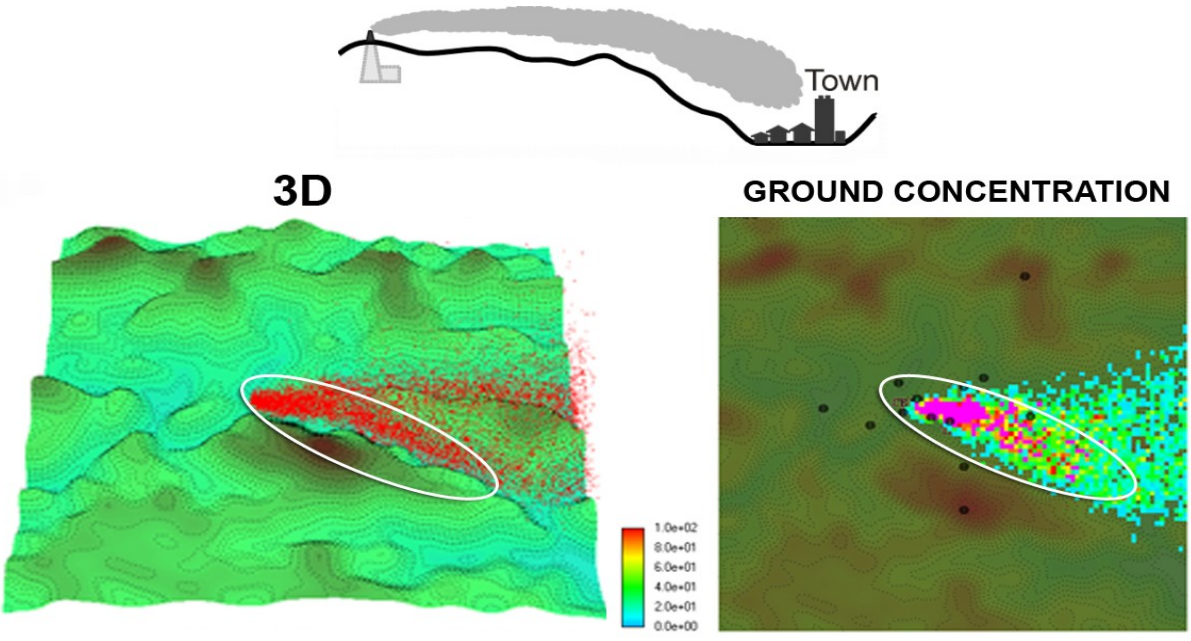


Figure 30: Drainage towards population centres – an example from the Zasavje region

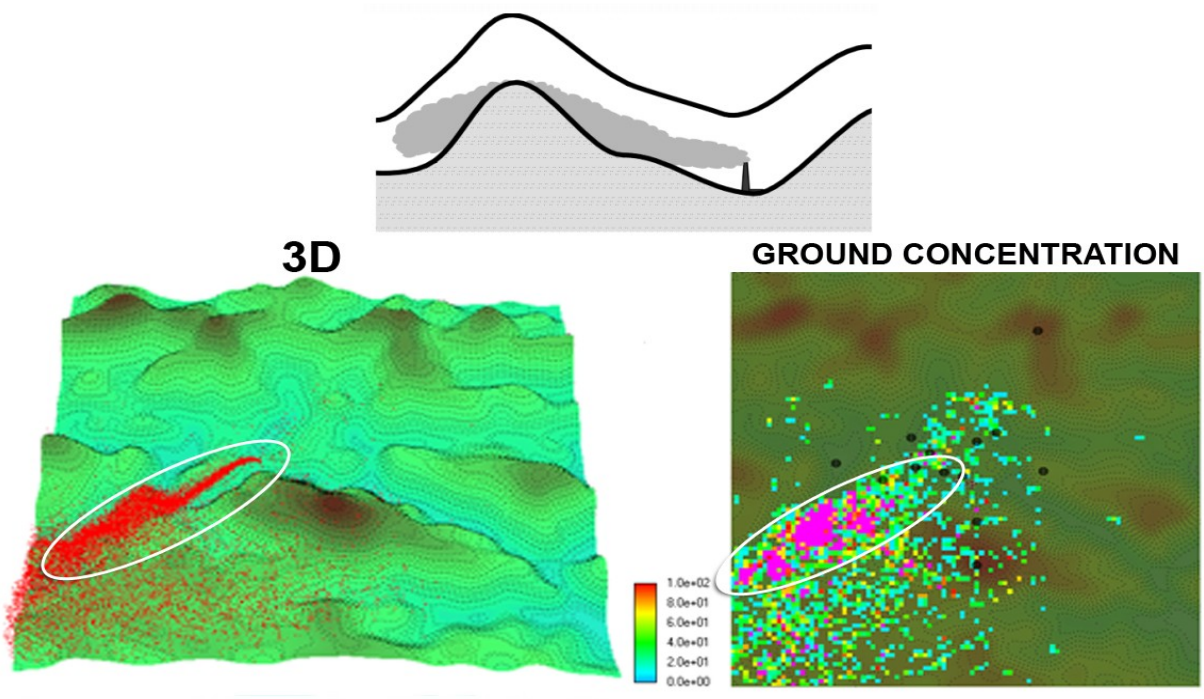


Figure 31: Persistence due to channelling – an example from the Zasavje region

2.2.2. Experimental measuring campaign

In the recent past the Zasavje region was facing serious air pollution. An experimental measuring campaign for regulatory purposes[94] was done to reconstruct the current air-pollution situation in the area, to quantify the expected reduction of SO_2 pollution by desulphurisation plants and to model the future scenarios (a new, planned gas-powered TPP). One year of on-line meteorological, air pollution and emission data was analysed in the measuring campaign. Across the area of interest it was mainly SO_2 and NO_2 pollution that exceeded the regulation limits.

The Zasavje region therefore represents a field data set for the evaluation of dispersion modelling in complex terrain[95]. In the area there is intense monitoring of the meteorological parameters (including one SODAR profile) ambient concentrations (9 stations) and one on-line emission station on the existing TPP stack. The on-line measured emission and ambient data for a one-year time interval from 1st of September 2005 until 1st of October 2006 were used to reconstruct the air-pollution situation across the area.

2.2.3. Sources of emission

The major sources of air pollution over the Zasavje region are a thermal power plant and a cement factory, both located near the town of Trbovlje near the Sava river (see Figure 32 and Figure 33). High air-pollution periods were present, especially during the winter inversions. The major source of the air pollution was a 360-m-high smoke-stack of the thermal power

plant, where coal rich with sulphur from a local mine is burned. The situation in the area was drastically improved by the installation of a wet desulphurisation plant in September 2006, after which the emissions from the smoke-stack dropped down to approximately 10% of the previous emissions.

For evaluation purposes, artificially generated high-level emission data is used in the following sections for the air-dispersion simulations, where a virtual source of emission is placed in the centre of the domain. The measured data can be used in this study only, because both the main air-pollution facilities are in the middle of purchasing an environmental license and publishing the data could cause some procedural problems.

Figure 32: Thermal power plant Trbovlje (left) and Cement factory Trbovlje (right)

2.2.4. Automatic environmental measuring system

There are two automatic environmental measuring systems located across the Zasavje region. The first system is owned by the Environmental agency of the Republic of Slovenia. Its code name is ANAS and it consists of three environmental measuring stations in the Zasavje region, represented by the green colour in Figure 33: Trbovlje, Zagorje and Hrastnik.

The second system is under the control of the thermal power plant . Its code name is EIS TET and it consists of seven environmental measuring stations, represented by a light blue colour in Figure 32: Kovk, Kum, Dobovec, Prapretno, Ravenska vas, Lakonca and Mrzlica. To measure the wind profile there is also an automatic SODAR measuring station installed near the thermal power plant, represented by a blue colour in Figure 33. The measuring parameters monitored on each of the automatic measuring stations are presented in Table 2. The real, measured meteorological data from the measuring system will be used for the air-dispersion simulations.

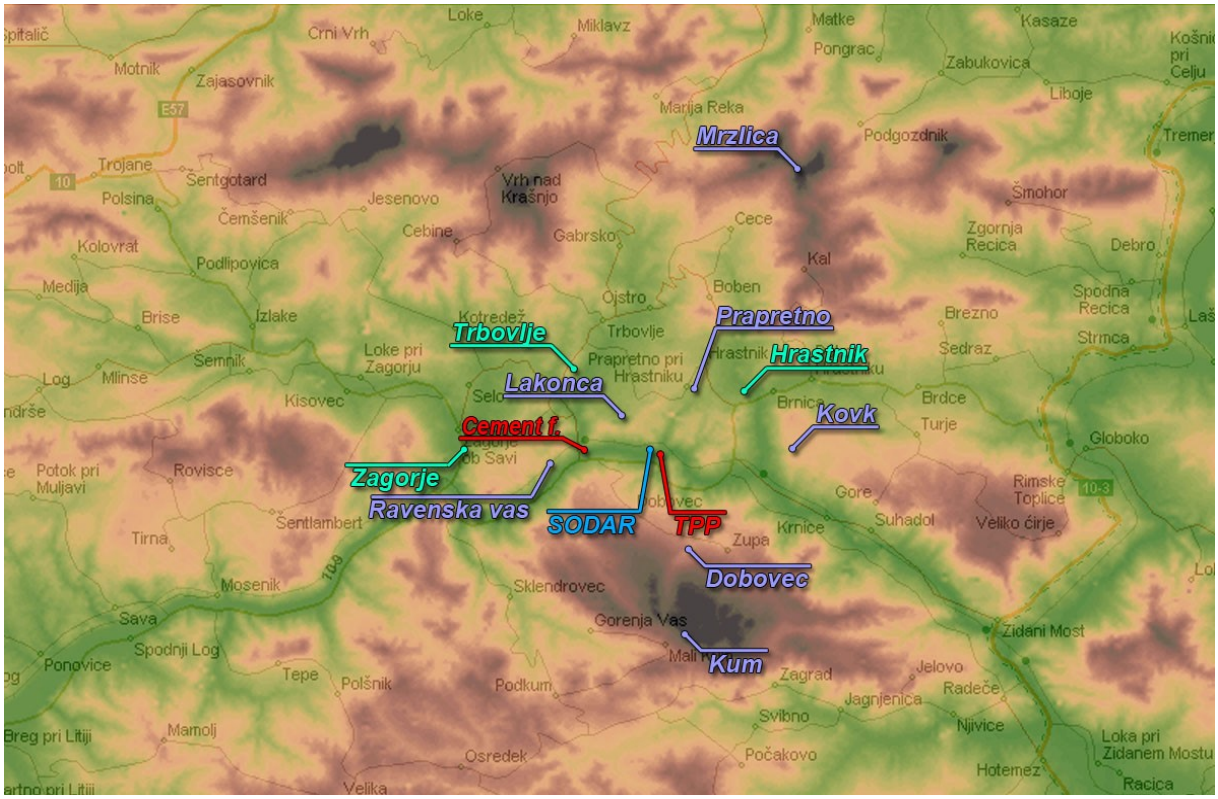


Figure 33: Locations of the sources and the automatic measuring stations across the Zasavje region

Table 2: Measuring parameters monitored across the Zasavje region

	Kovk	Dobovec	Kum	Ravenska vas	Prapretno
Air temperature	x	x	x	x	x
Relative humidity	x	x	x	x	x
Global solar radiation	x		x		
Wind	x	x	x	x	x
Dust PM10					x
SO2	x	x	x	x	
NO	x				
NO2	x				
NOx	x				
O3	x				

	Lakonca	Trbovlje	Zagorje	Hrastnik	Mrzlica
Air temperature	x	x	x	x	x
Relative humidity	x	x	x	x	x
Global solar radiation		x	x	x	
Wind	x	x	x	x	x
Dust PM10		x	x		
SO2		x	x	x	
NO		x			
NO2		x			
NOx		x			
O3		x	x	x	

Figure 34: Ambient automatic measuring stations at Ravenska vas (upper-left), Dobovec (upper-right), Zagorje (lower-left) and SODAR (lower-right)

2.2.5. Situation selection from the Zasavje region field data set

A database of measurements is available for the full duration of the measuring campaign, i.e., from the 1st of September 2005 until the 1st of October 2005. A situation that lasted from the 14th of October 2005 at 00:00 until the 16th of October 2005 at 00:00 was selected because of its complexity. This complexity demands greater computational resources for the reconstruction. It is used for the validation of the contributed methods presented in the following sections of this thesis and to demonstrate the efficiency of the enhanced Lagrangian particle-dispersion model, described in the following section. The SODAR measurements performed on the 14th of October 2005 are presented in Figure 35 and on the 15th of October 2005 in Figure 36. The arrows on the figures represent the horizontal wind component at different heights above the SODAR. The three-dimensional behaviour of the reconstructed plume presented in Figure 37 is consistent with SODAR measurements and it illustrates that the plume spread in all directions during this period of time. The results reveal that a relatively strong wind from the west was present over the domain until the morning of the 14th of October 2005. In the morning the wind slowly changed its direction from the opposite side and remained relatively constant until the evening of the 14th of October, when calm conditions occurred. After the occurrence of the conditions the calm persisted for practically

the whole of the following day, the 15th of October 2005, which caused the accumulation of air pollution in the domain. The accumulation of air pollution is very well presented in the lowest three 3D presentations of the plume in Figure 37. A lot of computational time was used for this reconstruction because the number of active particles almost reached the limit of the available computational resources during the calm.

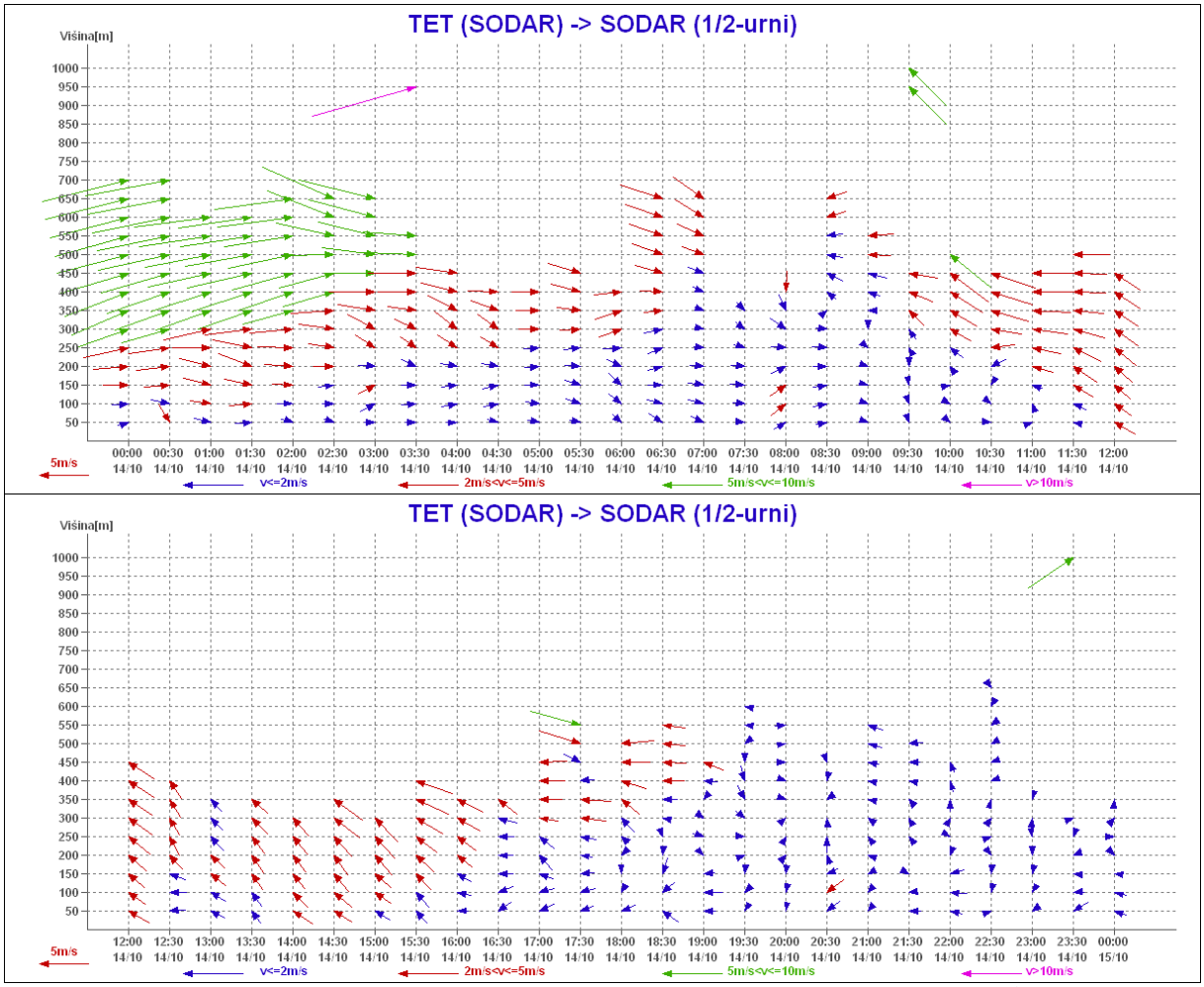


Figure 35: SODAR measurements performed in the Zasavje region on the 15^{h} of October 2005

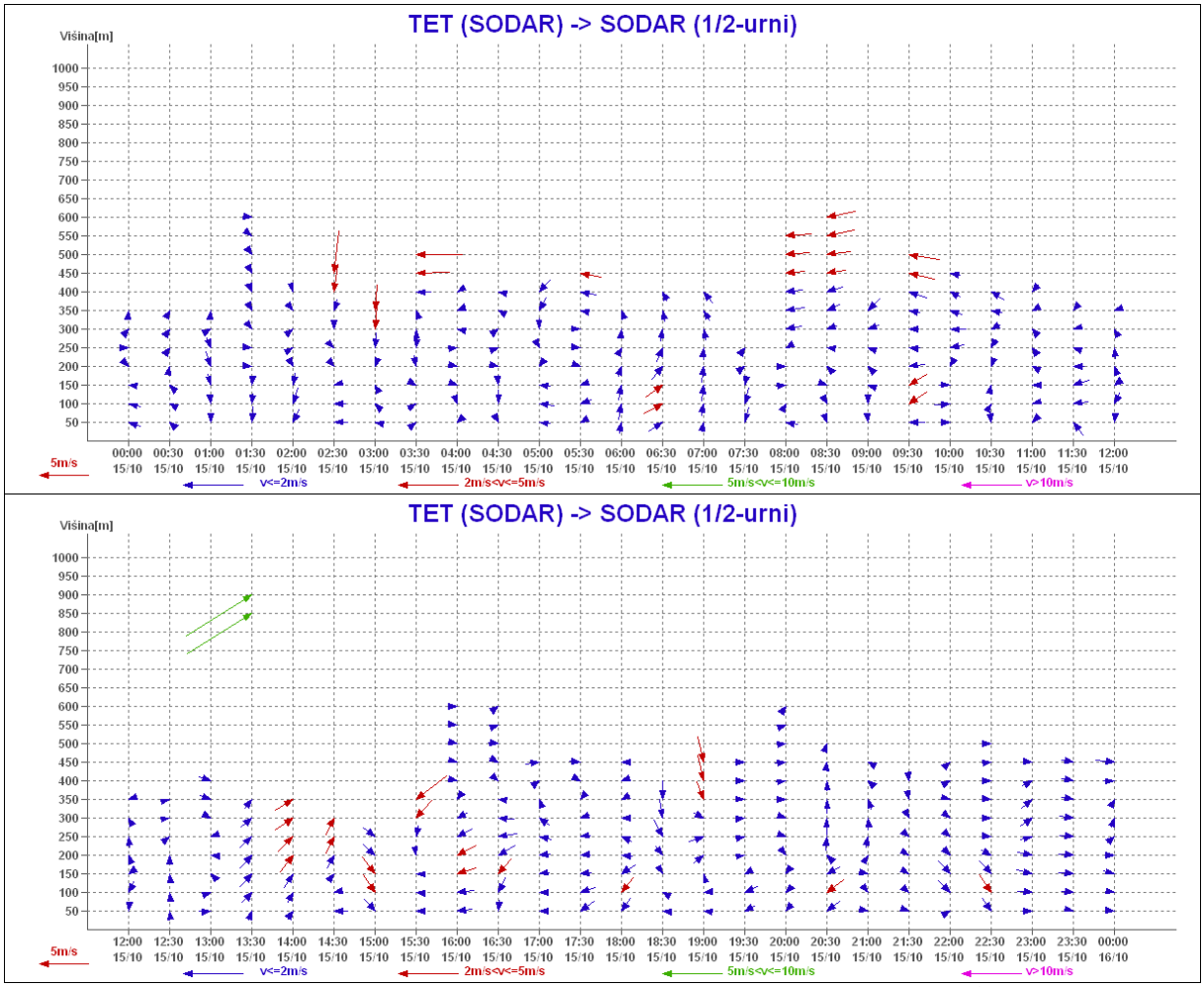


Figure 36: SODAR measurements performed in the Zasavje region on the14^{h} of October 2005

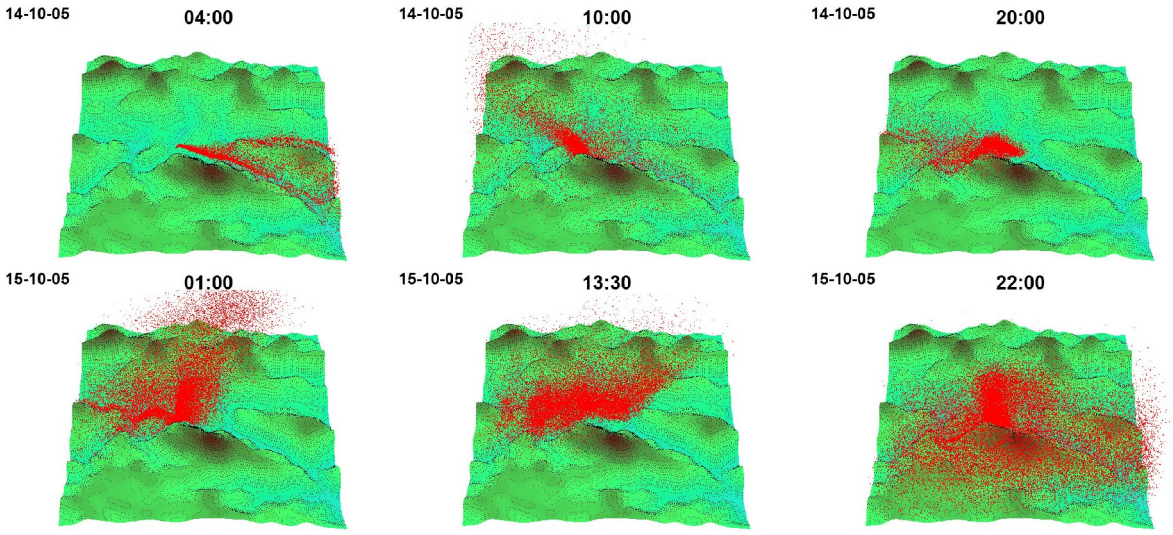


Figure 37: Three-dimensional representation of a plume spreading in all directions the across Zasavje region

3. LAGRANGIAN-PARTICLE AIR-POLLUTION MODELLING METHODOLOGY

An air-pollution model is the final result of the modelling process and is used in simulations to reconstruct the air pollution over selected domains of interest. In the following subsections of this section air-pollution modelling and the simulation process are described as well as the air-pollution (AP) model based on Lagrangian particle dispersion (LPD), according to the terminology used in a series of books *Air pollution modeling and it application* by Kluwer Academic/Plenum Publishers[96].

3.1. Air-pollution modelling

Air-pollution modelling is generally a cyclic process, as presented in Figure 38. The presentation is created on the basis of a description of the general modelling process[97]. air-pollution modelling is defined as an attempt to describe the functional relation between the emissions and the occurring concentrations[96] in the surroundings. It can give us a relatively complete and consistent description, which also includes an analysis of the causes (emissions sources) that lead to measured concentrations. The modelling in general begins with the definition of the problem. In air-pollution modelling this phase represents the determination of the area of interest, which usually consists of major air-pollution sources in the centre of the domain. Its width is determined according to the purpose of the application and is usually defined by legislation.

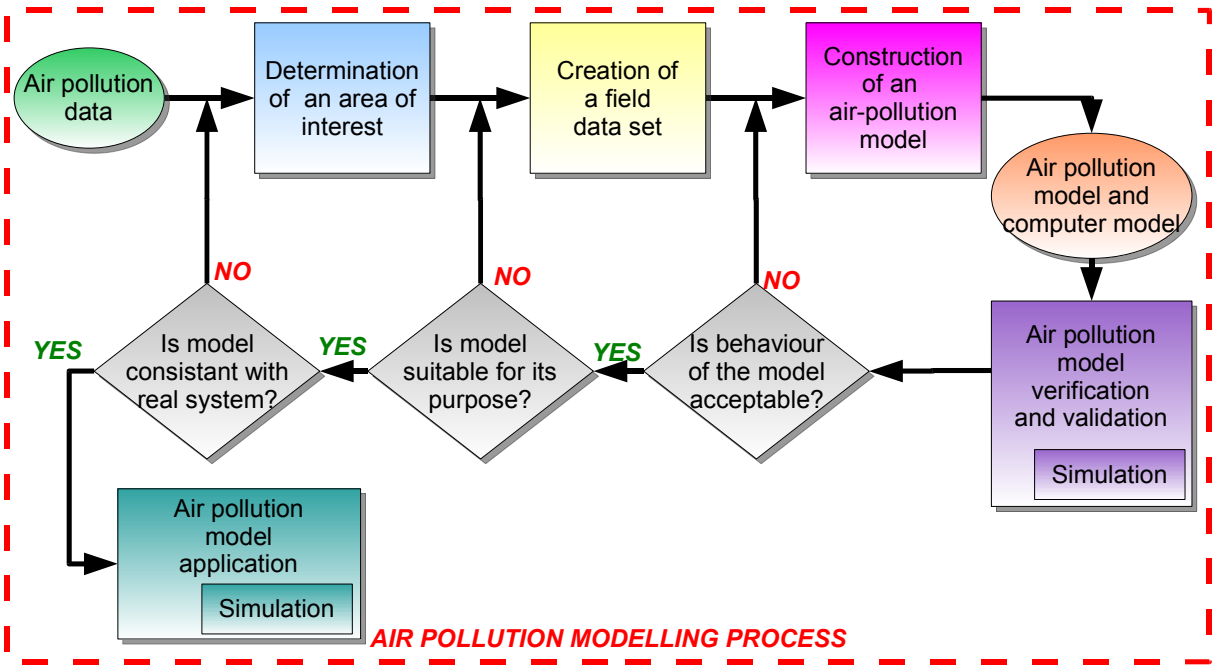
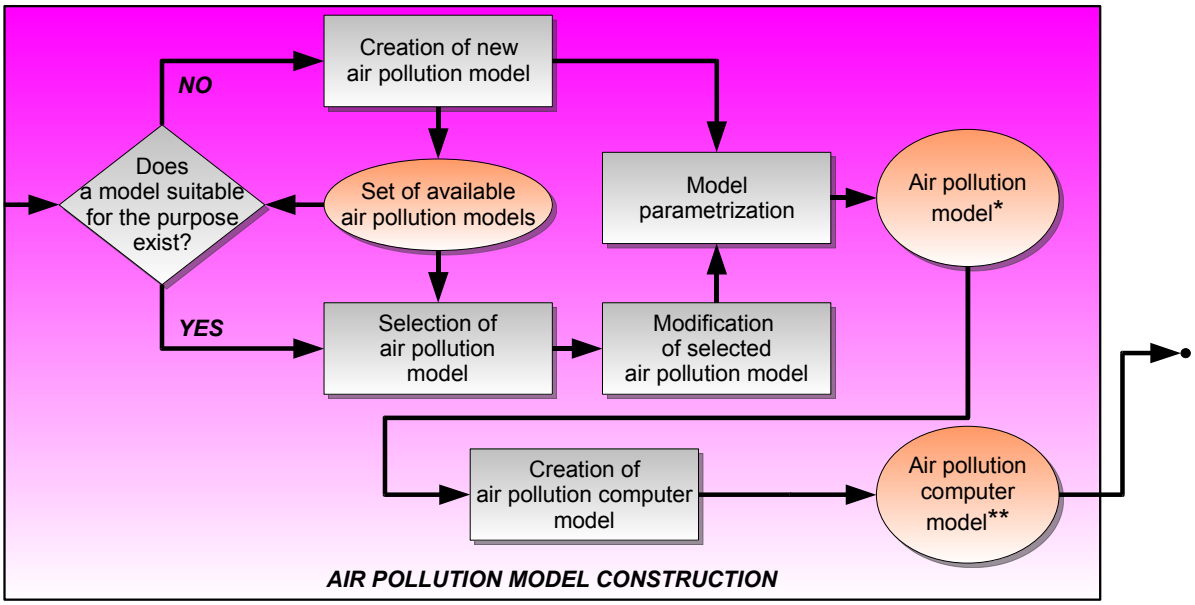


Figure 38: Air-pollution modelling process[97]

When the area of interest is determined, measured data from the domain is collected. From this collection of measured data a *field data set* is created. The field data set consists of all the

available emission data, the meteorological data and the topographical data. When some data is not available some measurement estimations can be used. The quality and quantity of data is defined by the selected air-pollution model. While some models for simple terrain require a relatively simple data set (i.e., in the most simple case only the emission data and the wind measurements at only one position near the source) other models for complex terrain require a relatively large and precise set of data (i.e., detailed topographical data, meteorological measurements in several positions spread across the area with the vertical wind profile and the precise emission data).

The model's construction is based on the application, the terrain and the available model concepts. It consists of several steps, presented in Figure 39. There are two results of the model construction phase: the air pollution mathematical model and the air pollution computer simulation model. In this thesis, for the air-pollution mathematical model a simplified term, *air pollution (AP) model*, is used and for the air-pollution computer-simulation model the term *air pollution (AP) computer model* is used.



*simplified term for *air-pollution mathematical model*
**simplified term for *air-pollution computer-simulation model*

Figure 39: Air pollution model construction

In the first step the existence of the AP model that can satisfy the requirements of the purpose is checked. If it does not already exist, I.e., an AP model that satisfies these requirements, in a set of available air-pollution models (i.e., the Lagrangian particle, the Eulerian, the Gaussian, etc.) a new AP model is constructed. Otherwise, the selected AP model can be modified to satisfy the requirements. In the next step a parametrization of AP model is performed, where the parameters of the model are determined according to the selected area of interest.

For simple terrain and steady-state conditions, simple models with a relatively small number of inputs are used, while for complex terrain conditions advanced models are used. The Lagrangian particle atmospheric dispersion model is generally accepted as the best, recently known, atmospheric dispersion model for use over complex terrain at the local, the regional and the global scales[9,10]. A large number of the model's parameters must be determined according to the selected area of interest, the available measured data and the application.

In the final step of the model's construction the AP computer model is finally designed to be used in a computer simulation. The AP computer-model design consists of three main steps: creating an atmospheric dispersion computer module, creating an input module that is used to pre-process and transform the input data into a suitable form for the atmospheric-dispersion computer module and creating an output module that is used to transform the output data into a suitable form for presentations and archiving.

In the next phase of the modelling the simulation is performed to verify and validate the constructed model. If the verification fails the modelling process returns to the model modification. When the model's behaviour becomes acceptable, the validation begins. If the model is not suitable for its purpose, the collection of additional data or the filtering of data is performed and the model is reconstructed according to new inputs. When the model is accepted for its purpose the consistency with the real air-pollution dispersion is evaluated. If the model meets the defined requirements of the application the modelling process ends and the model is ready to be used for an application.

The air-pollution model's application is the final phase of the air-pollution modelling process. The simulation is used to reconstruct air-pollution situations over a selected area of interest according to different practical purposes: air-pollution assessment of the existing sources of emissions in the domain, the planning of new industrial facilities in the domain, or for emergency warning systems in the case of accidental releases of air pollutants.

3.2. Air pollution simulation

Air-pollution simulation is used to reconstruct the air-pollution situation over the selected area of interest (domain) for the selected period of time. air-pollution situation is an air pollution that usually lasts for some defined period of time. It is determined by concentrations of certain species (i.e. SO_2) over area of interest. The period of time for the reconstruction is usually split into several equally long episodes of air pollution, as presented in Figure 40, where one simulation run consists of several air-pollution-episode reconstructions. For the regulatory purposes the duration of one air-pollution episode should be equal to the meteorological measuring time interval of half an hour[98,99]. Each air-pollution-episode reconstruction usually consists of three main steps, presented in Figure 41.

The statistical elaboration of an air-pollution situation is performed for different periods of time, defined by legislation, where several sub-sequential air-pollution episodes are reconstructed. The number of reconstructed air-pollution episodes depends on the length of the time interval for a statistical elaboration, i.e., for the elaboration of the air pollution of one day, 48 air-pollution episodes must be reconstructed (1 episode per ½ hour). The

reconstruction of one episode is made by performing one simulation run with the air-pollution model. During the simulation process the air-pollution model reconstructs the air pollution over the selected area of interest by considering the given topography of the domain and the available measurements of the meteorological and emission data. The results of the air-pollution reconstruction are produced for the visualization and statistical elaboration at the end of the simulation.

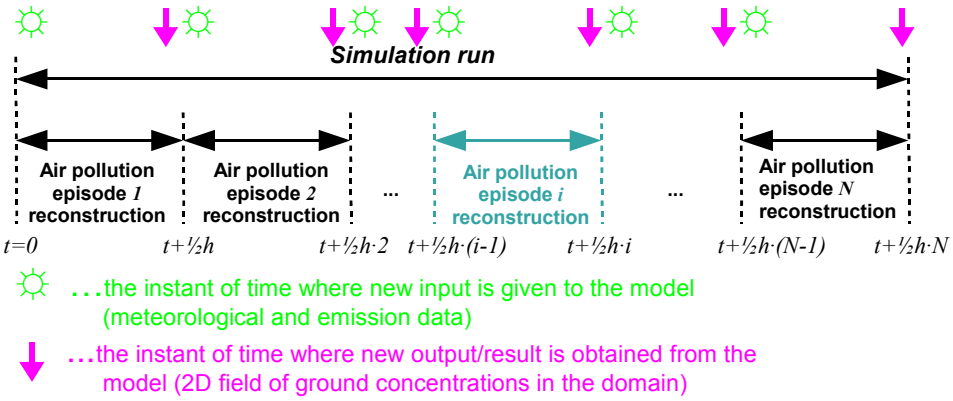


Figure 40: Simulation run

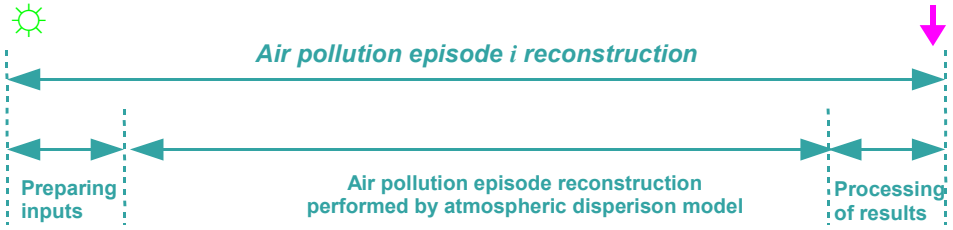


Figure 41: Reconstruction of a single air-pollution episode

3.3. Air-pollution model

The most often result of air-pollution modelling process for complex terrain domains is an air-pollution (AP) model based on Lagrangian particle dispersion (LPD) because it is generally accepted as the best recently known air-pollution reconstruction model at the local, regional and global scales[9,10].

The AP model consists of the main components, as presented in Figure 42. The main structure of the air-pollution model remains the same for all the research purposes of this thesis and it is not described in detail. Also, the main components of the air pollution are not described in detail because they are not the subject of this thesis. Except for the LPD model, which is described in subsection *1.4.6.The Lagrangian particle-dispersion model*. Detailed descriptions can be found in the literature[11,14,19,20,21,22,23,24,84,100,101,102,103].

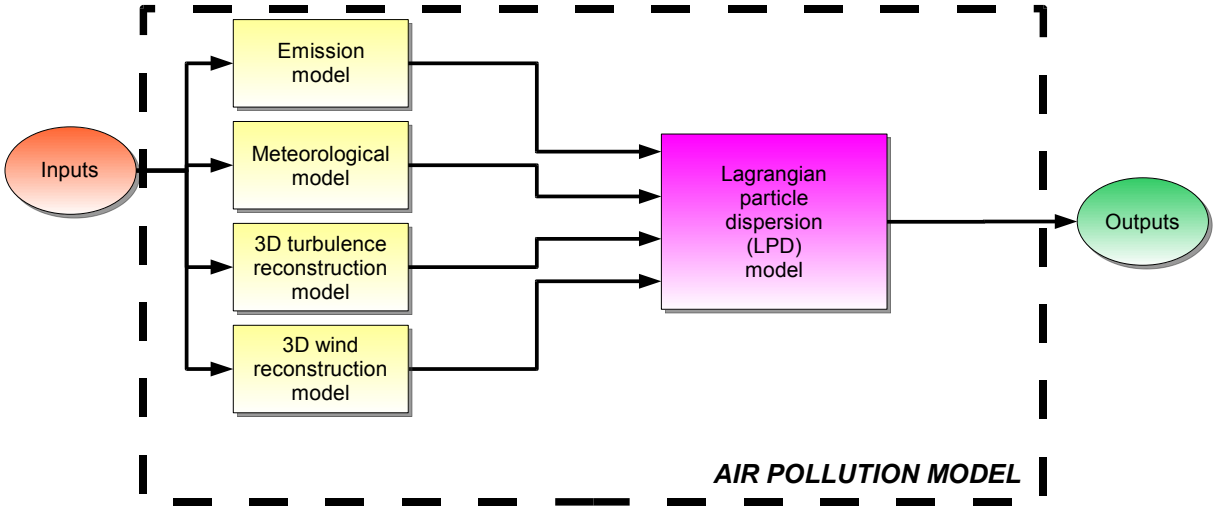


Figure 42: Air pollution model based on Lagrangian particle dispersion model

Inputs

There are several inputs that must be available for an AP model based on LPD:

- *topography data* of the area of interest, which consists of 3D topography data and a map of the land use,
- *current meteorological data* measured at several different locations, equally spread over the area of interest, which consists of the basic meteorological parameters (air temperature, relative humidity, air pressure, solar radiation, precipitations, wind speed and wind direction), the vertical wind profile and the vertical temperature profile,
- *current emission data* from all known sources of emission over the area of interest, which consists of a relatively detailed description of the emission sources, like their position, height and diameter, exit temperature and velocity of the emitted gas and

emission rate,

- *initial state of air pollution in the domain*, which consist of a description of the already dispersed air pollution, this input is optional and when it is not available a clean atmosphere over the domain is assumed.

Outputs

There are two main outputs of the AP model based on LPD:

- *the final state of the air pollution in a domain* that consists of air pollution dispersed from emission sources and the additionally dispersed initial air pollution
- the 3D grid of concentrations in the domain where the domain is split into a 3D grid of equally distributed cells, where the concentration for each cell is determined by the AP model.

3.4. Air-pollution computer model

The air-pollution (AP) computer model scheme presented in Figure 43 does not differ significantly from the AP model scheme. The presented description is based on the actual AP computer model that is designed according to the presented Lagrangian particle-modelling methodology. For all the practical experiments in this dissertation the AP computer model *AriaIndustrie,* designed by *ARIANET s.r.l.,* is used as the base into which the modifications are built. The selected AP computer model uses the LPD computer model *Spray* , described in the user's manual[103] and the paper[11] by Tinarelli et al.

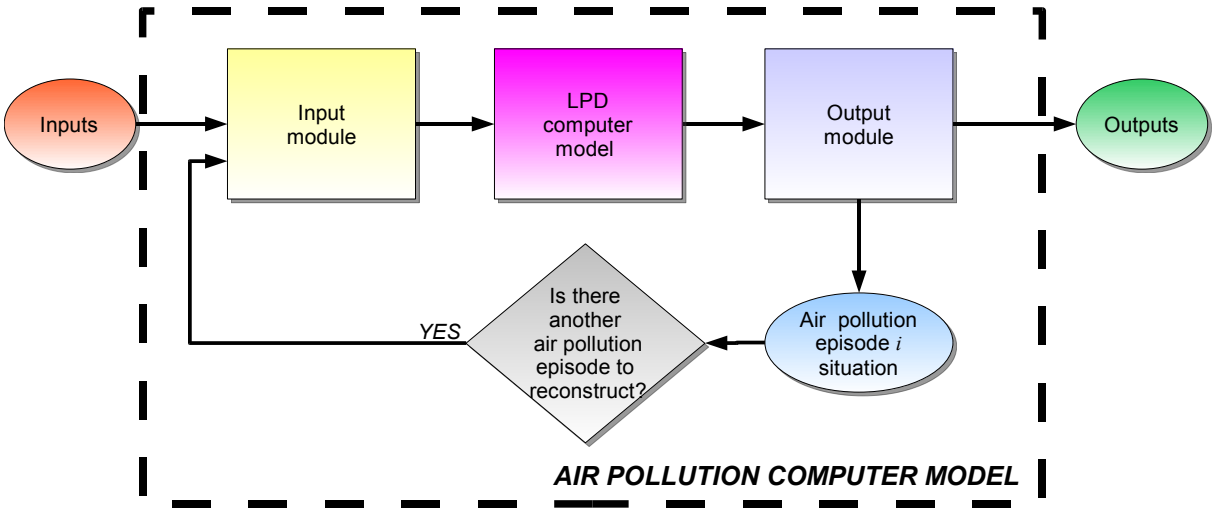


Figure 43: air-pollution computer model

The AP computer model consists of three main components, presented in Figure 43[11,14,103]. The *Input module* modifies and generates the input data to be in an appropriate input form for the *Lagrangian particle dispersion (LPD) computer model*. When the inputs are a prepared

simulation using the LPD computer model is executed. At the end of the simulation the outputs of the model are processed by the *Output module,* which generates results according to the requirements of the application that the AP computer model was designed for, i.e., for on-line systems or for off-line elaborations. The present AP computer model does not include the computational improvements proposed by Schwere et al.[10].

The following data must be available for the *Input module*:

- *topography data* of the area of interest where besides the 3D topography data (Digital Elevation Model) of the domain also 2D Corine Land Cover data of the domain (map of the land use based on the interpretation of satellite images) is required,
- *meteorological data* measured at several different locations, equally spread over the area of interest, where besides the air temperature, relative humidity, air pressure, solar radiation, precipitations, wind speed and wind direction as well as the vertical wind profile from SODAR and the vertical temperature profile from RAS are required,
- *emission data* from all known sources of emission over the area of interest, where each source of emission is described by static (position, height and diameter of source) and dynamic (exit temperature and velocity of emitted gas, emission rate in kg/s) parameters.

For the simulation with the *LPD computer model,* besides inputs prepared by the *Input module*, also the initial state of the air pollution in the domain must be available and several simulation run parameters (settings) must be defined. When the LPD computer model is finished the outputs of the model are processed by the *Output module.* According to the different practical purposes, the module can generate results like:

- 2D ground concentration fields in the form of ASCII files or pictures,
- 3D presentation of plume with particles,
- ground level wind field presentations,
- others according to the demands of the application that the modelling system is designed for.

General modules of the computer air-pollution model are described in detail in the proceeding subsections.

3.4.1. Input module

There are several inputs that must be prepared before the air-pollution episode reconstruction with the AP computer model is performed[11,14,103]. Some of the inputs are constant, like terrain elevation, locations of measuring stations and emissions sources, and some can vary with time, like the meteorological and emission parameters. The input processing and preparation is done in several sub-steps, as presented in Figure 44.

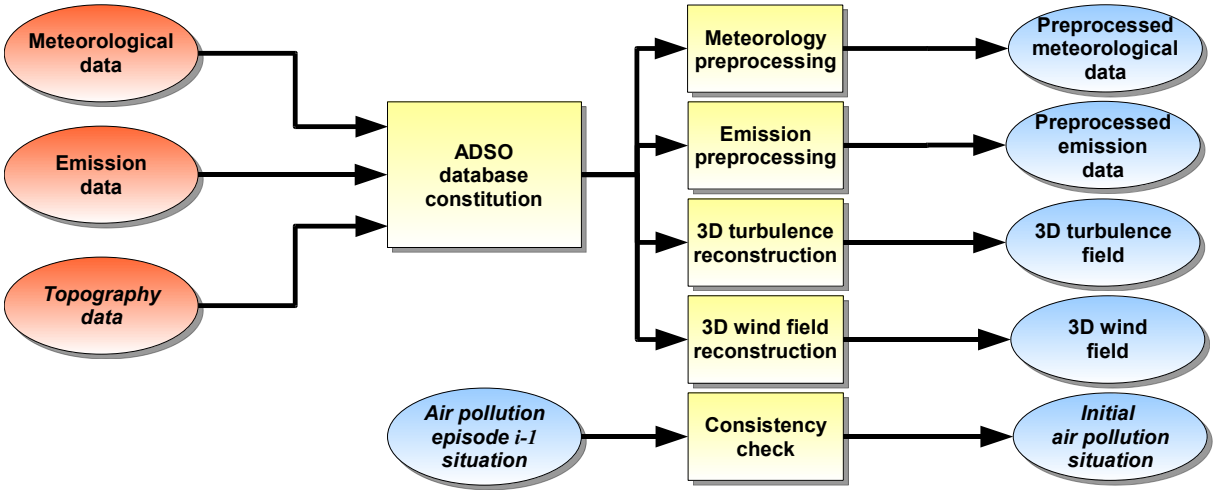


Figure 44: Sub-steps in processing and preparing model inputs

A database of the last meteorological conditions needs to be created. It must consist of the measured meteorological data from automatic measuring stations across the modelling domain and also of the SODAR and RAS measuring data.

There must also be a database about the current emission parameters created. It must consist of the current source positions and the amount of emitted species for each source must be provided. In the experimental model from the validation kits described in the previous section only one species type is used: sulphur dioxide (SO_2).

The topography data consist of the 3D topography data (Digital Elevation Model) of the domain and the 2D Corine Land Cover data of the domain. The topography description consists of a 2D grid of cells. This also means that the resolution of the topography description depends on the size of one grid cell (i.e., if the domain is 20x20 km wide and consist of 100x100 cells, the size of one cell is 200x200 m) which represents the final resolution of the inputs and indirectly also the final resolution of the outputs. From this data two-dimensional arrays of the *roughness length* z_0, *albedo* and *Bowen Ratio* (the ratio between the sensible and the latent heat fluxes over the portion of the terrain represented by the grid cell) are generated and given as the input to the LPDM. A time series of local meteorological data covering the whole episode period must be supplied, containing solar radiation fluxes, air temperatures close to the ground, and vertical temperature gradients. All this information is used by the parametrization schemes[104] to build the Surface Layer and Boundary Layer scale parameters over each grid cell (i.e., the Mixing Height h_{mix}, the friction velocity, the Monin Obukhov length L, and the convective velocity scale). The vertical profiles of the turbulence variables are then generated using stability-dependent similarity relationships[105].

When meteorological, emission and topographic databases are prepared, a special ADSO[102] database must be constituted. The ADSO is a specific database for atmospheric flows and pollutant dispersion over a complex terrain with the following capabilities:

- handling topography at different scales,
- use of many types of vertical coordinates, specific to meteorology (most of them

terrain-following),

- dealing simultaneously with 3D vector fields (wind field) and 3D scalar fields (temperature),
- handling several types of sensors, at different locations and reporting at different times,
- permitting special processing for particular sensors (e.g.: SODAR, SF6 tracer, aircraft, etc.).

ADSO incorporates some functions of the GIS (Geographical Information System), some DBMS (DataBase Management System) functions, but it is very much oriented towards the production of 3D fields for wind, turbulence and pollutant concentration in the atmosphere.

The constitution of the ADSO database triggers the constitution of a special database used as an input into the wind reconstruction system. The 3D wind field modelling system reconstructs a three-dimensional wind field in the defined domain. A special application is used afterwards to extract only the ground wind field from the reconstructed 3D wind field.

Within the selected AP computer model *AriaIndustrie* used for experiments in this thesis, the diagnostic non-divergent model *MINERVE*[102] is mainly used for this scope. In the selected AP computer model the outputs from other wind field and turbulence simulators, such as the prognostic non-hydrostatic meteorological code *RAMS*[106], can also be used.

The initial state of the air pollution should also be available as an input into the proceeding step. Usually, the final state of the air pollution from the previous air-pollution episode reconstruction is used. If the initial state of the air pollution is not available the initial state without air pollution is assumed.

3.4.2. Lagrangian particle-dispersion computer model

The Lagrangian particle-dispersion (LPD) computer model is a three-dimensional model designed to simulate the airborne pollutant dispersion[14]. A conceptual model of the LPD computer model is presented in Figure 45. The initial air-pollution situation consists of a set of particles that already exists in the domain (particles that were emitted in the previous air-pollution episodes). The emission data is used to generate new particles. The number of new particles and the mass of the new particles are inversely dependent; for the same emission a larger number of lighter particles or a smaller number of heavier particles can be released. The number of particles is controlled by the *PDNC* parameter (defined and described in details in the following subsection *3.6.Determination of acceptable simulation results*). Other properties of the new particles (i.e., initial velocity, initial position, etc.) are also determined according to the emission data. When the simulation of the dispersion of the particles (new and existing) in the domain is finished, the output of model is the final air-pollution situation and the 3D grid of concentrations[103]. The final air-pollution situation has the same form as the initial air-pollution situation, so that it can be used in the next air-pollution reconstruction as an input. The 3D grid of concentrations is determined by box-counting the concentration

estimation (described and discussed in detail in the following subsection *4.2.Method for estimation of a cell concentration based on kernel density*).

The LPD computer model is able to take into account the spatial and temporal inhomogeneities of both the mean flow and the turbulence[103]. Using this model the concentration fields generated by a point, linear, area or box sources can be simulated. The behaviour of the airborne pollutant is simulated through "virtual particles", whose mean movement is determined by the local wind and the dispersion is determined by the velocities obtained as a solution of Lagrangian stochastic differential equations. Different parts of the emitted plumes are therefore exposed to different atmospheric conditions, which allows more realistic reproductions of the complex phenomena (low wind speed conditions, strong temperature inversions, flow over topography, presence of terrain discontinuities such as land-sea or urban-rural) that are hard or impossible to simulate with more traditional approaches like the Gaussian approach[11]. For the experiments performed in this thesis the LPD computer model *Spray,*[103] designed by *ARIANET s.r.l.,* is used.

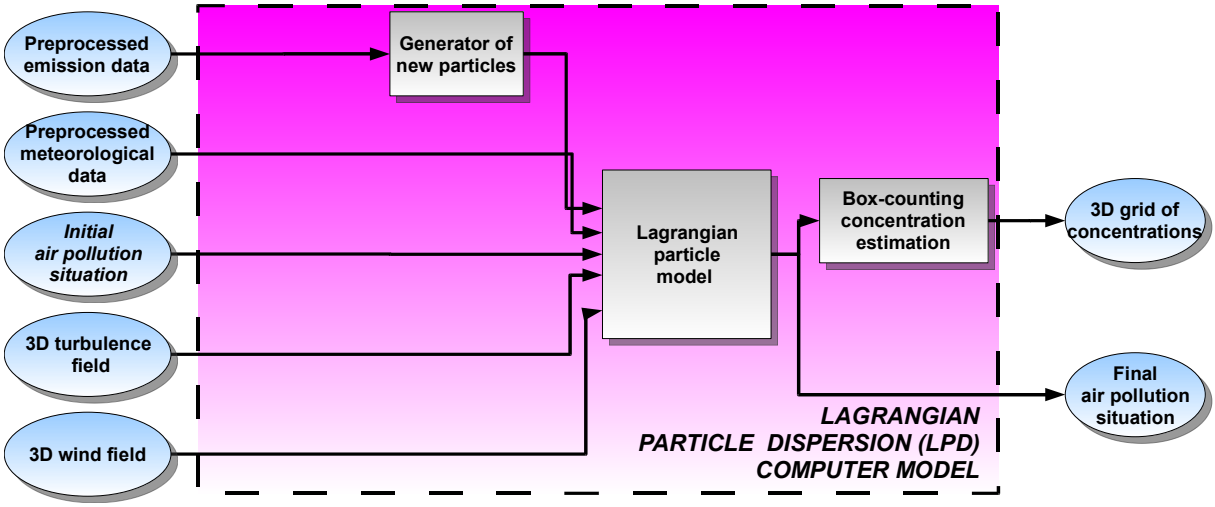


Figure 45: Lagrangian particle-dispersion (LPD) computer model

The LPD computer model simulates the air-pollution dispersion inside a computational domain, bounded on the upper side by a horizontal plane and on the lower one by an orographic function $z_g(x,y)$ that is computed through a bilinear interpolation of the Digital Elevation Model, given on a two-dimensional grid and defined by the user[11,14,103]. Thomson's 1987 scheme with Gaussian random forcing[24] has been adopted to describe the particle-velocity fluctuations and to generate the average concentrations at the ground at the same horizontal resolution as used by meteorological reconstructions. The LPD computer model takes into account both the cross-correlations between all the Cartesian components of the turbulent velocities and the vertical skewness[11]. The mean flow is represented by wind vectors, defined on three-dimensional grids in a terrain-following reference system x,y,s where the vertical coordinate s is defined, as follows, in equation (3.1).

$$s = \frac{z - z_g}{z_{top} - z_g} \tag{3.1}$$

z ...Cartesian vertical coordinate
z_{top} ...domain top level
z_g ...topography height about sea level

The wind at the particle positions is linearly interpolated using the values at the neighbouring points given on this grid[14,103]. The turbulent velocities u'_x, u'_y and u'_z are updated on the fixed Cartesian reference frame. Data representing the turbulence are described by three-dimensional arrays in the *x,y,s* system. The plume rise undergone by the hot-stack plumes is simulated by means of Anfossi's formulation,[101] which takes into account the horizontal and vertical variations of both the mean wind and the atmospheric stability[103].

The selected LPD computer model takes as its input a time series of three-dimensional wind fields over a complex topography, generated by an external meteorological code in a binary format[11,14,103]. The source parameters and the run options must be prepared by the user and the *Input module*. Multiple point, linear or areal sources can be configured, each emitting different non-reactive chemical species. The total number of particles is selected by the user, where the *PDNC* parameter (defined and described in detail in the following subsection *3.6.Determination of acceptable simulation results*) can be controlled and the particle mass for each species is, during the computer simulation, automatically calculated as a function of the emission characteristics.

3.4.3. Output module

The 3D concentration fields of each species used in the simulation for different levels above the ground are reconstructed from data about particle positions. After that a special database for a graphical tool that generates the graphical presentation is created. From this database the presentations of the surface wind streams, the ground concentrations and the three-dimensional representation of particles are generated. Typical results from this step will be used and presented in the following sections. The sub-steps in the processing of the results are presented in Figure 46.

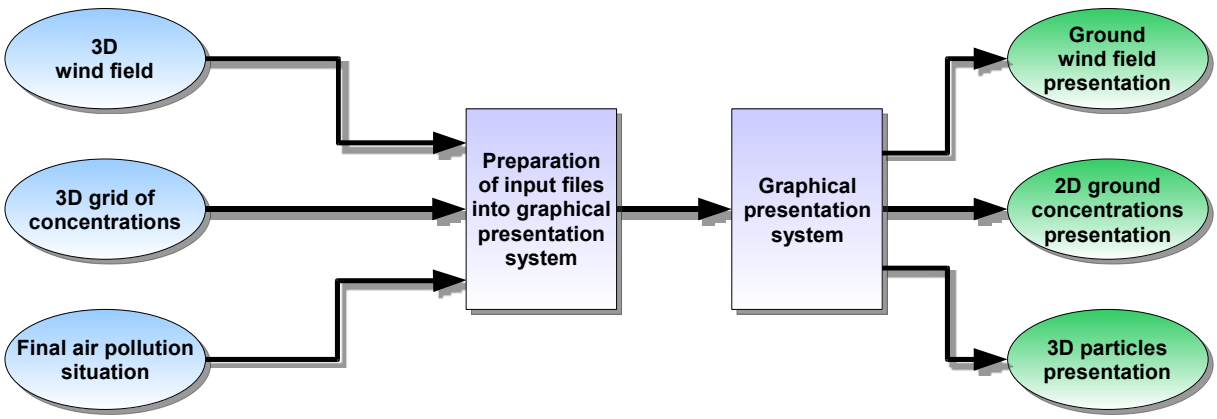


Figure 46: Sub-steps in the interpretation of the results

3.5. Evaluation methods

The proposed methods presented in Section 4 contribute into the existing air-pollution modelling methodology and integrated into the enhanced Lagrangian particle dispersion model to make it possible to obtain the same reliable results with a shorter computational time . To evaluate the quality of the simulation results performed with the AP computer model based on the LPD several evaluation methods are contributed, based on adjustments to the existing evaluation methods[81,107,78,108,18]. The result of one air-pollution episode reconstruction is a 3D concentration field, out of which only the 2D ground-level concentration field (a matrix of size *MxN* grid cells) is relevant to this study. Two 2D ground concentration levels are presented in Figure 47: the left is obtained with a larger number of particles than the right.

To estimate the quality of the result of each air-pollution episode reconstruction the *reference concentration field* needs to be defined. It is common practice[109] that this reference concentration field is obtained by using a very large number of particle trajectories for the full three-dimensional model simulation. In our case the reference concentration field is a two-dimensional ground-level field. It is the result of the air-pollution episode reconstruction with the highest possible number of used particles according to the available computational resources (i.e., for the performed experiments about 3,000,000 particles are used for the reconstruction of one air-pollution episode, which is almost 10 times more than for usual applications). All the other results of each simulation are compared to this defined reference concentration field using several statistical analysis methods.

A simple example is presented in Figure 47 where the left concentration field is taken as a *reference.* The quality of the right concentration field can be estimated by comparing it with the left *reference concentration field*. The comparison can be performed by using the proposed evaluation method described in the following paragraphs.

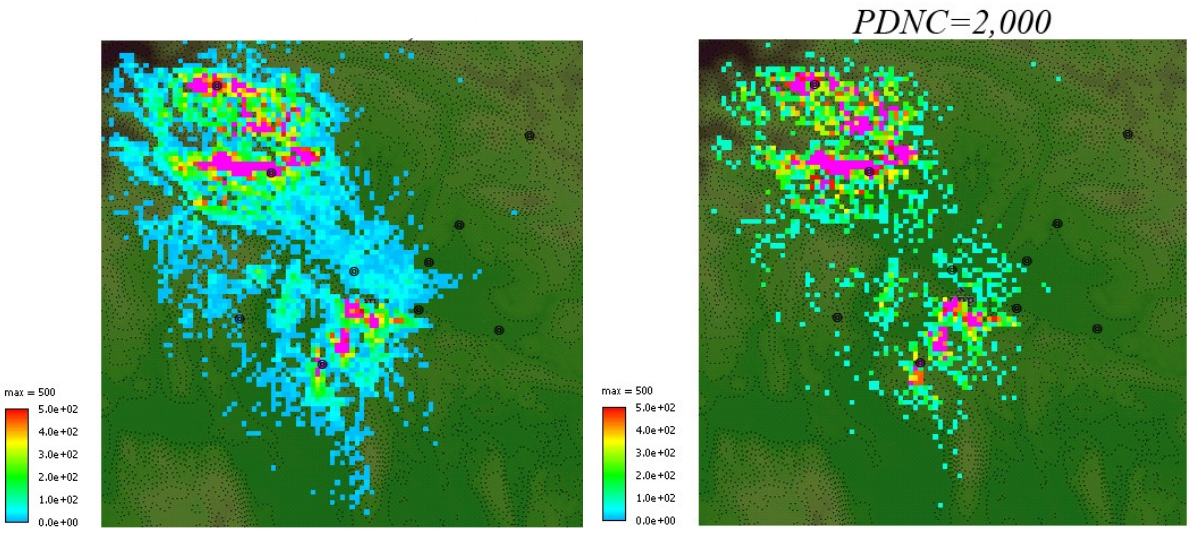


Figure 47: Example of similar 2D ground concentration fields

3.5.1. Correlation coefficient

The correlation coefficient[110] is calculated according to equation (3.2). It is a number between -1 and 1. If there is no relationship between the reconstructed concentration field and the reference concentration field, the absolute value correlation coefficient is 0 or very close to 0. As the strength of the relationship between the reconstructed concentration field and the reference concentration field increases, so does the absolute value of the correlation coefficient. A perfect fit gives a coefficient of 1 and a strong negative linear correlation gives a coefficient of -1. Thus, the higher is the absolute value of the correlation coefficient, the better are the results. It is very important that the absolute values must be compared to when the correlation is evaluated (i.e. A -0.9 correlation is a significantly better correlation than -0.1).

$$r=\frac{N\cdot M\sum_{n=1}^{N}\sum_{m=1}^{M}(C_{n,m,ref}\cdot C_{n,m})-(\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m,ref})(\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m})}{\sqrt{N\cdot M(\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m,ref}^{2})-(\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m,ref})^{2}}\sqrt{N\cdot M(\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m}^{2})-(\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m})^{2}}} \tag{3.2}$$

N ... number of grid cells in east-west direction in 2D plane near the surface
M ... number of grid cells in north-south direction in 2D plane near the surface
$C_{n,m}$ ...concentration at grid cell {m,n} of reconstructed concentration field
$C_{n,m,ref}$ ...concentration at grid cell {m,n} of reference concentration field

3.5.2. Fractional bias

The fractional bias is a measure of the performance recommended by the U.S. EPA. The general expression is given in equation (3.3), as defined in a paper by De Haan[109]. The mean reconstructed concentration is defined by equation (3.4) and the mean reference concentration by equation (3.5). It was selected because it has two desirable features. First, the fractional bias is symmetric and bounded; the values for the fractional bias range between -2.0 (extreme over-prediction) to +2.0 (extreme under-prediction). Second, the fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels. The values that are equal to -0.67 are equivalent to an over-prediction by a factor of two, while the values that are equal to +0.67 are equivalent to an under-prediction by a factor of two. Model predictions with a fractional bias close to 0.0 are relatively free of bias.

$$FB=\frac{(\overline{C}_{ref}-\overline{C})}{0.5(\overline{C}_{ref}+\overline{C})} \tag{3.3}$$

$$\overline{C}_{ref}=\frac{\sum_{n=1}^{N}\sum_{m=1}^{M}C_{n,m,ref}}{N\cdot M} \tag{3.4}$$

$$\overline{C} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} C_{n,m}}{N \cdot M} \tag{3.5}$$

N ... number of grid cells in east-west direction in 2D plane near the surface
M ... number of grid cells in north-south direction in 2D plane near the surface
$\overline{C}$ ...mean reconstructed concentration
$\overline{C}_{ref}$ ...mean refrence concentration

3.5.3. Root mean square error

The root mean square error (RMSE) is a very often used measure of the difference between values predicted by a model or an estimator and the reference values from the phenomenon being modelled or estimated. These differences are also called residuals. The RMSE of an estimated concentration field *C,* with respect to the "true" concentration field C_{true}, is defined as the square root of the mean squared error defined in equation (3.6).

$$RMSE = \sqrt{\frac{1}{N \cdot M}\sum_{n=1}^{N}\sum_{m=1}^{M}(C_{n,m,ref} - C_{n,m})^2} \tag{3.6}$$

N ... number of grid cells in east-west direction in 2D plane near the surface
M ... number of grid cells in north-south direction in 2D plane near the surface
$C_{n,m}$ ...concentration at grid cell {m,n} of reconstructed concentration field
$C_{n,m,ref}$ ...concentration at grid cell {m,n} of reference concentration field

3.6. Determination of acceptable simulation results

The theory of the Lagrangian approach as described in subsection *1.4.6.The Lagrangian particle-dispersion model, and* is based on tracking the flow path of the point-like particles. Each particle that represent a trace species is tracked on its path through the atmosphere with this type of model according to equations (1.3) and (). The particles are moved according to the reconstructed mean wind field and are additionally subjected to the influence of turbulence. The required computational complexity is according to two equations (1.3) and () proportional to the total number of particles used in the simulation because the number of equations to be solved increases with the number of particles.

The quality of the simulation results performed with an AP computer model based on the LPD depends on the number of particles used in simulation. In theory, the use of a large number of particles should ensure good results. In the paper written by Graham and Moyeed[111] an investigation was preformed and an approach was presented to determine how reliable the results from Lagrangian simulations actually are. According to the recommendations from the paper the number of particles must be increased by two orders of magnitude if the precision is to be increased by one order of magnitude. But in practice, the number of particles is limited by the computer resources and the duration of the simulation. A larger number of particles usually needs more time for the simulation and a greater computational cost.

An experiment was made to determine how the AP computer model based on LPD depends on the total number of particles emitted in the air. Several simulation runs were performed for the the same time interval. At each simulation the number of particles was doubled but the meteorological and emission conditions are the same. The absolute number of particles in the domain varies during the simulation because it depends on the emission and meteorological conditions. For the presentations of the results a parameter *PNDC* is defined by equation (3.7), which is used to denote the *Particle Number Density Coefficient*. The defined parameter is proportional to the number of particles *N* emitted during the simulation for a certain emission. When the value of the *PNDC* is high a larger number of lighter particles are emitted (higher resolution), and when the value is low, a smaller number of heavier particles are emitted (lower resolution) for the same amount of emitted specie. The parameter is inversely proportional to the parameter *resolution*, as defined in equation (3.8).

The parameter *resolution* is used in the AP computer model to define the number and the mass of the emitted particles for a certain amount of emitted species (i.e., for the emission of *1kg/s* of SO^2 1 virtual particle is emitted in each time step of the simulation for *resolution=1* or 10 virtual particles are emitted in each time step for *resolution=0.1*).

$$PNDC \propto N \tag{3.7}$$

$$PNDC = \frac{1}{resolution} \tag{3.8}$$

To determine the dependence of the efficiency of the AP computer model according to the particle number density coefficient (*PDNC*) an experiment was made where several simulations are performed on the Šaleška region field data set air-pollution situation that lasted from 1st of April 1991 at 20:00 until 3rd of April 1991 at 00:00. The computational times used for each air-pollution episode reconstruction of each simulation are presented in the graph on the upper part of Figure 48 and on the lower part the dependence of the full time used for each simulation to the number of particles is shown. The presented results show that a very weak dependence of time used on the *PDNC* is observed when a small number of particles is used for the reconstruction, but when the number of particles significantly increases, a strong linear dependency can be determined. In the following figures the time used for each air-pollution episode reconstruction, additionally determined for each of three basic modules of the AP computer model, is shown. The absolute time used within the *Input module* is presented in Figure 49, where no dependency is observed between the time used and the *PDNC*. It is practically constant during all the air-pollution episode reconstructions for all the cases of the *PDNC*. *I*n Figure 50 the time used within the LPD computer model is presented. The results show a strong linear dependency on the time used and the number of particles used for the air-pollution episode reconstruction. This linear dependency becomes very strongly emphasized in the full time used for the air-pollution episode when a very large number of particles are used for the reconstruction (very large *PDNC*). This shows that the time used for a simulation run linearly depends only on the number of particles. In Figure 51 the dependency of the time used within the *Output module* is presented. The result shows that when a small *PDNC* is used, there is almost no dependency between the time used and the *PDNC*, because the time used is practically constant. But when the number of particles exceeds a certain limit, a strong linear dependency occurs. This linear dependency has a lower impact on the full time for one air-pollution episode reconstruction than the dependency of the LPD computer model because the measured values are one order of magnitude lower.
From the point of view of the time spent for each simulation run an optimal value of *PDNC=2* is selected. The optimal value has been selected according to the time that is spent to reconstruct the air pollution over the area of interest for a time period of one year. For the selected optimal point where 1.53 hours are spent for the reconstruction of one day, approximately 23 days are spent for the reconstruction of one year, which is still acceptable in practice.

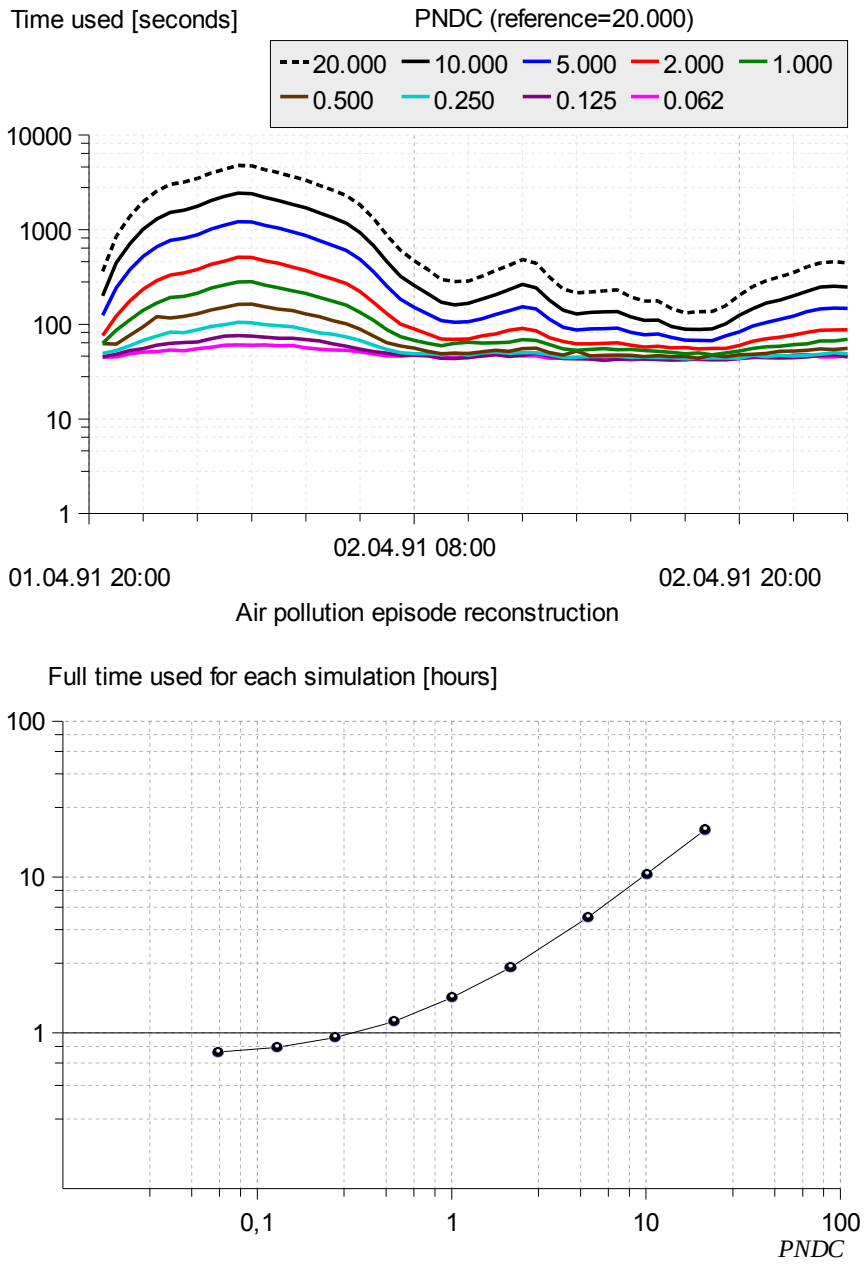


Figure 48: Full time used within the AP computer model (Input module+LPD model+Output module) simulations with different PDNC

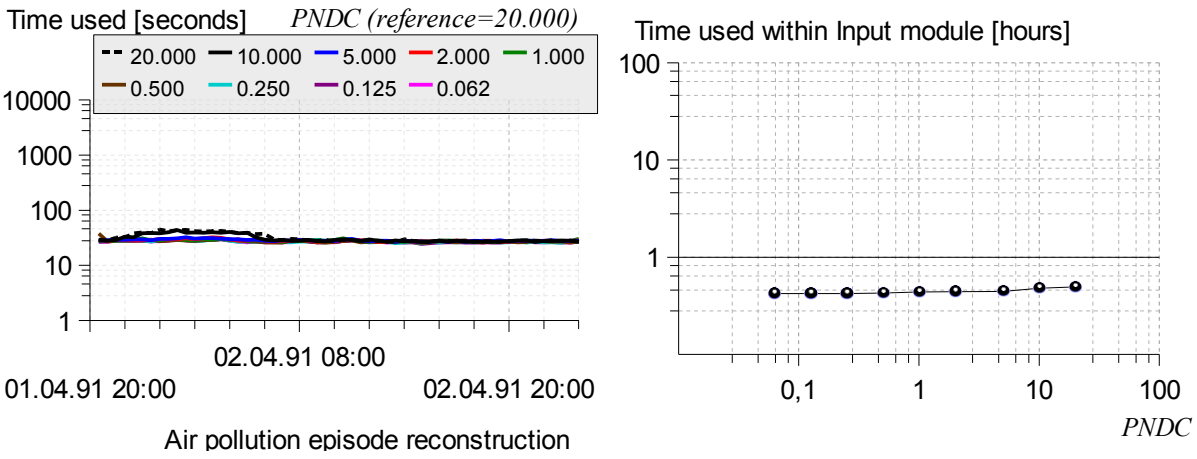


Figure 49: Time used within the Input module for each simulation with different PDNC

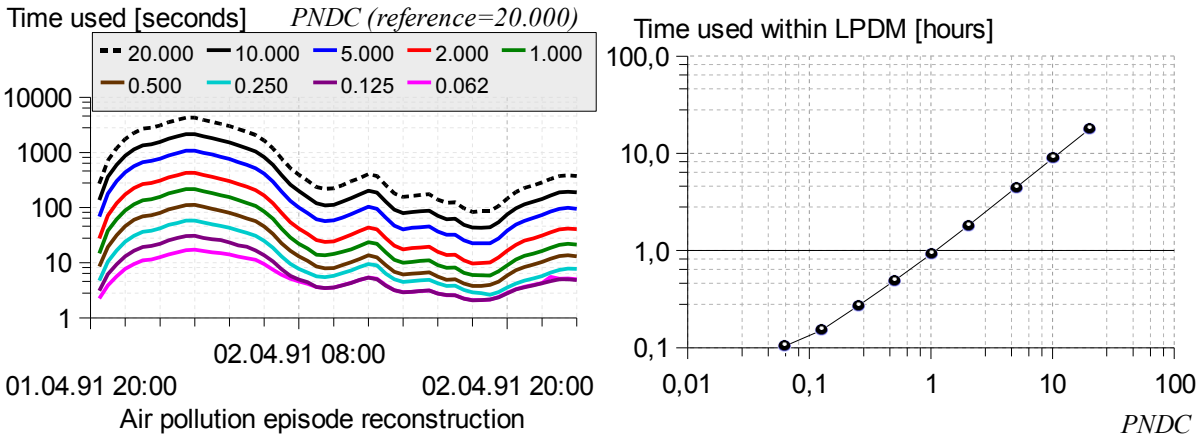


Figure 50: Time used within the computer LPD model for each simulation with different PDNC

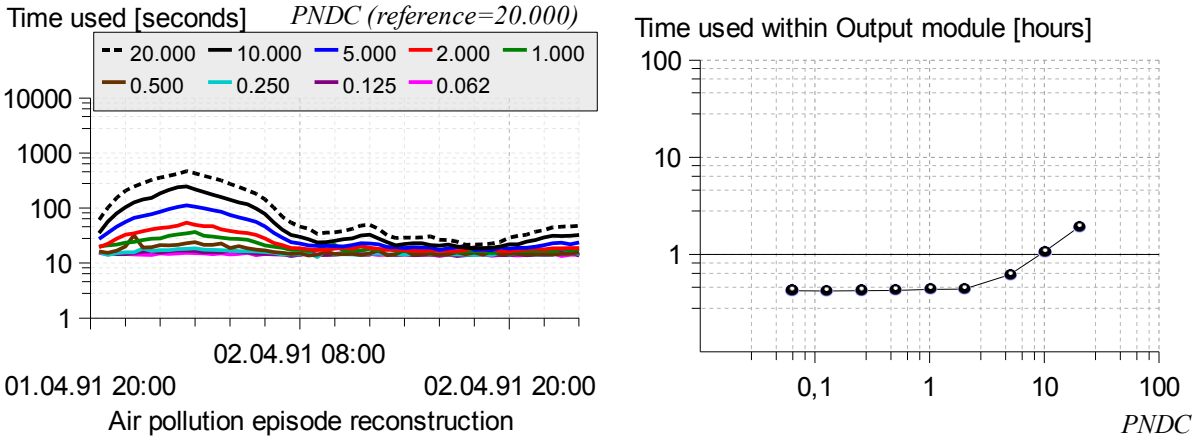


Figure 51: Time used within the Output module for each simulation with different PDNC

An evaluation of each result with a different number of used particles is performed to define the minimum acceptable *PDNC* that is necessary to achieve a good air-pollution reconstruction. Methods that are described in detail in subsection *3.5.Evaluation methods* were used for the evaluations. To estimate the quality of the result of air-pollution episode reconstruction the reference concentration field is defined: a ground level concentration two-dimensional field, which is the result of one air-pollution episode reconstruction with the highest number of used particles where the *PDNC* value was set according to the available computational resources *PDNC=10 (resolution=0.1).* Within this maximum value of the *PDNC* approximately 3,000,000 particles are used in each air-pollution episode reconstruction. All the other results of each air-pollution episode reconstruction are compared to this defined reference concentration field using several statistical analysis methods.

The results of the first comparison are presented in Figure 52 where the c*orrelation coefficient* between the reconstructed concentration field and the reference concentration field exponentially increases to some certain number of particles used in the simulation. When this point is reached, the increase in the number of particles has almost no influence on the quality of the results. So in practice the use of a too large number of particles results in an unnecessary consumption of computational power and wasting of the computational time. It is estimated from the graph presented on the left side of Figure 52 that the results are not improving significantly after the correlation factor is greater than 0.8. So in practice the results that correlate with the original result above the factor of 0.8 are completely acceptable.

The results of the second comparison are presented in Figure 53, where the *root mean square error* between the reconstructed concentration field and the reference concentration field exponentially decreases to some constant value of 0. Again, it is estimated from the graph presented in Figure 53 that the results are not improving significantly after the root mean square reaches some threshold value, which in our example is 15. So the results that are distant from the original result, below the threshold value are acceptable in practice.

The results of the third comparison, where the *fractional bias* is used for the evaluation, are presented in Figure 54. All the average results are almost completely free of bias according to the definition. The fractional bias variability decreases with a higher resolution when the number of particles increases. The value can be used to approximately estimate whether the result is in accordance with the original or is it completely under or over estimated.

From the point of view of the evaluated results the optimal value of *PDNC=5* can be selected. Unfortunately this value is not acceptable in practice. For this value of *PDNC=5* where 3,8 hours are spent for the reconstruction of one day, approximately 58 days are needed for the reconstruction of one year. The time spent for a reconstruction of the air pollution for one year is too high for practical purposes.

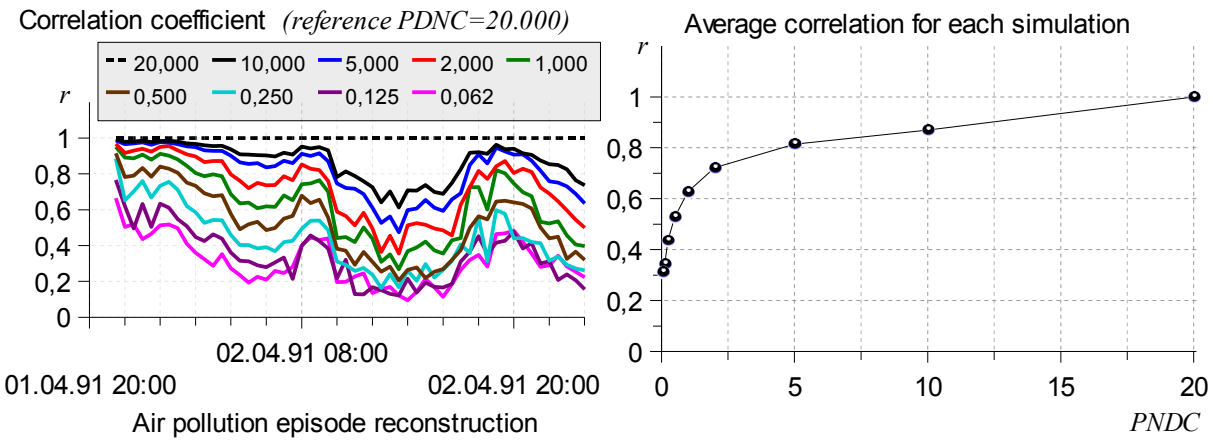


Figure 52: Correlation coefficient

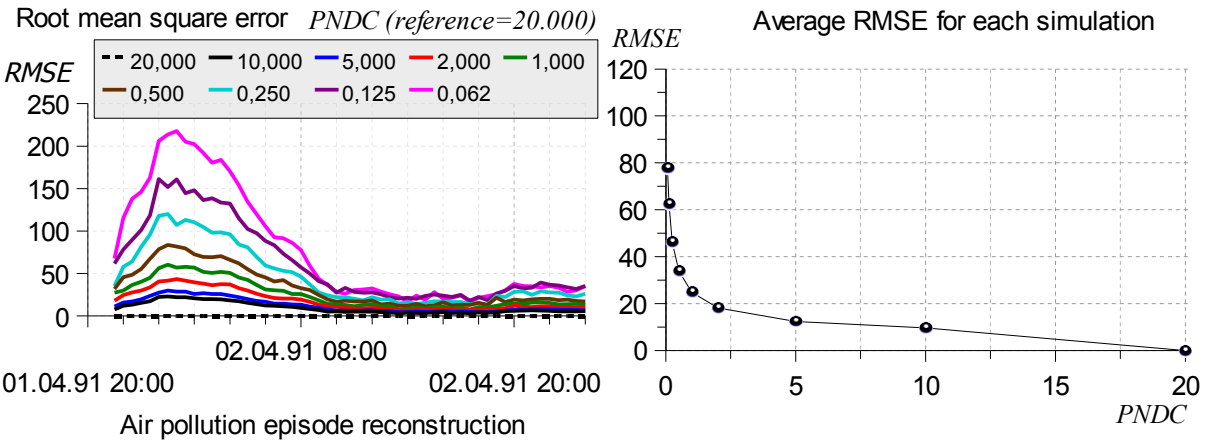


Figure 53: Root mean square error

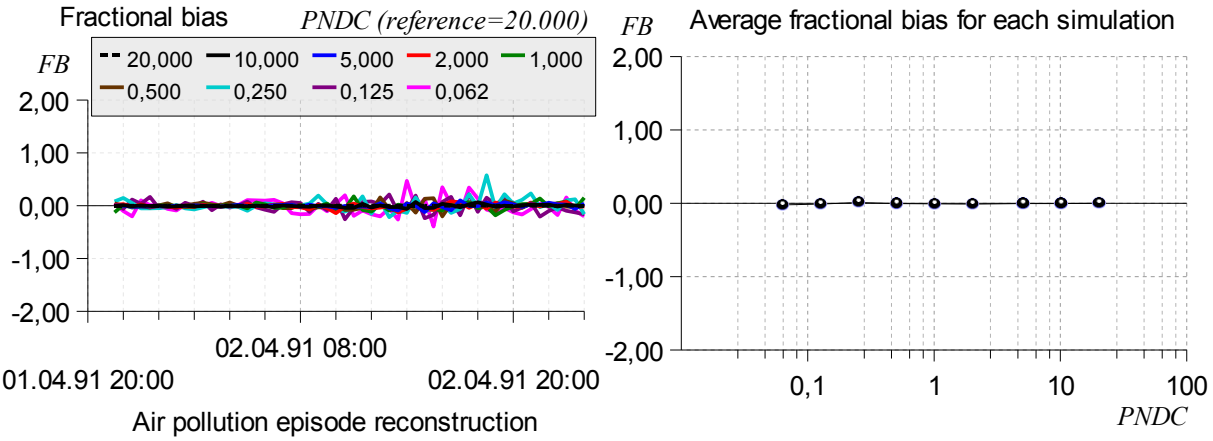


Figure 54: Fractional bias

The ground-level concentration results for one of the reconstructed air-pollution episodes on 1st of April 1991 at 23:30 that are used for the comparisons are presented in Figure 55. In the upper left corner is the result when the highest number of particles is used (reference concentration field, high *PDNC*) in simulation run and in the lower right corner is the result when the smallest number of particles is used (low *PDNC*). It can be seen that with the higher number of particles the ground concentration distribution is more smoothly dispersed than with the smaller number of particles.

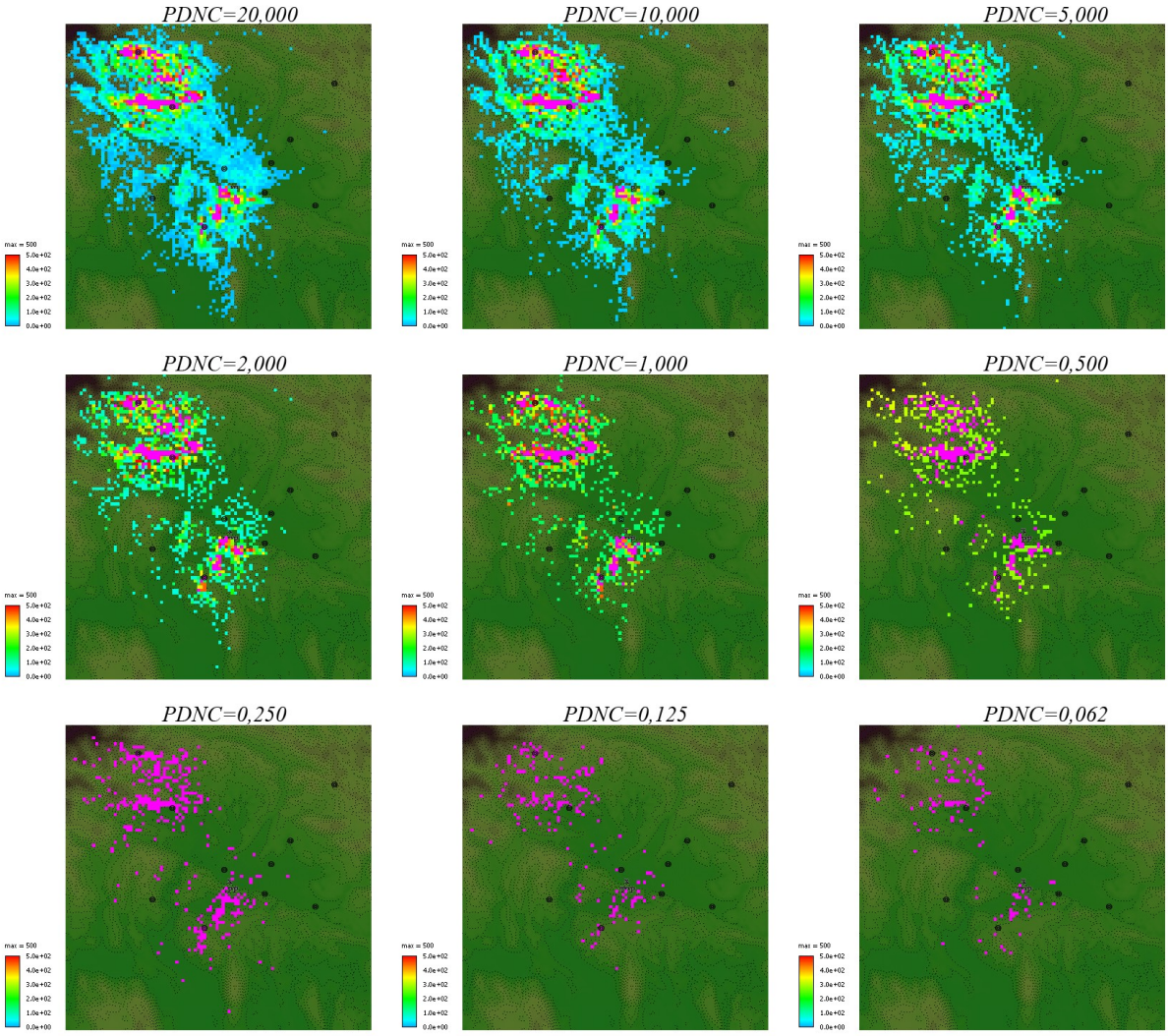


Figure 55: Ground-level concentration results of air-pollution episode reconstruction on 1st of April 1991 at 23:30 used for a comparison

3.7. Air-pollution computer model limitations

In this subsection the computational problems that can appear during the simulation within the AP computer model based on the LPD are demonstrated. The problems are directly related to the occurrence of situations with a large number of active particles. There are two kinds of computational problems that can occur due to the different operational purposes of the AP computer model:

- during the *off-line air-pollution* reconstruction simulations it often happens that the number of active particles exceeds the available *computer resources*
- during the *on-line* air-pollution reconstruction simulations the available *computational time* can be exceeded when a too large number of active particles occurs.

In both listed situations the simulation must be interrupted when the number of particles exceeds the limitations and the reconstruction of the current air-pollution episode is lost. And several proceeding air-pollution episode reconstructions are also corrupted due to the loss of the previous air-pollution situation, especially in low-wind situations when the accumulation of air pollution in the domain occurs.

The problem that occurs around the computational constraints of the AP computer model is illustrated with an example. A situation from the field data set described in subsection *2.1.5.Situation selection from the Šaleška region field data set* that lasted from the 1st of April 1991 at 20:00 until 3rd of April 1991 at 00:00 was selected for a demonstration. For the selected time interval two simulations were performed to reconstruct the air-pollution episodes:

- in the first *original* simulation the maximum allowed number of active particles in the domain is set to a value of 500,000 ,which ensured that the current number of particles in each air-pollution episode is not exceeded,
- in the second *constrained* simulation the maximum allowed number of active particles was set to a value of 200,000, where it is expected that the current number of particles in at least one air-pollution episode is constrained.

The results of both simulations are presented in Figure 56, where the number of active particles of each air-pollution episode reconstruction is compared between the *original* and the *constrained* simulation. The maximum number of active particles of 200,000 is reached at a reconstruction of the air-pollution episode 00:30, where during the reconstruction of the air-pollution episode the emission is interrupted and cut down to keep the number of active particles beyond the limit. In the next two proceeding episodes at 01:00 and 01:30 the situation repeats and the final number of active particles is again constrained according to the restriction on account of cutting down the emission during the simulation. As a consequence, the reconstruction of a proceeding air-pollution episode begins with a corrupted previous state of air pollution in the domain, which is underestimated due to the particle number constraint and part of the information is lost.

To evaluate the results of the constrained simulation several comparisons were performed, where the first comparison is between the unconstrained and the reference simulation (*Original* in blue colour) and the second is between the constrained and the reference simulation (*Constrained* in red colour). The comparison of the correlation coefficient presented in Figure 57 shows that the limitation begins at the reconstruction of the air-pollution episode 00:30, where first significant drop in the correlation is determined. The same result is observed during the root mean square error comparison presented in Figure 58. The fractional bias presented in Figure 59 shows that the results obtained after the limitation are underestimated, which is expected because of the limitation of the emitted particles and consequentially not properly simulating the highest values.

All the comparisons show that all the proceeding reconstructed air-pollution episodes are severely corrupted and the important information is lost. Reconstructed concentrations are lower than they would be obtained in the case where the limitations would not have occurred. Several air-pollution episodes must be reconstructed in the proceeding of the simulation before the lost information is negligible. The problem is even more severe in low wind conditions because several more proceeding episodes must be reconstructed to recover from the constrained situation than in strong wind conditions.

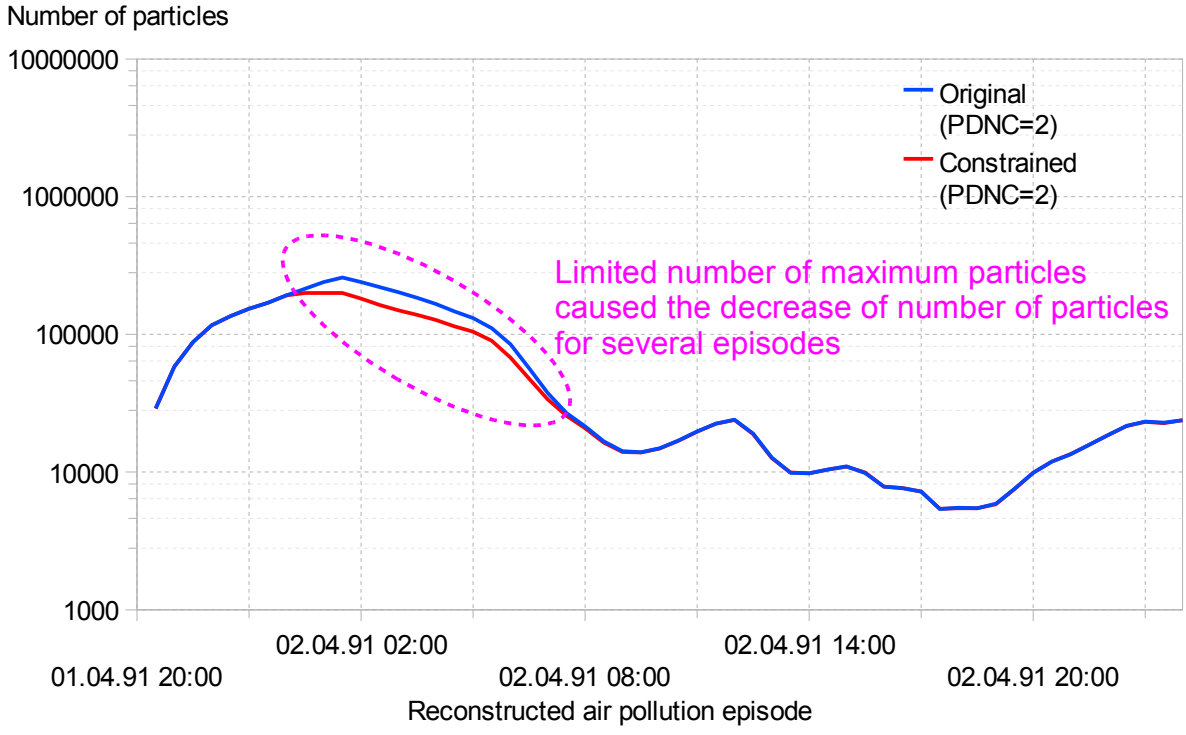


Figure 56: Comparison of the number of particles in the unconstrained and constrained simulation

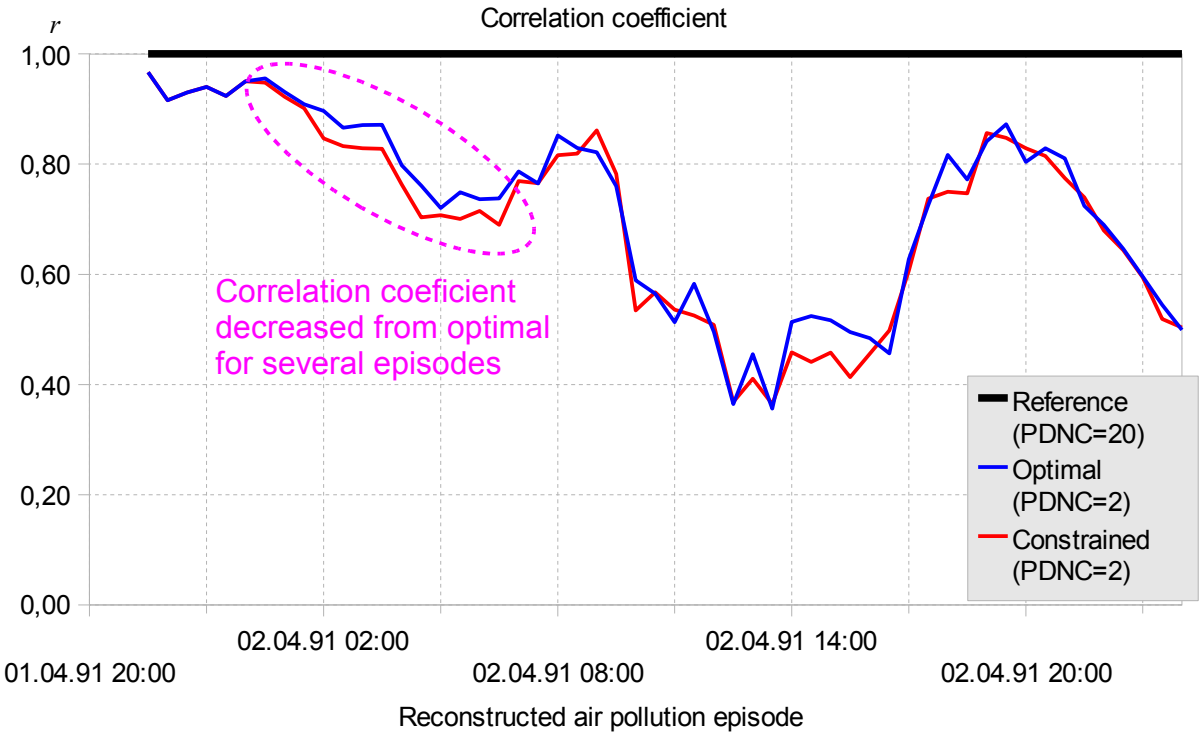


Figure 57: Correlation coefficient of the unconstrained and constrained simulation results compared to the reference simulation results

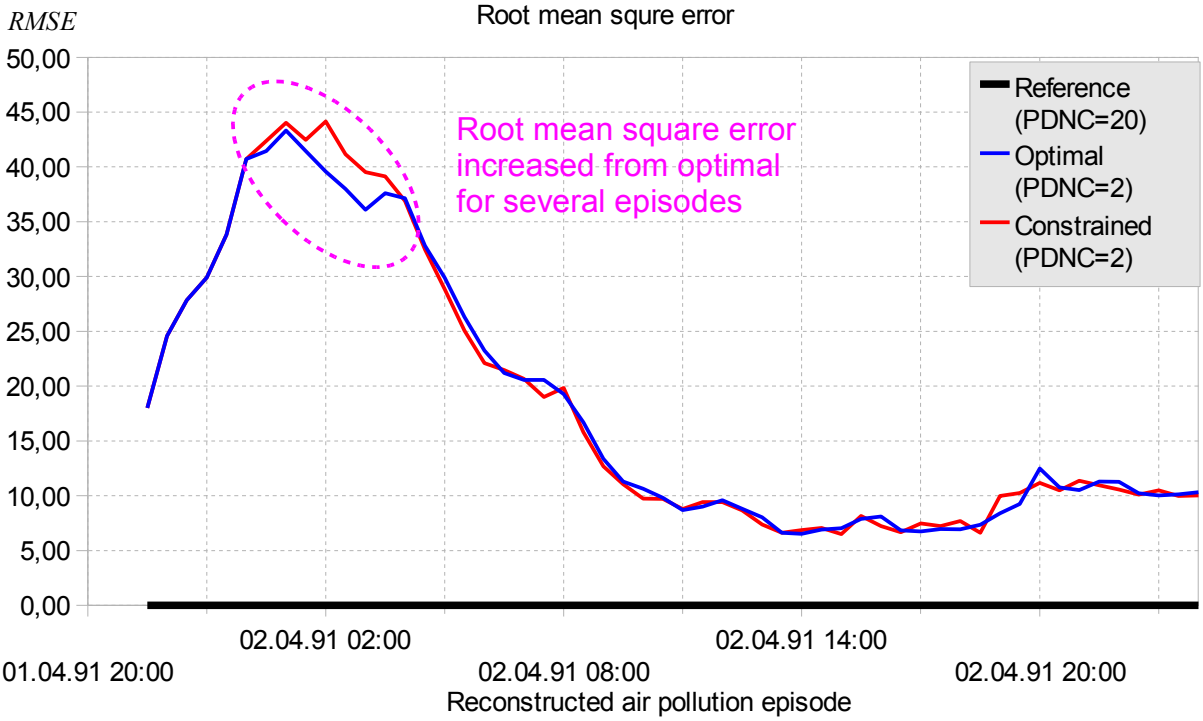


Figure 58: Root mean square error of the unconstrained and constrained simulation results compared to the reference simulation results

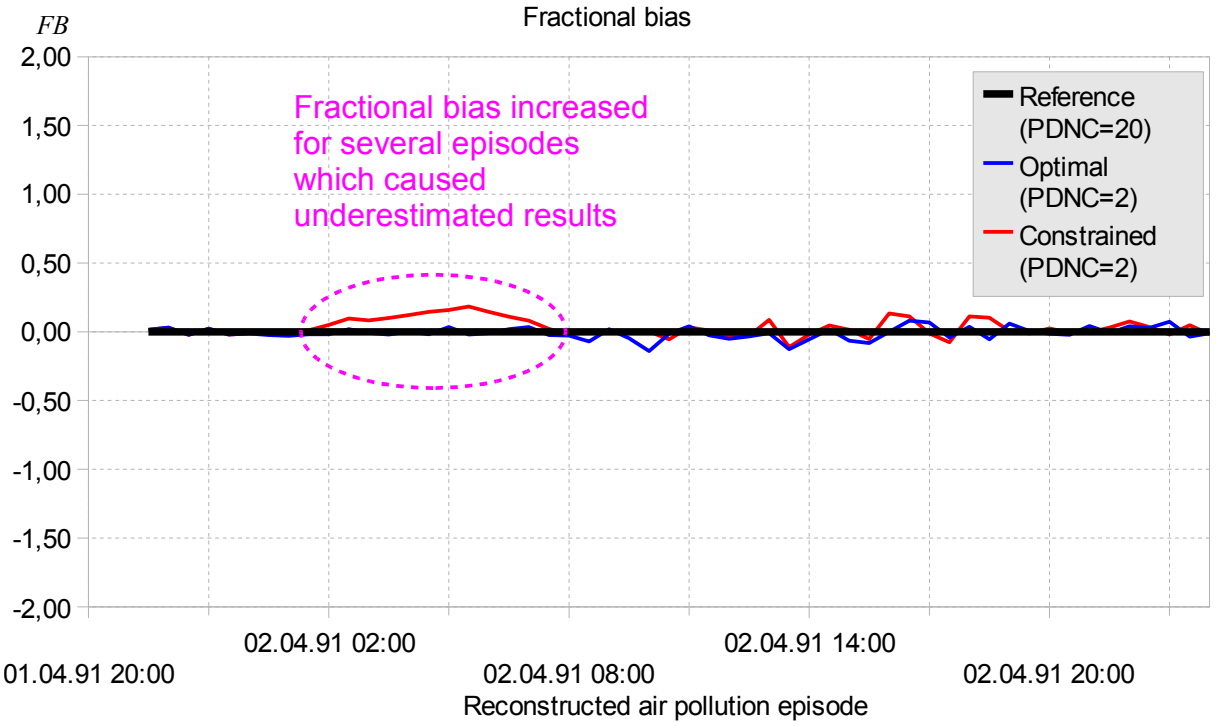


Figure 59: Fractional bias of the unconstrained and constrained results compared to the reference simulation results

4. PROPOSED IMPROVEMENTS IN AIR POLLUTION MODELLING METHODOLOGY

Three new methods to improve the computational efficiency are described in this section. The main advantage of these methods is that the basic physical properties of Lagrangian particle dispersion (LPD) model are not modified at all. All the parameters and methods of the air-pollution modelling methodology are preserved in their original form and no adjustments to the well-developed methodology are required. All the methods are proposed independently and can be separately integrated into the existing air-pollution modelling methodology to improve the computational performances to optimally exploit the available computational capabilities.

The first presented method is clustering, adopted to decrease the computational cost by decreasing the number of particles in the simulations.

The concentration estimation method based on the kernel density represents the second proposed method. It is adapted to substitute for the box-counting concentration estimation method and to improve the poor quality of the results when a smaller number of particles are used in the simulations.

To control the Lagrangian particle dispersion (LPD) and the clustering algorithm a third method is contributed. It consists of two main subsequent methods. In the first step the percentage of lost particles is predicted with the use of an artificial neural network based on the meteorology, the emission and the situation of air pollution at the end of previous episode reconstruction. In the second step the clustering parameters are determined by a decision-making method.

For each contributed method an algorithm is designed that is used for the software development of a computer programme. All the methods are transformed into computer programs to verify their performance. The performance of methods presented in this section are verified on the Šaleška region field data set.

4.1. Clustering method

4.1.1. Introduction

In practice the number of particles is constrained by the computer resources and the acceptable duration of the simulation. A larger number of particles usually results in more time being used for the simulation and to a greater computational cost. The clustering method to decrease the computational cost while preserving the quality of results is described in this section.

The number of particles in the area of interest during the simulation varies. It increases due to the emissions of new particles (species) from a different source in the area and decreases due to different meteorological conditions, especially the wind, which pushes the particles out of the domain. A smaller number of particles is also achieved due to their exposure to dry and wet deposition. In the current AP computer model the user can influence the number of

particles only through the emission from sources where the same emission can be simulated with a smaller number of particles by increasing the weight of the emitted particles. From the presented experiment in the previous subsection *3.6.Determination of acceptable simulation results* it was determined that this can be performed only to a certain level. When that level is reached the quality of the results begins to decline drastically.

During the first simulation experiments an idea occurred to control the maximum number of particles in the domain by modifying the current air-pollution modelling methodology with the use of clustering methods generally, described in subsection *1.4.8.Clustering*. This means that certain particles are joined according to some rules into new, heavier particles. These new particles are introduced into the domain and the old, lighter particles are removed. The properties of the new particles are composed of the properties of the old, lighter particles. In this way the total number of particles in the domain (the area of interest) is significantly reduced and therefore the reconstructions of the proceeding air-pollution episodes (after the clustering) is significantly less computationally demanding.

The concept of the clustering method is presented in Figure 60, where the conventional clustering is assumed. Within the conventional clustering method concept the number of clusters is equal to the maximum allowed number of particles. The particles are joined together only in clusters, which consist of more than one particle.

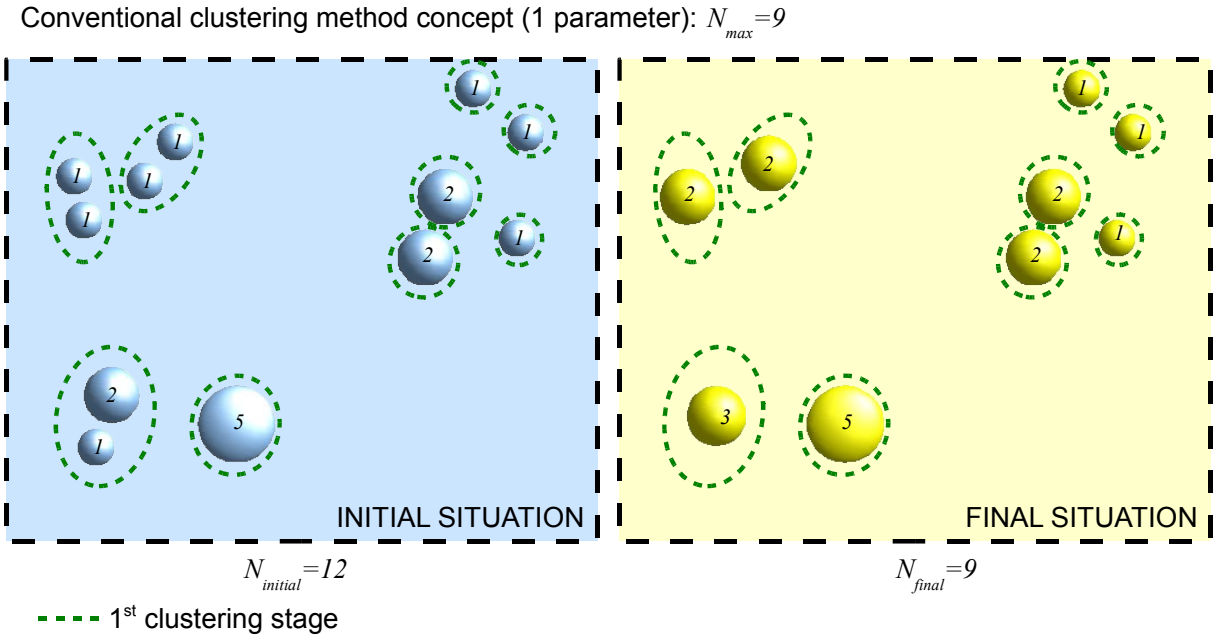


Figure 60: Conventional clustering method concept where initial particles are clustered into eight final clusters defined by one parameter and finally joined into new particles (one for each cluster)

The clustering method for different purposes was already used, by Melheim and presented in his paper,[88] to reduce the number of equations to be solved in particle dynamics. The main idea is that only particles that interact or may interact during the next global time-step are

integrated simultaneously, while the particles that are far from the other particles can be integrated alone. The cluster integration method was proposed in order to perform numerical simulations of collisional particle dynamics in the Lagrangian framework. The clusters of particles were made of particles that interact or may interact during the next global time-step. The cluster integration method was applied to the sedimentation of 5000 particles in two-dimensional box. A significant speed-up was reported at a factor of three orders of magnitude in the dilute regime and two orders of magnitude in the dense regime.

Clustering methods were proved to be a very effective tool also for environmental applications by Mlakar and Božnar[50] in a cluster analysis of wind fields and SO_2 concentrations based on the Kohonen neural network. The method was used for compressing a huge database of wind measurements in order to find correlations between winds and pollutant concentrations.

According to the recommendations from the literature[71], available clustering tools and practical experiences[50] a clustering software has been developed that incorporates two methods: the Kohonen neural network or the K-means clustering algorithm. A software has been developed in the C++ programming language using Qt[112] development tools and an Open Source Clustering library[113] for C++ developed by the Laboratory of DNA Information Analysis, Human Genome Centre, Institute of Medical Science, University of Tokyo.

4.1.2. Clustering criteria

The term feature is used for observations, data items or feature vectors. In this case the feature is a property of particles where each particle is described by its mass, position and velocity. The features must be defined to be used in a clustering process. Feature selection is a process where the optimal subset of features is selected. This can be achieved by eliminating those features that are redundant or do not contain enough relevant information for the clustering process[114]. According to the procedures described in the literature[32,57] a heuristic determination of the features for clustering is made. For a heuristic determination of the features all the facts about the patters and the task domain must be known in advance. In our case detailed information about the particles is given in the user's manual of the selected AP computer model *AriaIndustrie,* which is based on the LPD computer model *Spray*[103].

From preliminary tests of the used clustering tools on large sets of data it was determined that not more than four features should be used in the clustering process because of the constrained computational resources and the available computational time. In our case it is expected that the time used for the clustering should be smaller than the additional time used for the air-pollution episode reconstruction with the original particles in comparison to the clustered particles. In contrast it is the only reasonable solution to perform simulations with a large number of particles. From the set of all the features several features were determined to have less relevant information for the clustering process:

- the mass of each species, because all particles have the same mass when they are

introduced into domain through emission.

- "plume rise" parameters because they are related to the 'plume rise' effect of the hot plume to which the particle initially belongs. After a certain time the plume-rise mechanism is switched off and the properties become irrelevant for the simulation. The particles that are emitted at the same time have equal values of "plume rise" parameters, so the "plume rise" parameters and age are practically redundant.
- the velocity of the particle because the particles that are relatively close to each other also have similar velocity that depends on the velocity of wind. So the velocity and position are practically redundant.

So, finally, the selected features are:

- x, y and z position in the domain
- age a of the particle.

To compare the similarity of a input pattern and a cluster centre (centroid or weight vector of the node) a distance measure must be defined. According to recommendations from the literature[71] the Euclidean distance is selected as defined in equation (4.1).

$$d_{i,c} = \sqrt{\|x_i - x_c\|^2 + \|y_i - y_c\|^2 + \|z_i - z_c\|^2 + \|a_i - a_c\|^2} \tag{4.1}$$

$d_{i,c}$ ...distance between particle i and cluster centre c
x_i, y_i, z_i ...position of particle i
a_i ...age of particle i
x_c, y_c, z_c ...position components of cluster centre j
a_c ...age component of cluster centre j

4.1.3. Implementation

After the feature selection has been completed a clustering software is being developed and designed to be integrated into the existing LPD computer model. At the end of each air-pollution episode reconstruction a file is created where the final air-pollution situation is described. In this file the position and properties of each particle at the end of simulation run are saved. This file is used as the input into the proceeding air-pollution episode reconstruction to define the initial state of the air pollution at the beginning of the next reconstruction. The clustering software is developed to intercept this file, reduce the number of particles in this file with the clustering method, and pass the modified file back to the LPD computer model for the next proceeding air-pollution episode reconstruction.

In the file where the air-pollution situation is saved, each particle is described by several properties:

- x, y and z position in the domain,
- age a of the particle,
- "plume rise" parameters,

- mass of the particle,
- velocity of the particle.

When the integration has been completed a first experiment is made to analyse the new computational performance of the clustering application and modified LPD computer model with an integrated clustering application. Figure 61 presents highlighted modification of the LPD computer model scheme that is presented in previous section in Figure 45.

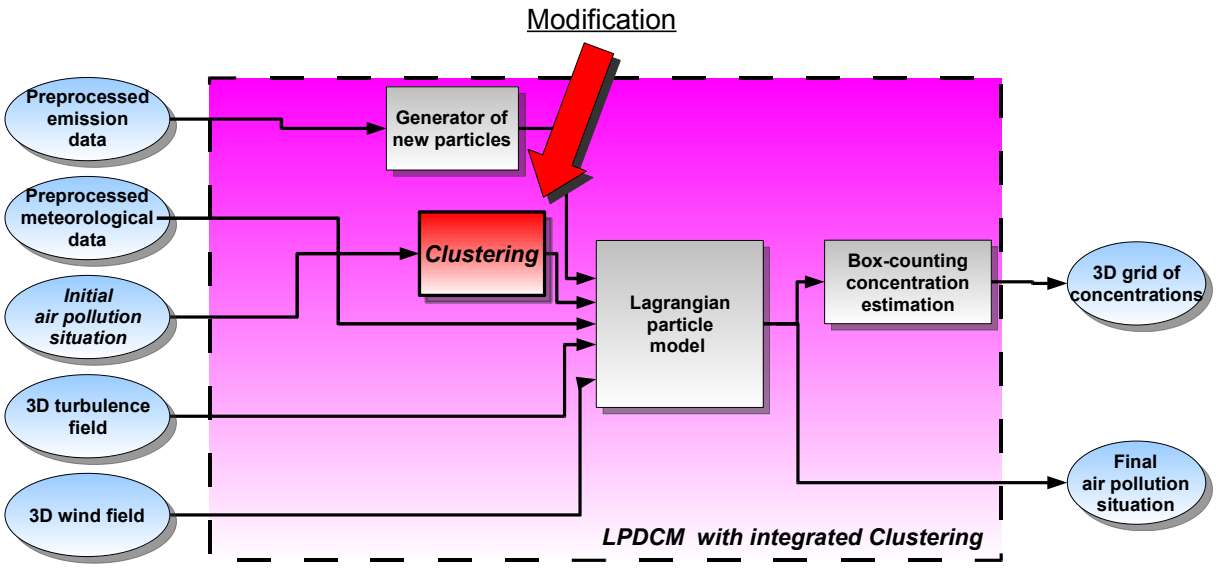


Figure 61: LPD computer model scheme with integrated clustering

In this experimental simulation an air-pollution situation containing about 200,000 particles is used. This large set of particles is clustered several times into a different number of clusters. The time used for clustering with both used clustering methods according to the final number of clusters is presented on a graph depicted in Figure 62. The results of the clustering where the SOM algorithm was used are presented with a green colour and the right results where the K-MEANS algorithm was used are coloured in red. Additional parameters for the SOM algorithm were set to ensure reliable results: the number of iterations to $N_{iter}=50$ and the initial learning rate $\alpha=0.02$. No additional parameters were set to the K-MEANS algorithm. The time used for the clustering is linearly dependent on the number of clusters within the SOM algorithm and exponentially is dependent within the K-MEANS algorithm, which was expected from the theory about both clustering techniques.

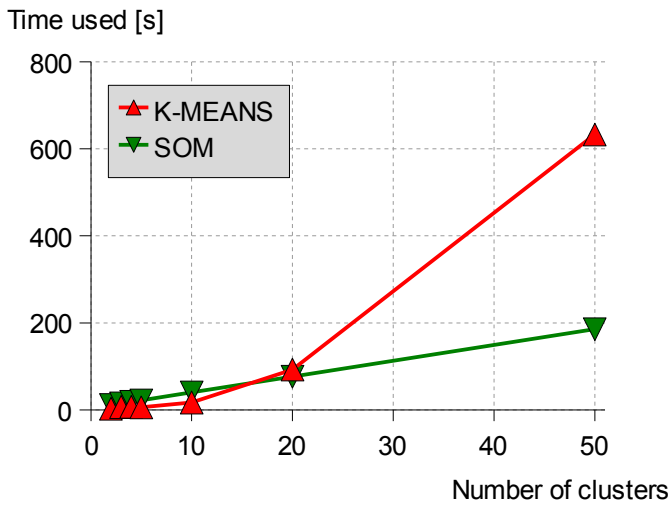


Figure 62: Dependence of the time used for clustering according to the number of clusters

The presented results of the experiment show that for practical purposes the clustering of a large dataset into more than 5 clusters is unsuitable. The main idea was to directly cluster the initial, for example 200,000 particles, into about 20,000 particles, as shown in Figure 60 where the conventional clustering concept is illustrated. This is unachievable in practice, especially when the K-MEANS algorithm is used, while using the SOM algorithm the result can be achieved, but its quality would not be sufficient. To overcome the deficiency of the presented clustering application it will be upgraded, based on the developed hierarchical clustering algorithm described in the proceeding subsection.

4.1.4. Hierarchical clustering method

To achieve a high computational efficiency, the clustering is performed hierarchically, where a large number of particles is clustered into several large clusters. Each large cluster in afterwards sub-clustered into several clusters and so on is the clustering performed until small sub-clusters with particles that really belong together are obtained. At the end of the clustering procedure the obtained small sub-clusters are joined into new particles.

The concept of the hierarchical clustering method is illustrated in Figure 63 using a simple example. In the first stage of clustering all the particles are grouped into two clusters. The particles of each cluster are in the second step again grouped into two clusters. This process of grouping particles of a cluster into two sub-clusters is performed for several clustering stages until the number of particles in each sub-cluster is less than or equal to two particles. Finally, the particles in each sub-cluster are joined into new particles. In this basic method two parameters are used to control the clustering process:

- N_{sub}, sub-cluster count, which defines into how many sub-clusters the current cluster will be separated, for practical purposes the value of this parameter should not exceed a value of 4,

- N_{size}, cluster size, which defines the maximum number of particles that can be joined together (of how many small particles should a new larger particle consist).

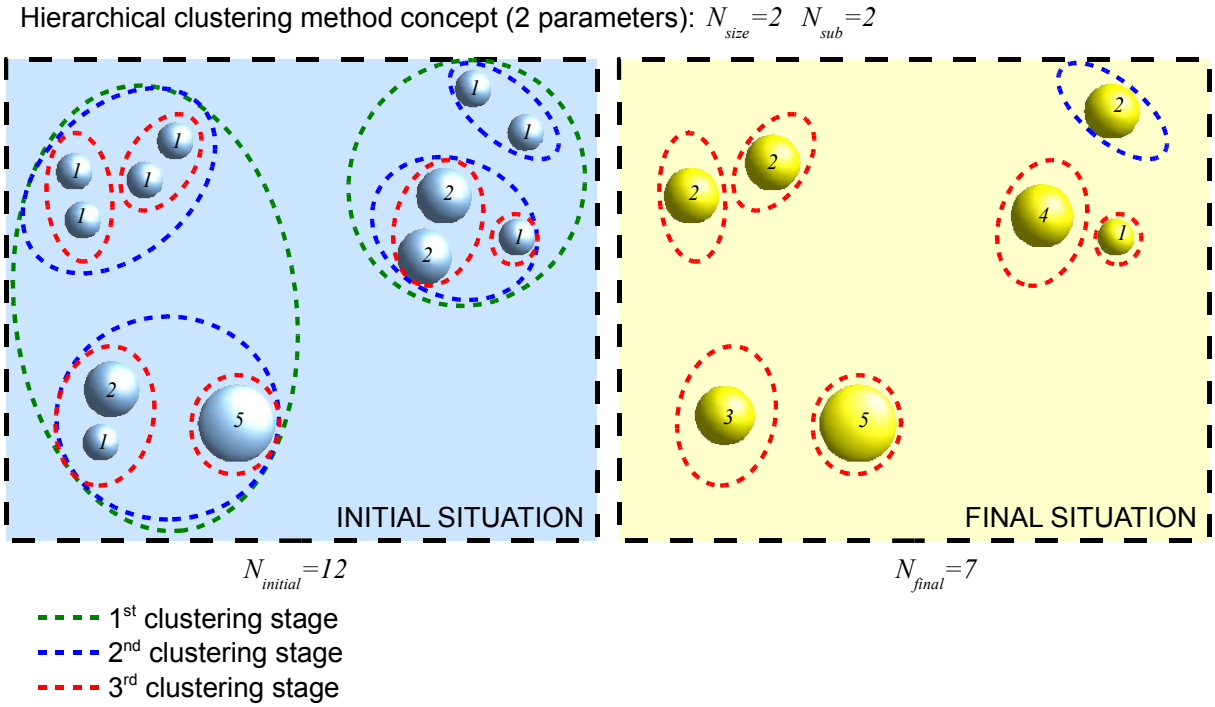


Figure 63: Illustration of the hierarchical clustering method concept where the final number of clusters is obtained in several clustering stages, in each clustering stage the particles in each cluster are clustered into two sub-clusters until the size of the sub-cluster in less than or equal to two

The hierarchical clustering method based on the SOM and K-MEANS clustering algorithms is presented in Figure 64 and the recursive clustering function used in basic method in Figure 65.

In practice, when a large dataset of 200,000 particles is used, it is expected that when the parameters are, for example:

- N_{sub}=2; N_{size}=2; the result consists of approximately 100,000 particles,
- N_{sub}=3; N_{size}=2; the result consists again of approximately 100,000 particles, the computational time is shorter, but more memory resources are used,
- N_{sub}=2; N_{size}=4; the result consists of approximately 50,000 particles.

The term approximately is used because the final size of the dataset is not dived exactly by a factor of 4 when the N_{size} parameter is 4. The N_{size} parameter defines only the upper limit of the particles that a new particle consist of, which means that the new particle can be constructed from 1, 2, 3 or 4 particles in this example. When, for example, two particles are far away from other but close enough one to another, the clustering algorithm on the first level these two particles joins into a new particle, while the remaining dataset of particles is being

additionally clustered. So, the final number of particles can only be approximately estimated, because the exact final number of particles depends on the current distribution of the particles in the domain.

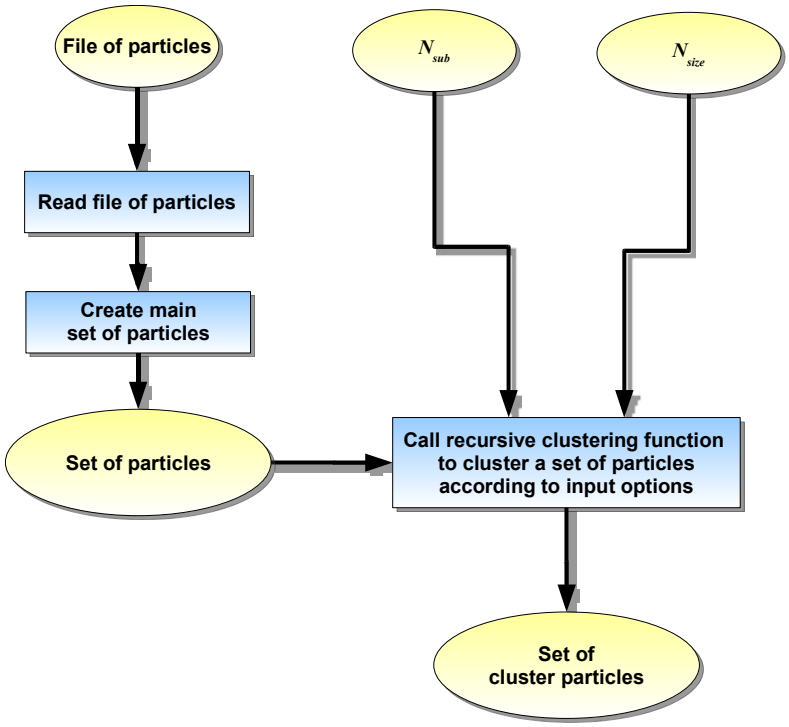


Figure 64: Flow chart of the hierarchical clustering-method implementation where particles are clustered by calling the recursive clustering function and two parameters are used

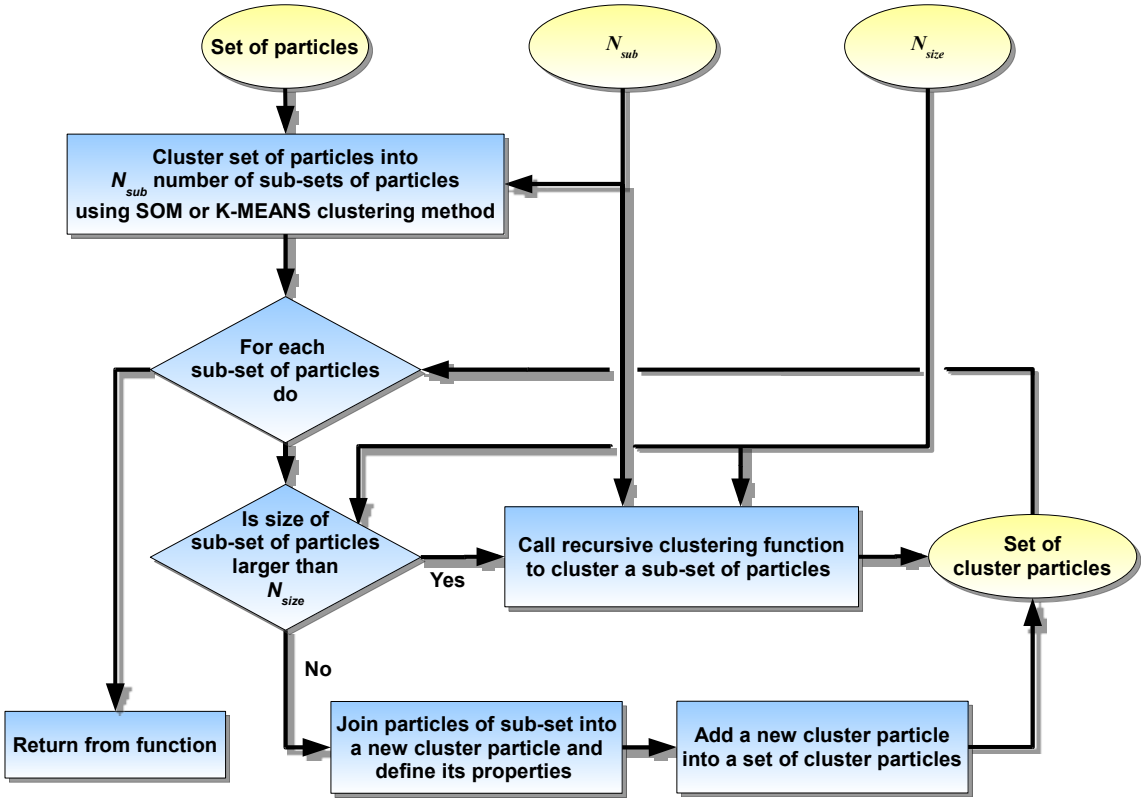


Figure 65: Flow chart of the recursive clustering function implementation which is called recursively until the number of particles in obtained sub-clusters is less than N_{size}

Based on the presented hierarchical clustering method a software is designed and integrated into the LPD computer model for the validation of the method. Several simulations with different clustering parameters have been performed to test the performance of the developed method and to determine the influence of the clustering algorithm on the quality of the results. To determine the quality of the results a comparison to the reference ground-level concentration fields are made according to the developed methods described in detail in subsection *3.5.Evaluation methods*. For all the simulations only the results obtained by the K-MEANS clustering algorithm are presented because the results obtained by the SOM are practically the same.

In Figure 66 the result of the time used within the Lagrangian particle-dispersion computer model (*LPDM)* and in Figure 67 the time used within the *Clustering module* are presented for air-pollution episode reconstructions with the different clustering parameters N_{sub} and N_{size}. In all the presented simulations the particles have been emitted with the parameter *PDNC=2*. Comparisons show that with the increase of the values of the clustering parameters, the time used for the simulation and the clustering decreases. The results also show that the time used for the clustering only depends on the dispersion of the particles and very little on the number of particles. The two peaks in Figure 67 are the results of changing the meteorological conditions when the wind started to change its direction and the particles become widely dispersed over the domain, as presented in Figure 26 from the subsection *2.1.5.Situation selection from the Šaleška region field data set*.

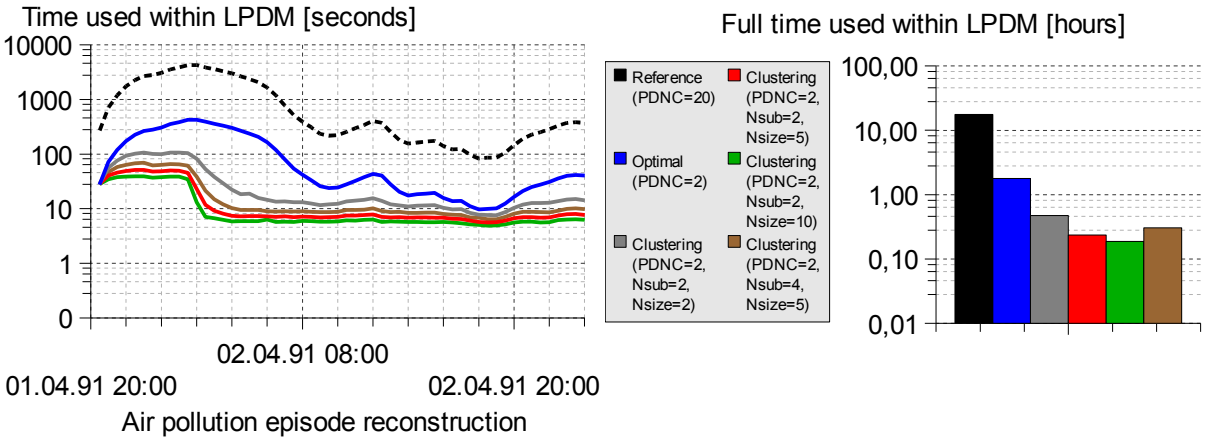


Figure 66: Time used within the LPDM according to the different clustering parameters (N_{sub}=variable, N_{size}=variable)

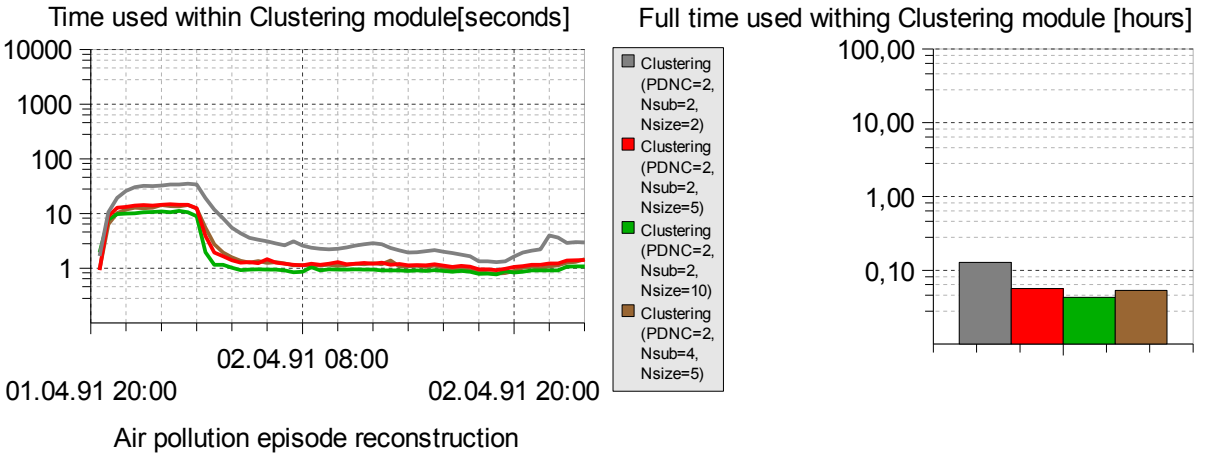


Figure 67: Time used within the clustering module according to the different clustering parameters (N_{sub}=variable, N_{size}=variable)

In Figure 68 a correlation factor, in Figure 69 a root mean square error and in Figure 70 a fractional bias are presented. Qualitative comparisons to the reference concentration field show that the results are more correlated when certain meteorological conditions appear in the domain. The trend of the correlation coefficient of the simulations performed with clustering compared with the correlation coefficient trend of the optimal simulation is strongly correlated. The correlation coefficient of the clustering simulations is much worse because the number of particles is strongly reduced due to the clustering algorithm. Also, the results of the root mean square error and the fractional bias are significantly worse, especially for the simulations where the N_{size} and N_{sub} parameters have lower values. The ground level concentrations for several intervals where the correlation between the results is good are presented in Figure 71, where a relatively strong wind from the south-east is present. Fewer correlated results are obtained when other meteorological conditions appeared. Several intervals where the correlation is poor are presented in Figure 72, where calm weather occurred and the air pollution started to accumulate in the domain. On both presentations in Figure 71 and in Figure 72 it is clear that the number of particles is approximately inversely

proportional to the parameter N_{size}.

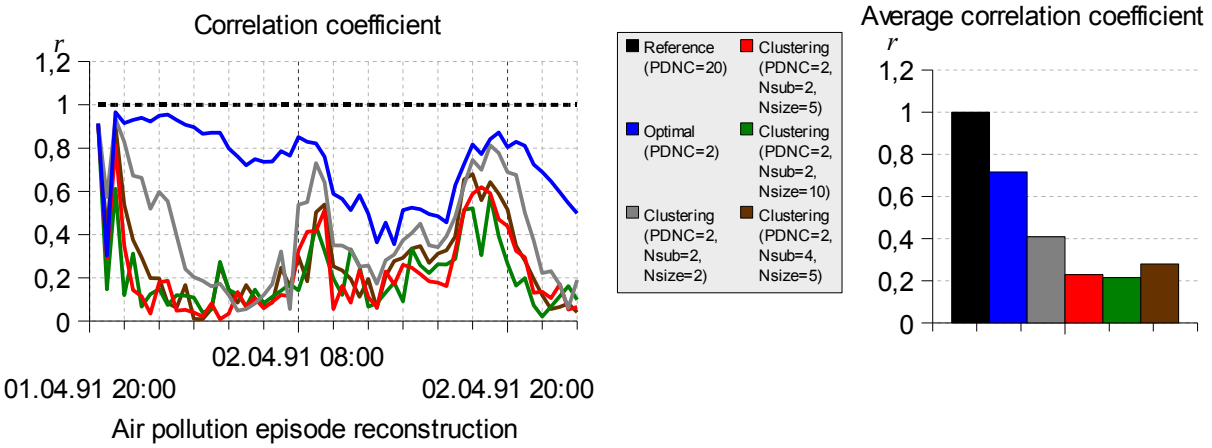


Figure 68: Correlation coefficient for different clustering parameters

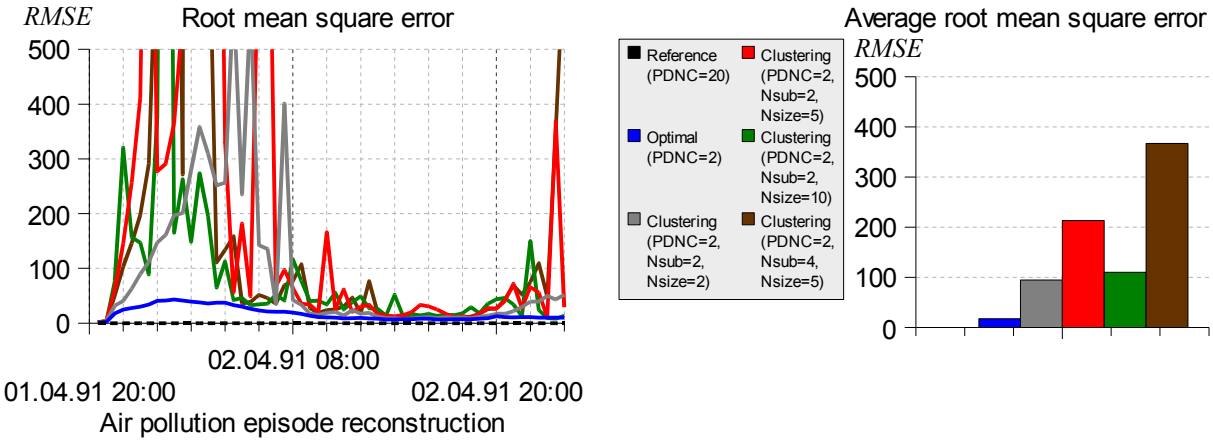


Figure 69: Root mean square error for different clustering parameters

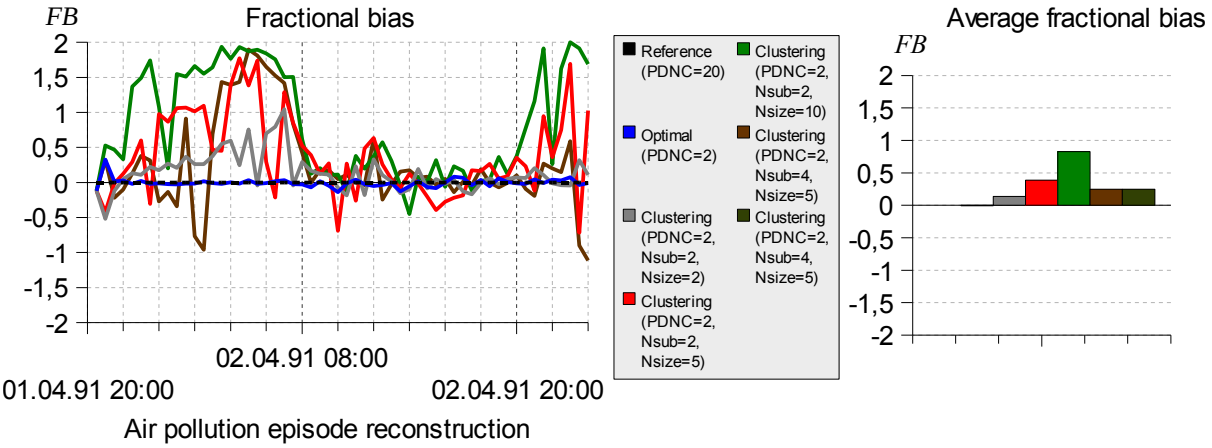


Figure 70: Fractional bias for different clustering parameters

Comparisons show that when the parameter N_{size} increases, the time used for the simulation and clustering decreases. This is because the number of particles in the simulation becomes increasingly reduced when more particles are joined together. With the increase of the parameter N_{size} the quality of the results also decreases.

Amplification of parameter N_{sub} reflects in an increase of the used computational time. This is expected because of the results presented in Figure 62, where it is determined that clustering in more clusters results in more computational time being used.

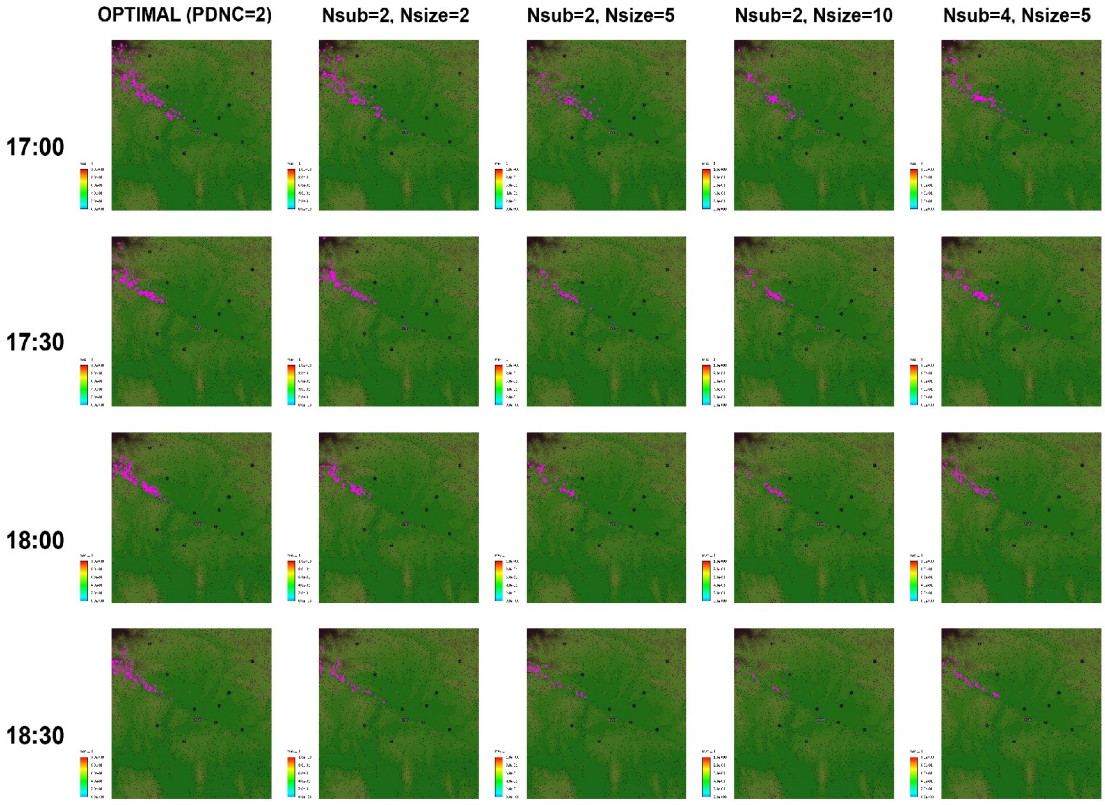


Figure 71: Good correlation between the original and hierarchical clustering results

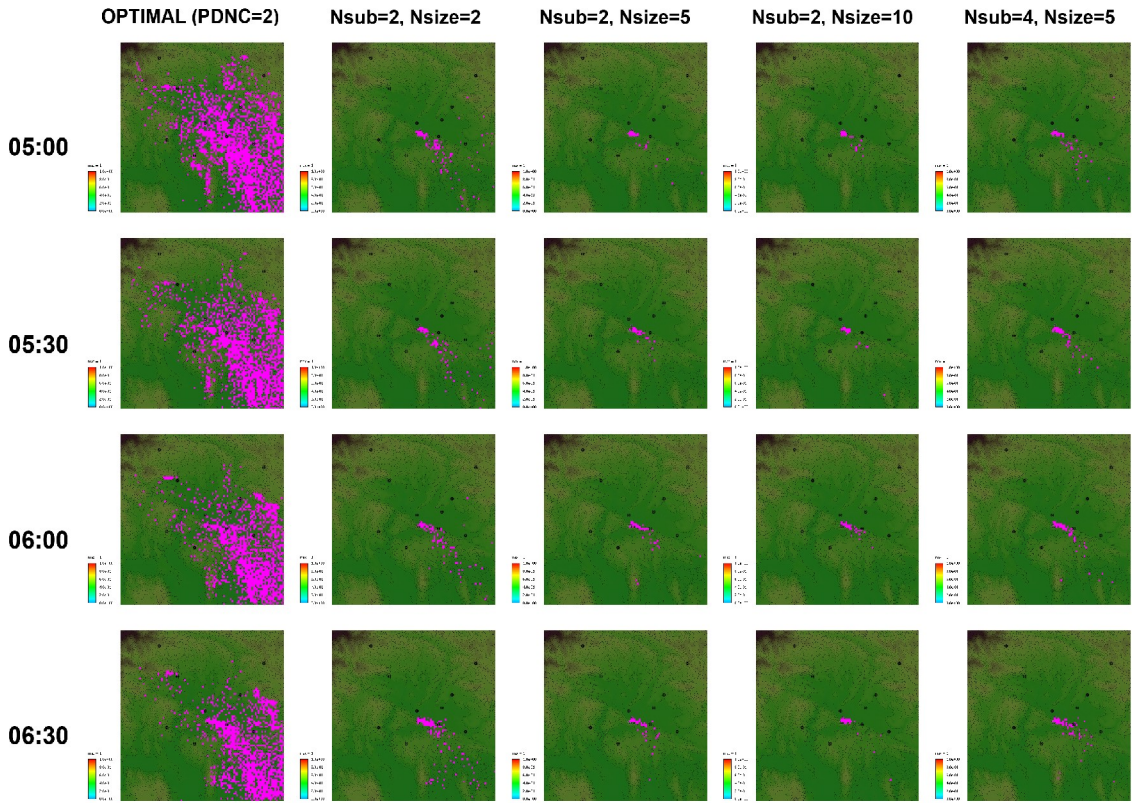


Figure 72: Poor correlation between the original and hierarchical clustering results

4.1.5. Hierarchical clustering method with additional parameters

Previously obtained results show that the reduction of particles by this clustering is too intense, which is reflected in the poor quality of the results. So, the next idea is to reduce after each air-pollution episode reconstruction the number of particles in the domain only to some defined maximum number. The clustering method is improved by adding the additional parameter N_{max}, which defines the upper limit of the final number of particles. If the current number of particles is not exceeded, none of the particles is joined and replaced with a new particle. In this new, advanced algorithm only those clusters of particles are joined into a new particle, which are closer to each other according to a distance measure.

Another additional parameter m_{max} is introduced into the new hierarchical clustering method with additional parameters that excludes the particles from clustering that exceed a certain mass. This parameter is provided to exclude the hypothetical growth of very large and heavy particles, which can lead to physical nonsenses.

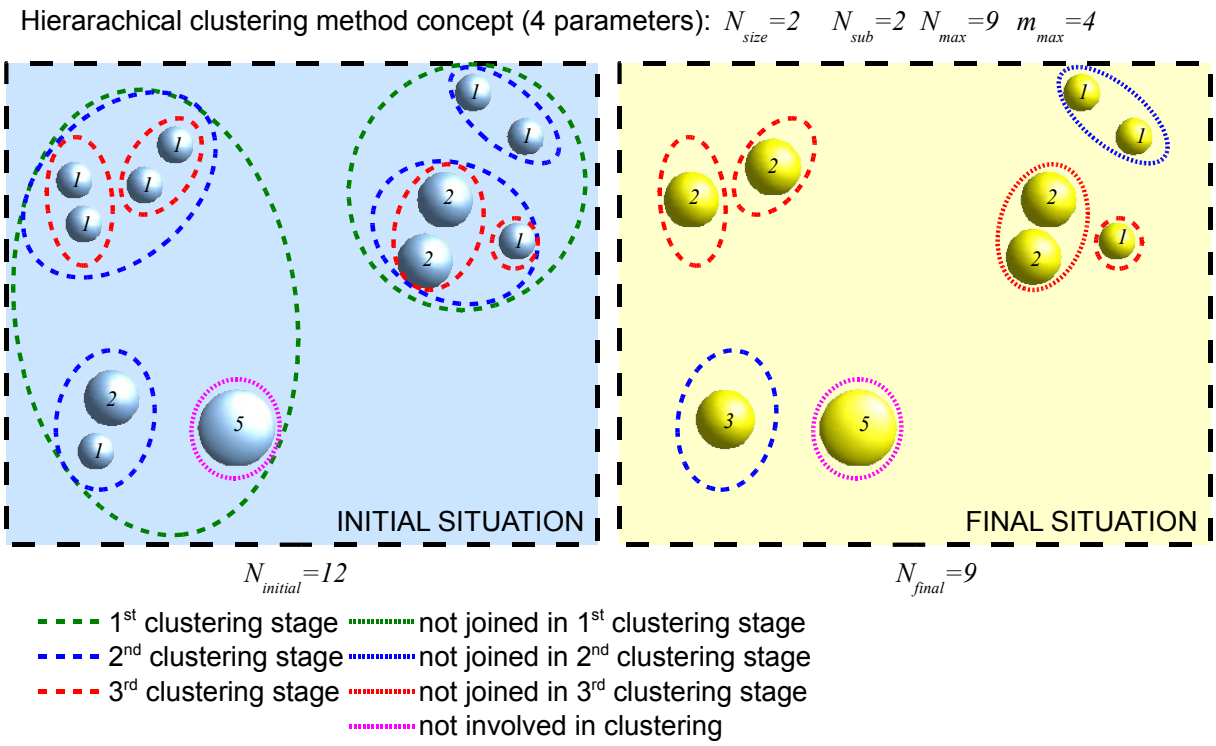


Figure 73: Illustration of the hierarchical clustering method concept with additional parameters where the large particle is not involved in clustering and only the closest particles are joined together to satisfy the maximum number of requested final particles

The concept of the hierarchical clustering method with additional parameters is illustrated in Figure 73 using a simple example. Before the clustering begins the particle with a mass larger than 4 is excluded from the clustering process. In the first stage of the clustering all the particles are grouped in two clusters. The particles of each cluster are, in the second step, again grouped into two clusters. This process of grouping particles of a cluster into two sub-clusters is performed for several clustering stages until the number of particles in each sub-cluster in less than or equal to two particles. Finally, only those particles in each sub-cluster are joined into new particles that are close enough according to the defined distance metrics. The number of particles joined together depends on the defined maximum number of final particles. In this example only three pairs of particles are joined together.

The final, hierarchical clustering method with additional parameters is presented in Figure 74, while the recursive clustering function remained the same as that presented in Figure 65.

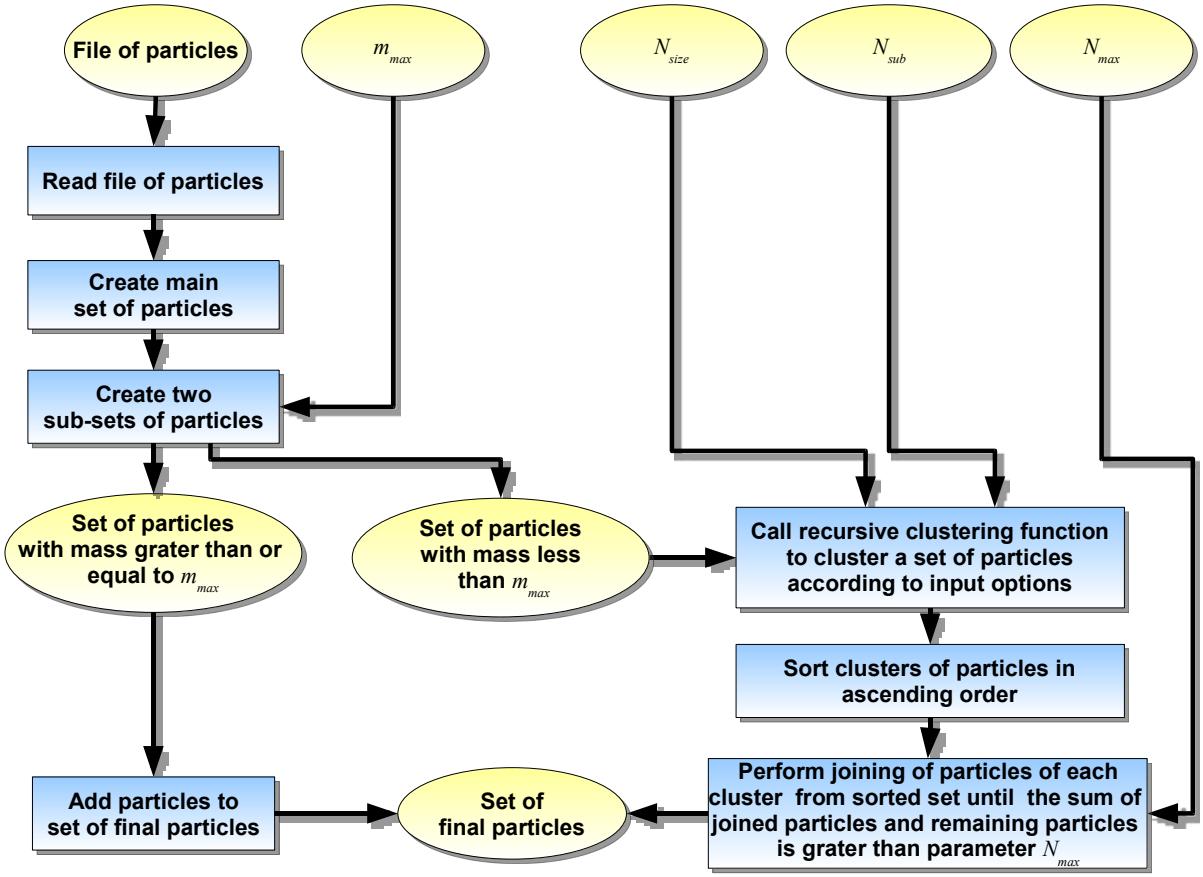


Figure 74: Flow chart of implementation of hierarchical clustering method with additional parameters where parameter m_{max} excludes heavy particles from clustering and parameter N_{max} reduces the number of particles that must be joined into new particles

Based on the presented, hierarchical clustering method with additional parameters, a software has been designed and integrated into the LPD computer model for a validation of the method. With an advanced software clustering the previous experimental simulations have been repeated with constant parameters $N_{size}=2$, $N_{sub}=5$ and $m_{max}=0.1$, while the parameter N_{max} was changing. For all the simulations only the results obtained by the K-MEANS clustering algorithm are presented because the results obtained by the SOM are practically the same.

In Figure 75 a result of the time used within the Lagrangian particle-dispersion computer model (*LPDM)* and in Figure 76 the time used within the *Clustering module* are presented for all the experimental simulations. The obtained results show that the used computational time is inversely proportional to to the parameter N_{max}, which is used to constrain the maximum number of particles in the result of a simulation. But because another parameter m_{max} is introduced into the advanced algorithm to prevent the constitution of abnormally large particles the computational time cannot be significantly reduced with a parameter N_{max} after a certain threshold is reached. Figure 55 shows that the decrease of the parameter N_{max} under values of 25,000 maximum particles does not have any significant effect on the computational time.

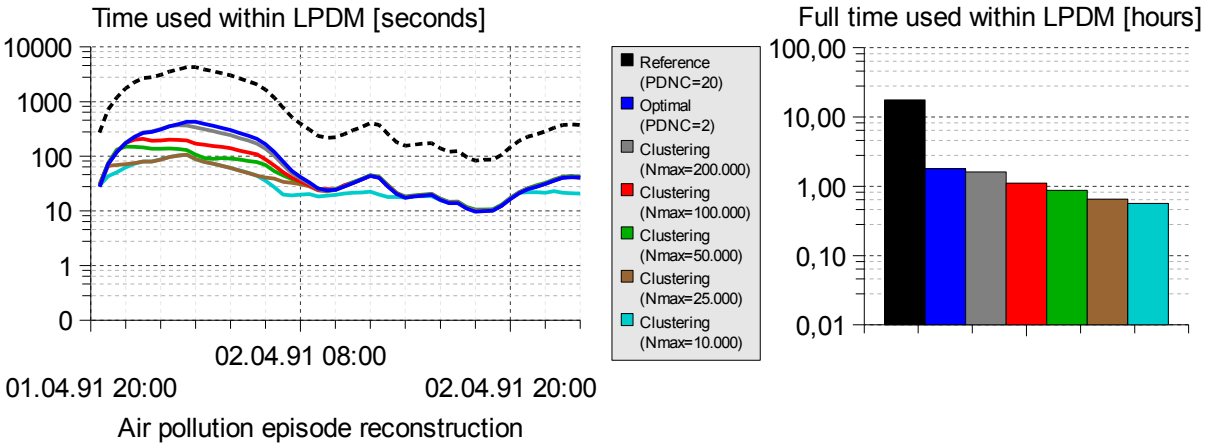


Figure 75: Time used within the LPDM for different clustering parameters (N_{sub}=2, N_{size}=5, m_{max}=0.1, N_{max}=variable)

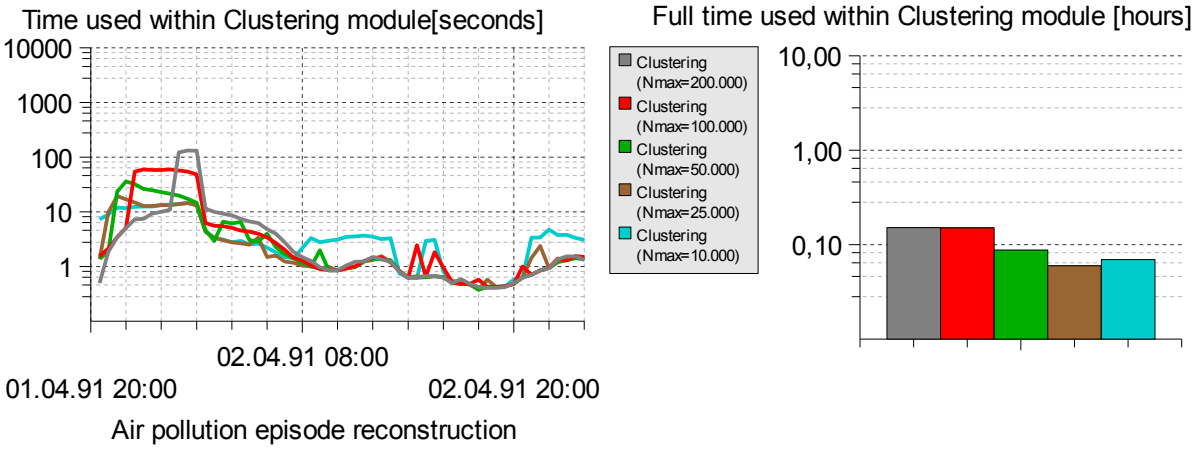


Figure 76: Time used within the clustering module for different clustering parameters (N_{sub}=2, N_{size}=5, m_{max}=0.1, N_{max}=variable)

In Figure 77 a correlation factor, in Figure 78 a root mean square error and in Figure 79 a fractional bias are presented with a different set of parameters N_{max}. The correlation coefficient and the root mean square error are inversely proportional to the parameter N_{max}. The results obtained with the hierarchical clustering algorithm with additional parameters are better correlated to the reference than in the case of using the hierarchical clustering algorithm without additional parameters, mainly due to the parameter m_{max}, which prevents the construction of abnormally large particles. The improvement is also reflected in the fractional bias, which is slightly scattered around zero values.

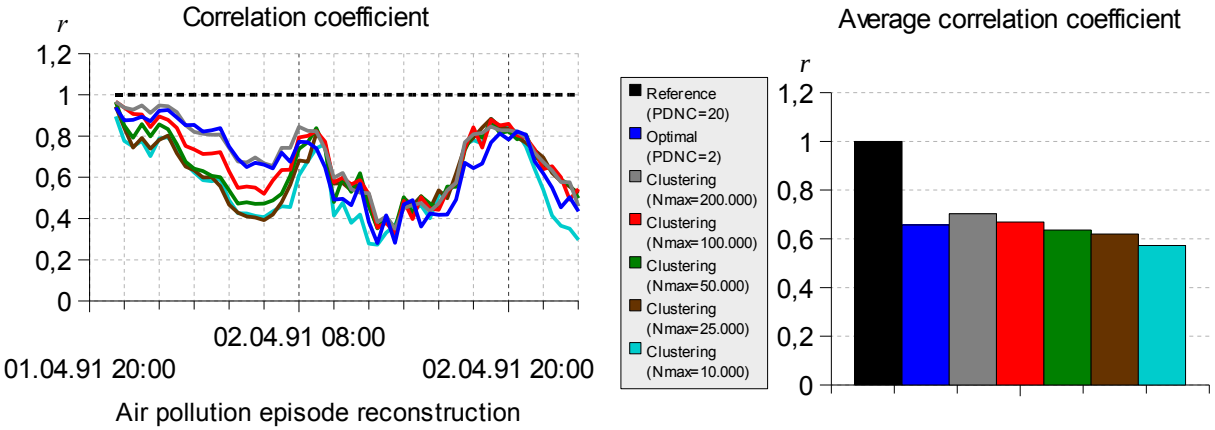


Figure 77: Correlation coefficient for different clustering parameters

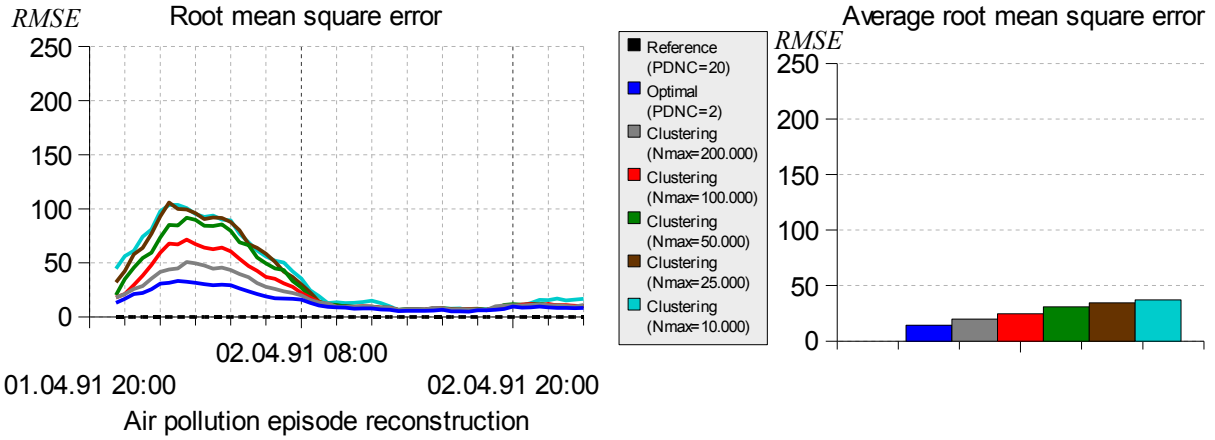


Figure 78: Root mean square error for different clustering parameters

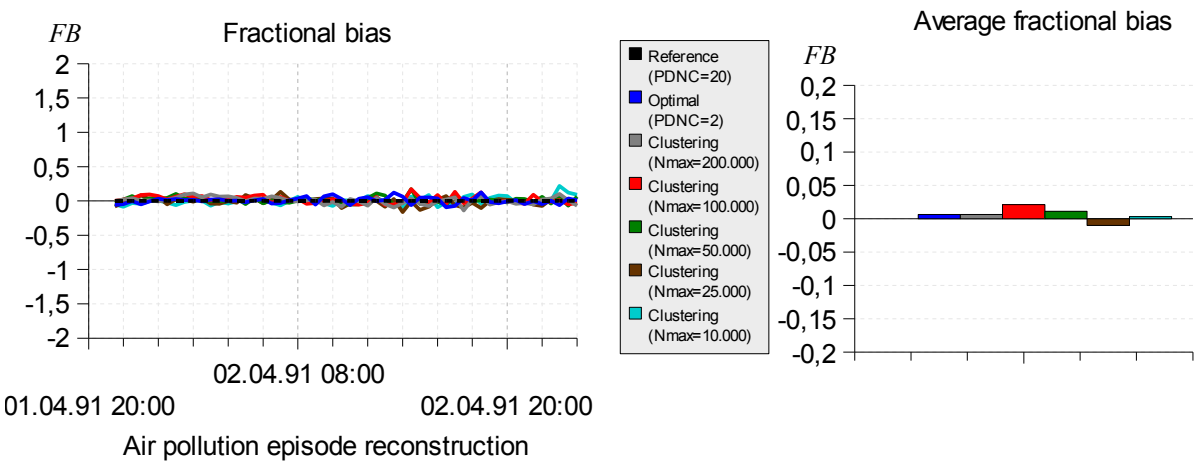


Figure 79: Fractional bias for different clustering parameters

The ground-level concentrations for several intervals where the correlation between the results is good are presented in Figure 80, and the less well correlated results are shown in Figure 81. From both figures it can be seen that the results obtained with the clustering algorithm are bubbled and less smooth than the original.

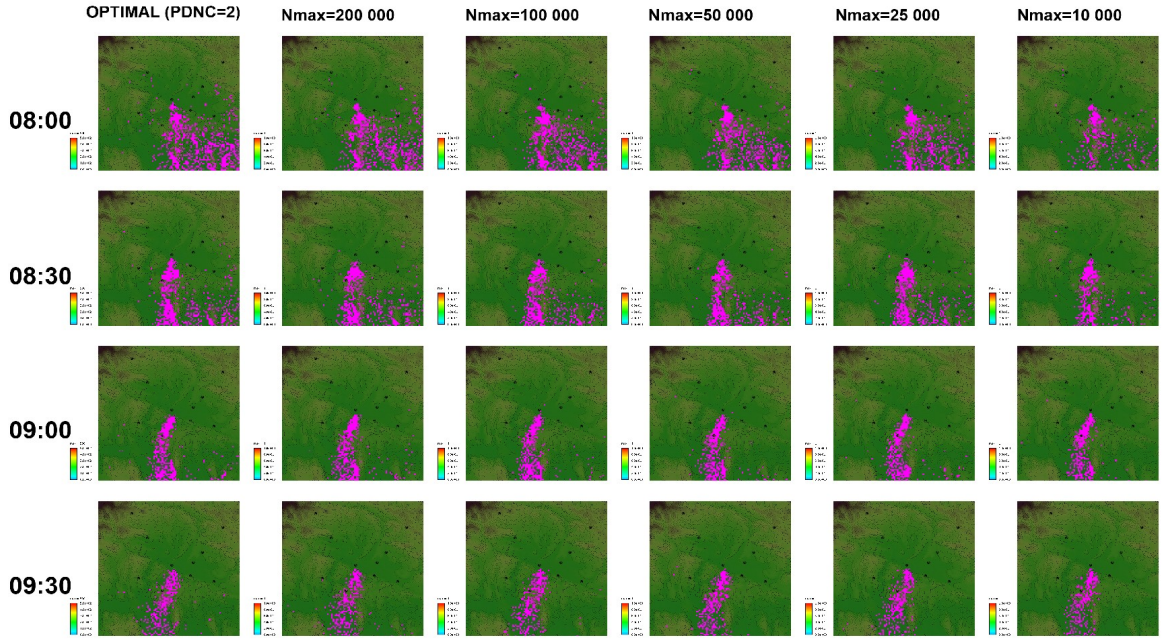


Figure 80: Well-correlated results between original and hierarchical clustering method with additional parameters

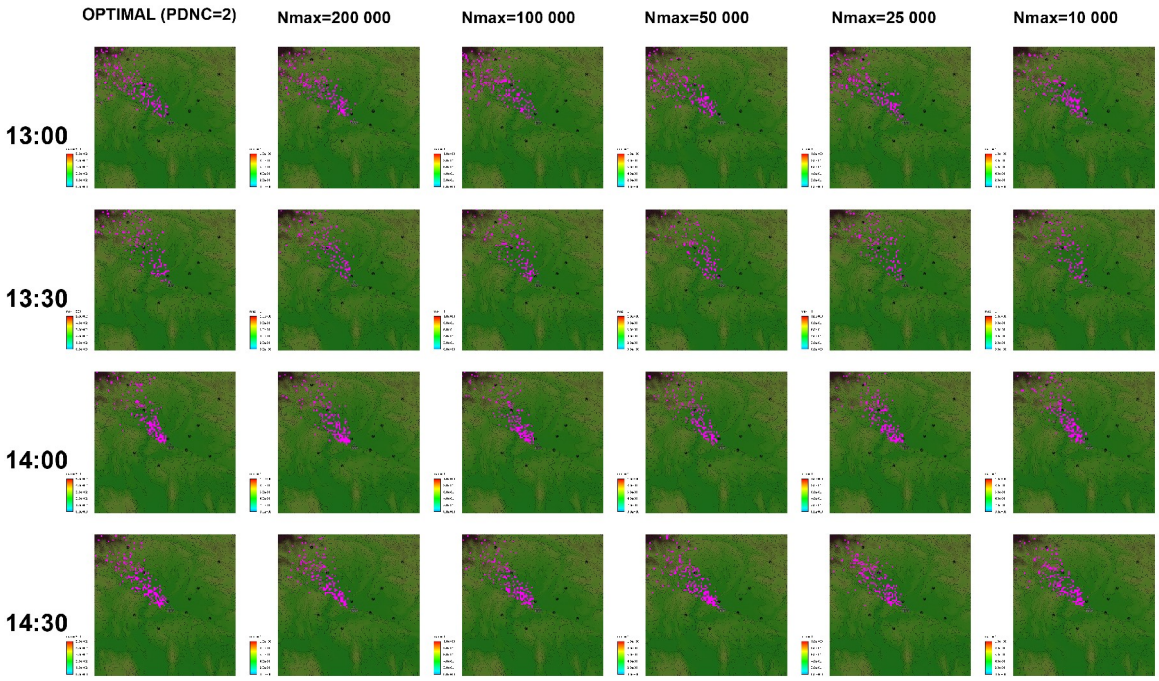


Figure 81: Weakly correlated results between original and hierarchical clustering method with additional parameters

4.1.6. Discussion

The clustering method is proposed for an air-pollution modelling methodology based on Lagrangian particle dispersion. It was adopted to decrease the computational cost by decreasing the number of active particles in the simulations and therefore to prevent the system shut down in the cases where otherwise an overflow of particles would be reached. From the proposed clustering method and results of its validation we can conclude that four basic parameters must be optimally set to achieve satisfactory results:

- N_{sub}*: H*igher values of parameter are reflected in an increase of the used computational time. It is concluded not to use values larger than 4, for practical purposes a value of 2 is chosen.
- N_{size}*: W*hen the parameter N_{size} is increased, the time used for the simulation is decreased, but with the increase of the parameter also the quality of results decreases. The optimal value of 2 is chosen from the obtained results, but the value should not exceed a number of 10 to ensure the quality of the results.
- m_{max}*:* A parameter is introduced into the clustering algorithm to prevent the hypothetical growth of very large and heavy particles, which can lead to physical absurdities. Its value depends on the mass of the emitted particles. In our example, where the emitted particles' weight is 0.01 kg, the parameter is set to 0.1 kg, which means that not more that 10 initially emitted particles can be joined into new particles.

- N_{max}: The computational time used is inversely proportional to the parameter N_{max}. But because another parameter m_{max} is used by the advanced algorithm to prevent the constitution of abnormally large particles the computational time cannot be significantly reduced with the parameter *Nmax* after reaching a certain threshold. From the results it was clear that decrease the of the parameter N_{max} under the values of 25,000 maximum particles does not have any significant effect on the computational time.

According to the finally acquired results presented in the figures, the hierarchical clustering method with additional parameters can be used in practice only for the limitation of a very large number of particles, when the number of particles exceeds abnormal values when extreme situations occur, for example:

- failure of the desulphurization plant when the emissions increase by an order of magnitude,
- when very stable meteorological situations occur, where low winds are present and the air pollution starts to accumulate in the domain.

A severe limitation of the number of particles in the reconstructions at typical situations with the clustering method is not recommended because the quality of the results becomes very poor. To preserve the good quality of the results only a slight limitation is concluded. It is suggested that another method should be integrated into the current modelling system to improve the final results. A comparison of original results and the results obtained with the clustering algorithm showed that the results obtained with the clustering algorithm are bubbled and less smooth than the original. The same effect occurs when not enough particles are used in the reconstructions. The results could be improved according to the recommendations from the research work of De Haan,[109] where point-like air pollution concentrations at certain locations were reconstructed with the AP model based on the LPD and smoothed with the use of a kernel density concentration estimation. The concentration estimation methods based on the kernel density provide smooth results and are not computationally expensive. The presented principle by De Haan is adapted and adjusted in the next subsection to be used for the reconstruction of the concentrations of each cell in the domain, especially to improve the quality of the reconstruction of a two-dimensional ground-concentration field. It is integrated into the air-pollution modelling methodology to improve results obtained with the hierarchical clustering method with additional parameters.

4.2. Method for estimation of a cell concentration based on kernel density

4.2.1. Introduction

In air-pollution modelling the methodology based on the Lagrangian particle dispersion concentrations are estimated by counting the particles in a cell that has a rectangular shape. The term "box counting" is used in the literature to denote this concentration estimation method. A study was performed by De Haan,[109] where the effects of the different size and position of the boxes on the estimated point concentration was investigated. The outcomes of the study showed that if the boxes size are small, the concentration distribution becomes very noisy, having a large variance, and that if the sizes of the boxes are too large, the concentration becomes over smoothed, having a large bias. To minimise the sum of the variance and the bias, the kernel estimation method was proposed as an alternative, which also allows the number of particles to be reduced by an order of magnitude as compared to predictions made by the box counting method. The most important parameter of the kernels, the bandwidth, was outlined and its value was determined from the standard deviations of the particle position distribution. In the study six different kernel shapes were compared: Gaussian, Epanechnikov, bi-weight, tri-weight, quad-weight and quint-weight. For the near source point-like estimations the tri-weight kernel approach was recommended, and for the intermediate to far-field point-like estimations the quad-weight kernel was recommended.

The presented study by De Haan[109] was focused on point-concentration estimations. For regulatory purposes the ground-level concentrations for a certain area of interest must be reconstructed. The focus must be extended from point concentration estimations to cell concentration estimations. Usually, the area of interest is split into a grid of rectangular boxes or cells, and for each cell at ground level the concentration is estimated by counting the particles caught in a certain cell. When not enough particles are used in the Lagrangian simulation, the concentrations between the neighbouring cells can become very different and the concentration distribution becomes very noisy over the area. This effect is illustrated in the subsection *3.6.Determination of acceptable simulation results*. From the obtained results it can been seen that the results become bubbled and less smooth when a smaller number of particles are used for the simulation. Because less heavier particles are released during the simulation, the concentrations in certain cells become very high, while the concentrations in neighbouring cells are zero. From the point of view of an inexperienced observer some erroneous conclusions could be made, like, for example, that only half of the village or town has been exposed to air pollution. A similar effect is also faced when clustering is introduced to the LPD computer model to reduce the computational expense.

A method for the concentration estimation is proposed in the following sub-sections to reduce the presented effect. The method is complementary to clustering. It can be used independently to improve the results when not enough particles are used in the simulation. Another possibility is a combination with the clustering method when the number of particles after the clustering process is significantly reduced.

4.2.2. Development of a method for a a cell concentration estimation

Theoretical background

To improve the results of the ground-level concentration estimations a method proposed by De Haan is adopted and expanded from the point-concentration estimations to the cell-concentration estimations[109]. In the study by De Haan[109] a quad-weight kernel was recommended as being optimal, but the results of the other kernels were not falling behind significantly. The cell-concentration estimation method presented in this thesis is based on a Gaussian density kernel. A comparison of the different kernels is not a matter for this study; it was already addressed in the study by De Haan[109].

The cell-concentration estimation method algorithm begins with the definition of a three-dimensional Gaussian function for each particle in the area of interest. The integration over the area below the kernel of each three-dimensional Gaussian function approximately equals the mass of the particle that belongs to that Gaussian function, as defined in equation (4.2). The standard deviation in each direction of the three-dimensional Gaussian is a function of the age of the particle and also of its mass, as defined in equation (4.3). This dependence is determined from the assumption that the older particles should be spread across the area more widely and also heavier particles should also be more spread because they could be a product of several particles joined by the clustering process. The position of each three-dimensional Gaussian function is placed into a domain with its centre equal to the position of the particle as defined in equation (4.4).

$$\int_{-3\sigma_x}^{+3\sigma_x}\int_{-3\sigma_y}^{+3\sigma_y}\int_{-3\sigma_z}^{+3\sigma_z}\frac{1}{\sqrt{2\pi}\sigma_x}\frac{1}{\sqrt{2\pi}\sigma_y}\frac{1}{\sqrt{2\pi}\sigma_z}e^{-\frac{1}{2}\left(\frac{x-x_0}{\sigma_x}\right)^2-\frac{1}{2}\left(\frac{y-y_0}{\sigma_y}\right)^2-\frac{1}{2}\left(\frac{z-z_0}{\sigma_z}\right)^2}\approx m_p \tag{4.2}$$

$$\sigma_i = f(t_p, m_p); i = x, y, z \tag{4.3}$$

$$x_0 = x_{particle}; y_0 = y_{particle}; z_0 = z_{particle} \tag{4.4}$$

m_p ...mass of particle
t_p ...age of particle

In practice, the above equations mean that a particle was modified from a point-like shape to the shape of a bubble with a mass more densely concentrated around its centre. The obtained three-dimensional Gaussian functions or bubbles are in the process of area concentration estimation spread between several cells of the domain according to its position, mass and age, and not only to one cell as in the simple box counting method.

For an illustration of distributing a Gaussian function into several cells, a one-dimensional example is used. A one-dimensional example is presented in this subsection because of its simplicity and because it can be easily applied in two or three dimensions. A distribution of a one-dimensional Gaussian function between cells in one dimension is presented in Figure 82. The mass of a single particle that is distributed into the cell c_i is defined by equation (4.5), where *erf()* is used to denote an error function or a cumulative distribution function.

$$m_{c_i} = \frac{1}{\sqrt{2\pi}} \int_{i \cdot w}^{(i+1) \cdot w} e^{-\frac{1}{2}\left(\frac{x - x_0}{\sigma_x}\right)^2} = erf\left((i+1) \cdot w\right) - erf\left(i \cdot w\right) \tag{4.5}$$

$erf\,()$...cumulative distribution function
w ...width of the cell

And when the above example is extended into a three-dimensional form the equation (4.6) is obtained.

$$m_{c_{i,j,k}} = m_{c_i} \cdot m_{c_j} \cdot m_{c_k} \tag{4.6}$$

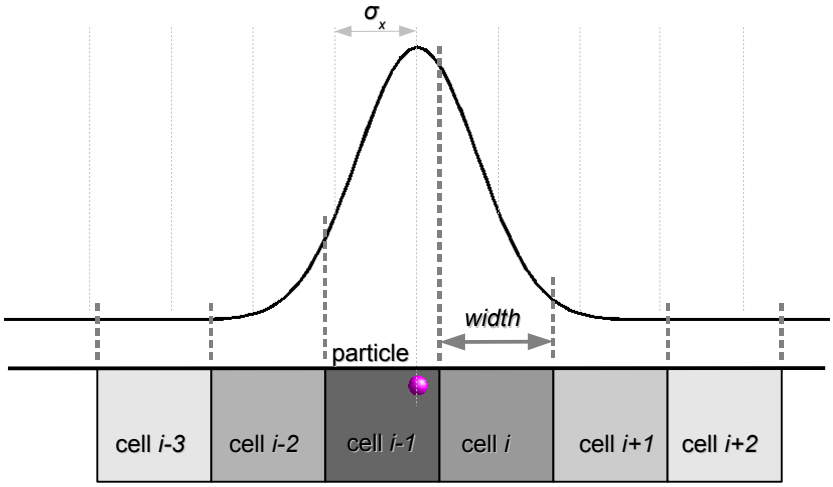


Figure 82: Distribution of a one-dimensional Gaussian function between cells in one dimension

For practical purposes the mass that extends over the area of 3σ is neglected because the integral has very low values and its contribution to the sum of masses in each cell is negligible. Computation would only increase the computational expense, but the effect on the final result is very small.

In the box counting method the concentration in each cell is estimated by adding the masses of the particles positioned in the cell and divided by the sum of the volume of the cell, as defined in equation (4.7).

$$c_{c_{i,j,k}} = \frac{\sum_S m_p}{V_{i,j,k}} \tag{4.7}$$

S ...set of particles at position of $cell(i, j, k)$
$V_{i,j,k}$ ...volume of $cell(i, j, k)$
m_p ...mass of particle from set S

In the kernel-density estimation method the mass of each particle is distributed between several cells around the position of the particle and these partial masses of each particle are summed in each cell and dived by the volume of the cell to obtain the concentration, as defined in equation (4.8).

$$c_{c_{i,j,k}} = \frac{\sum_{O} m_{c_{i,j,k}}}{V_{i,j,k}} \tag{4.8}$$

O ...set of particles around the position of $cell\,(i\,,j\,,k\,)$
$m_{c_{i,j,k}}$ ...part of mass of particle from set O

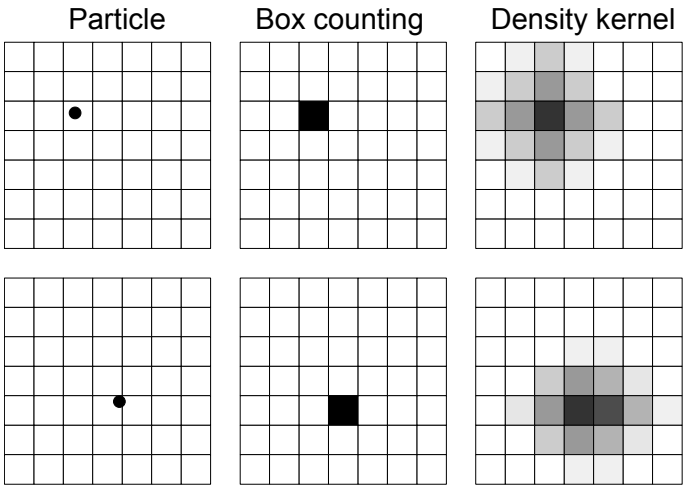


Figure 83: Illustrated comparison of box counting and cell concentration estimation method based on a kernel density

For example, when only one particle in a two-dimensional domain is present, as shown on the left side of Figure 83, the calculated concentration in the case of the box counting estimation method affects only one cell, as presented in the middle of Figure 83. And in the case of the cell estimation method based on kernel density the mass of the particle is spread between the neighbouring cells smoothly, as presented on the right side of Figure 83. It is expected that the advanced kernel density estimation method performs especially well when the particle is positioned near the border of the cell, where it is obvious that the estimated concentration between the neighbouring cells should not be sharp, as presented in Figure 83.

Ground reflection

A special treatment of the particles that are close to the ground is made to prevent the loss of part of the mass of a particle that falls below the ground level due to the kernel function. The basic principle is presented in Figure 84. It is based on the idea of taking the dispersion from a punctual release near an impermeable boundary. In this case the situation with one release and boundary is equivalent to another with two releases (original and virtual release that is actually the original release reflected over the boundary) and no boundary[26]. In our case the ground level represents a typical impermeable boundary. For each particle near the ground an additional virtual particle is introduced below the ground. It is actually the original particle reflected over the ground level. For example, the contribution of one particle to the ground concentration consists of its original position and of its virtual reflection as presented in Figure 84. What actually goes below the ground due to the kernel function is taken back by

means of a reflected particle, whose quantity above the ground is the same as the quantity previously lost below the ground.

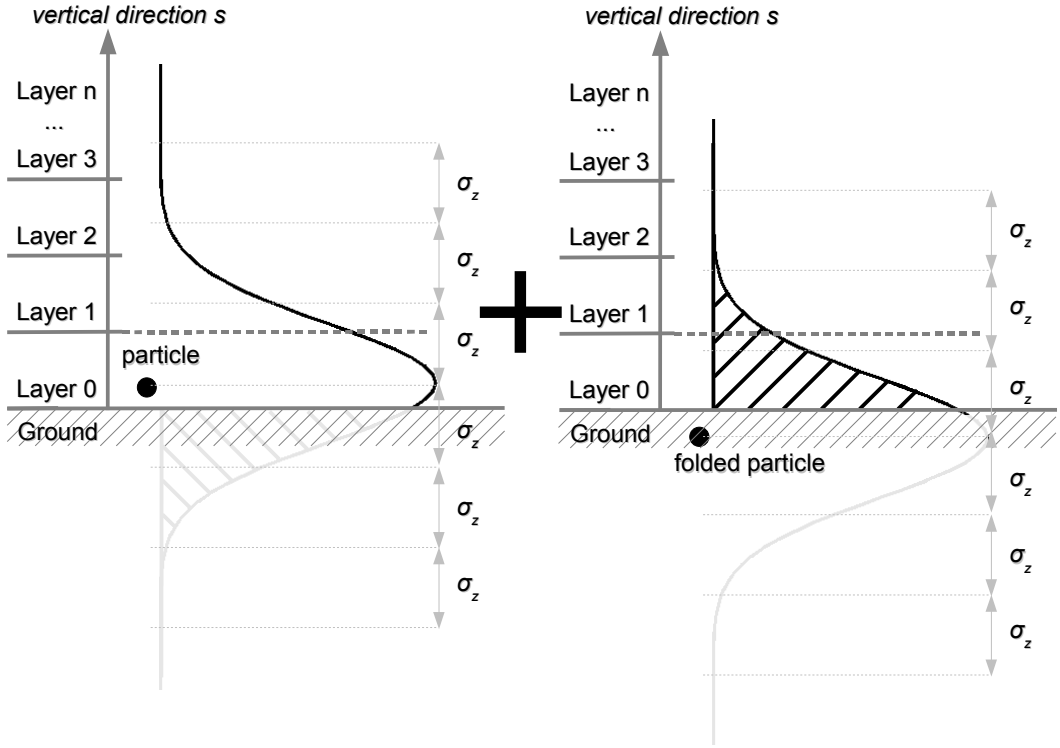


Figure 84: Consideration of a particle near the ground where the final ground level concentration consists of a contribution from the original particle and its reflected (folded) virtual particle

Implementation

The software for a ground-level concentration estimation based on the kernel density is developed in the C++ programming language using the Qt[112] development tools and the open source mathematical library GSL[115] for the C++ developed in the GNU project coordinated by the Free Software Foundation Inc.. The application is developed to reconstruct the ground-level concentrations, where the height of the ground-level layer is set as an input parameter. The software is developed and designed to be integrated into the existing LPD computer model. Figure 85 presents highlighted modification of the LPD computer model scheme that is presented in previous section in Figure 45.

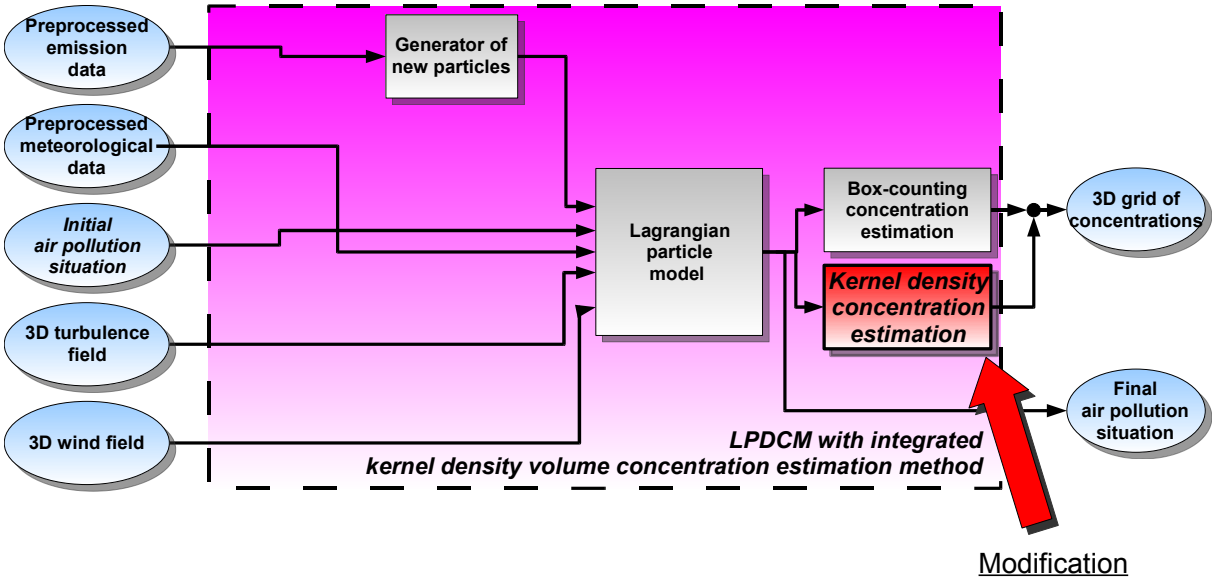


Figure 85: LPD computer model with integrated the kernel-density cell concentration estimation method

There are several inputs that must be provided before the software can be executed:

- *final air-pollution situation* described in the file generated by the AP computer model at the end of each simulation run. In this file the position and the properties of each particle at the end of the simulation run is saved. This file is also named the *restart* file. In this file each particle is described by several properties, among the most important for the concentration estimation are the x, y and z positions of the particle in the domain, the age of the particle, and the mass of each particle.
- *topography*, which contains the orography of the domain defined by the altitude of each ground cell in meters above sea level.
- σ_x, σ_y *and* σ_z are the parameters that define the initial value for spreading in the x (west-east), y (north-south) and z (vertical) directions.

- *the h* parameter defines the height of the ground-level layer.

The three-dimensional domain is, in the air-pollution computer model, split into a three-dimensional grid of cells, which are evenly distributed in the horizontal x and y directions, while in the vertical direction the distribution is performed using the terrain-following coordinates, as defined in the AP computer model manual,[103] where the spatial coordinates are *x, y* and *s*. The transformation from Cartesian coordinates is defined by equations (4.9, 4.10, 4.11), where z_t defines the top of the domain and z_g the altitude of the ground level above sea level. In the terrain-following reference system, the surface is not a horizontal plane, but follows the topographical profile. The *s* constant planes at higher levels also tend to reproduce the topography below, but this effect decreases when moving far from the ground level. The *s=1* level represents a horizontal plane.

$$x = x \tag{4.9}$$

$$y = y \tag{4.10}$$

$$s = \frac{z - z_g(x, y)}{(z_t - z_g(x, y))} \tag{4.11}$$

4.2.3. Evaluation of the developed density kernel concentration estimation

To validate and evaluate the performance of the developed application and the dependence on the input parameters, several simulations were performed and the results are compared to the original reference ground concentration field, as defined in detail in subsection *3.5.Evaluation methods.* Simulations are performed on a Šaleška region field data set air-pollution situation that lasted from 1st of April 1991 at 20:00 to 3rd of April 1991 at 00:00.

The first simulation was performed to generate reference ground concentration fields for all the reconstructed air-pollution episodes. To accomplish this task, the particle number density coefficient is set to the maximum possible value of *PDNC=10.0* according to the available computational resources. The results of the time used for the first simulation are presented in Figure 50. After that a second simulation is performed, where the particle number density is decreased to the optimal value of *PDNC=0.25*, which is determined the point of view of time spent for the simulation. The optimal value has been selected according to the time that is spent to reconstruct the air pollution over the area of interest for the time period of one year. For the selected optimal point where 0,23 hours are spent for the reconstruction of one day, approximately 3,5 hours are spent for the reconstruction of one year, which is really quick in practice. A comparison of the results for the first and the second simulations presented in Figure 50 shows that significantly less computational time is used for the second simulation. Unfortunately, but as expected, the consequence of this profit is the very poor correlation presented in Figure 52 and the large root mean square error presented in Figure 53.

To improve the poor results the developed kernel density concentration application is used.

But before the simulation is performed with the advanced AP computer model, several experiments are made to determine the optimal input parameters σ_x, σ_y and σ_z for the selected Šaleška region field data set. In all experiments the horizontal parameters σ_x and σ_y are set equally. In all the presentations of the results both parameters are presented as a single independent parameter $\sigma_{x,y}$, while the second independent parameter is vertical σ_z.

In the performed experiments three cases of optimal inputs into the kernel density concentration application are determined. Each case represents the optimal inputs of σ_x, σ_y and σ_z for a single simulated time interval. Three of the used simulated time intervals are selected according to the obtained results of the correlation coefficient presented in Figure 52.

A poor correlation coefficient is obtained at the simulated time interval at 13:30 hour. in Figure 87 the results of the comparisons between the reference and "kernel density estimated" ground concentrations are presented. For the comparisons different "kernel density estimated" ground concentrations were used, where the ground concentration field was estimated with different input parameters. The dependence of the correlation according to the horizontal input parameters $\sigma_{x,y}$ is presented in the first graph of Figure 87. The second graph presents the root mean square error, and the third graph fractional bias. Each curve on the graphs presents the dependence for one vertical input σ_z value. From the obtained results we can see that the correlation increases and the root mean square error decreases to some optimal values of $\sigma_{x,y}=250$ *and* $\sigma_z=20$. The optimal point is determined where the correlation coefficient between the reference and the estimated ground concentration reaches its highest value. After the optimal point is reached the correlation starts to decrease and the root mean square error starts to increase. The comparisons also show that with the larger values of the input parameters the fractional bias slightly decreases, and then slowly converges to a constant, which gives slightly underestimated results. With the comparisons of fractional bias the optimal input parameters cannot be estimated, but the results give us the assurance that at optimal values of the inputs the results will be reliable. In Figure 86, besides the reference ground concentration, three examples of "kernel density estimated" ground concentrations are also presented: the over-smoothed, optimal and the under-smoothed.

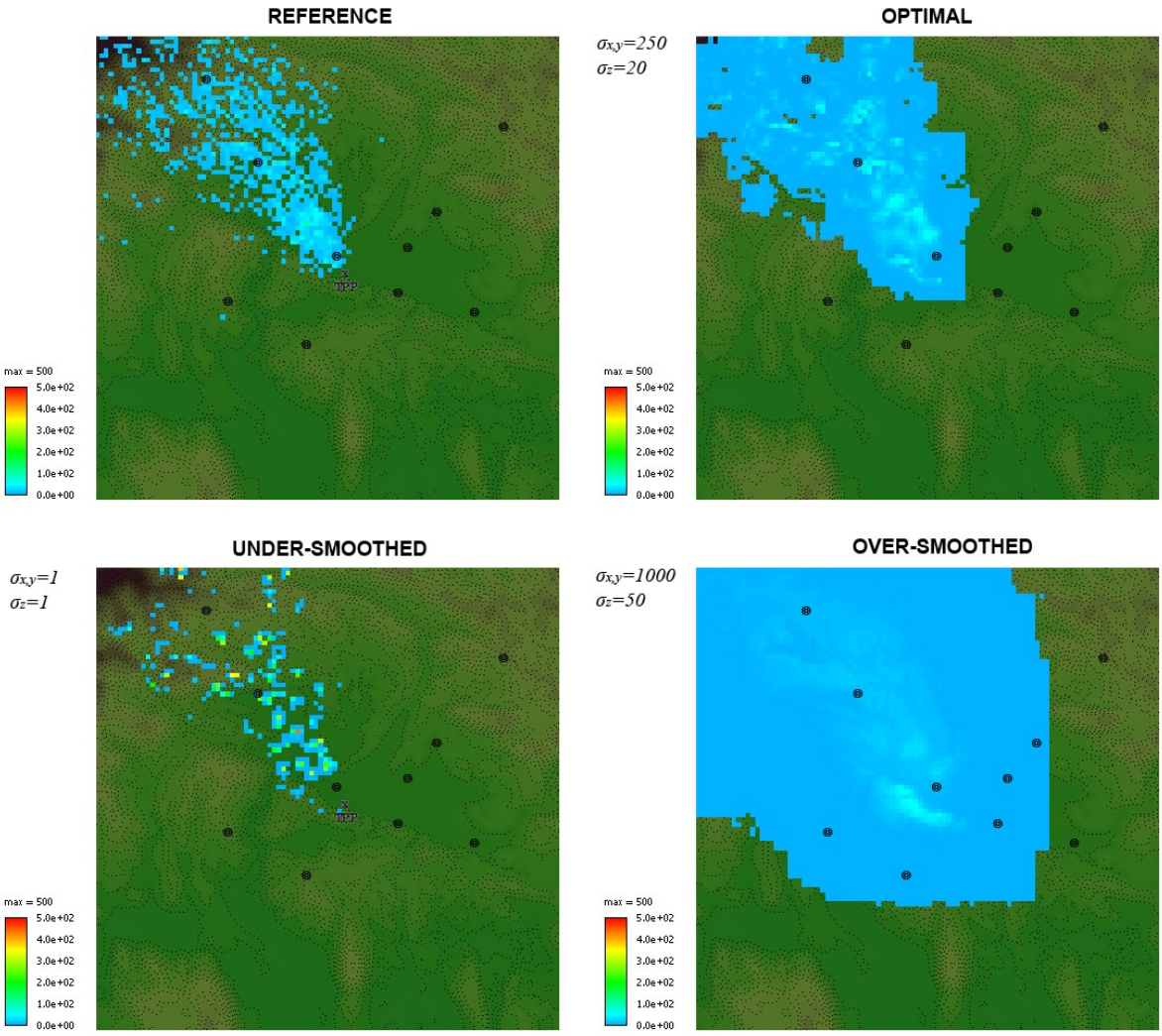


Figure 86: Examples of density kernel concentration estimations for a poor correlation at the simulated time interval 13:30

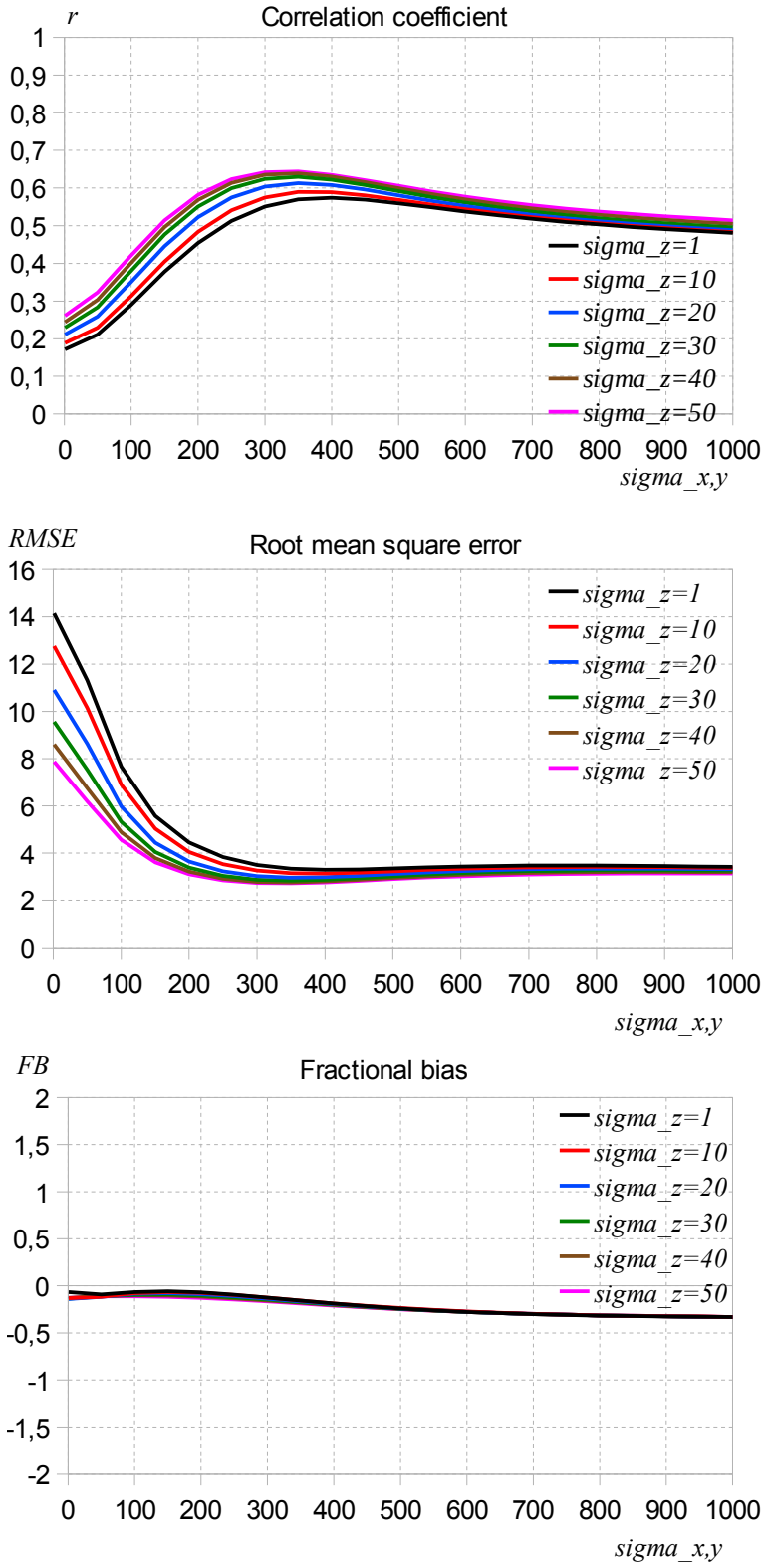


Figure 87: Comparisons of the density kernel concentration estimations for a poor correlation at the simulated time interval 13:30

A good correlation coefficient is obtained at the simulated time interval at 08:30 hour. In Figure 89 the results of the comparisons between the reference and "kernel density estimated" ground concentrations are presented. The dependence of the correlation is presented on the first graph of Figure 89, the second graph presents the root mean square error, and the third graph the fractional bias. From the obtained results we can see that the trend in the results is the same as that obtained in the previous case, only that different optimal values of $\sigma_{x,y}=200$ *and* $\sigma_z=20$ are determined. In Figure 88, besides the reference ground concentration, three examples of "kernel density estimated" ground concentrations are also presented: the over-smoothed, the optimal and the under-smoothed.

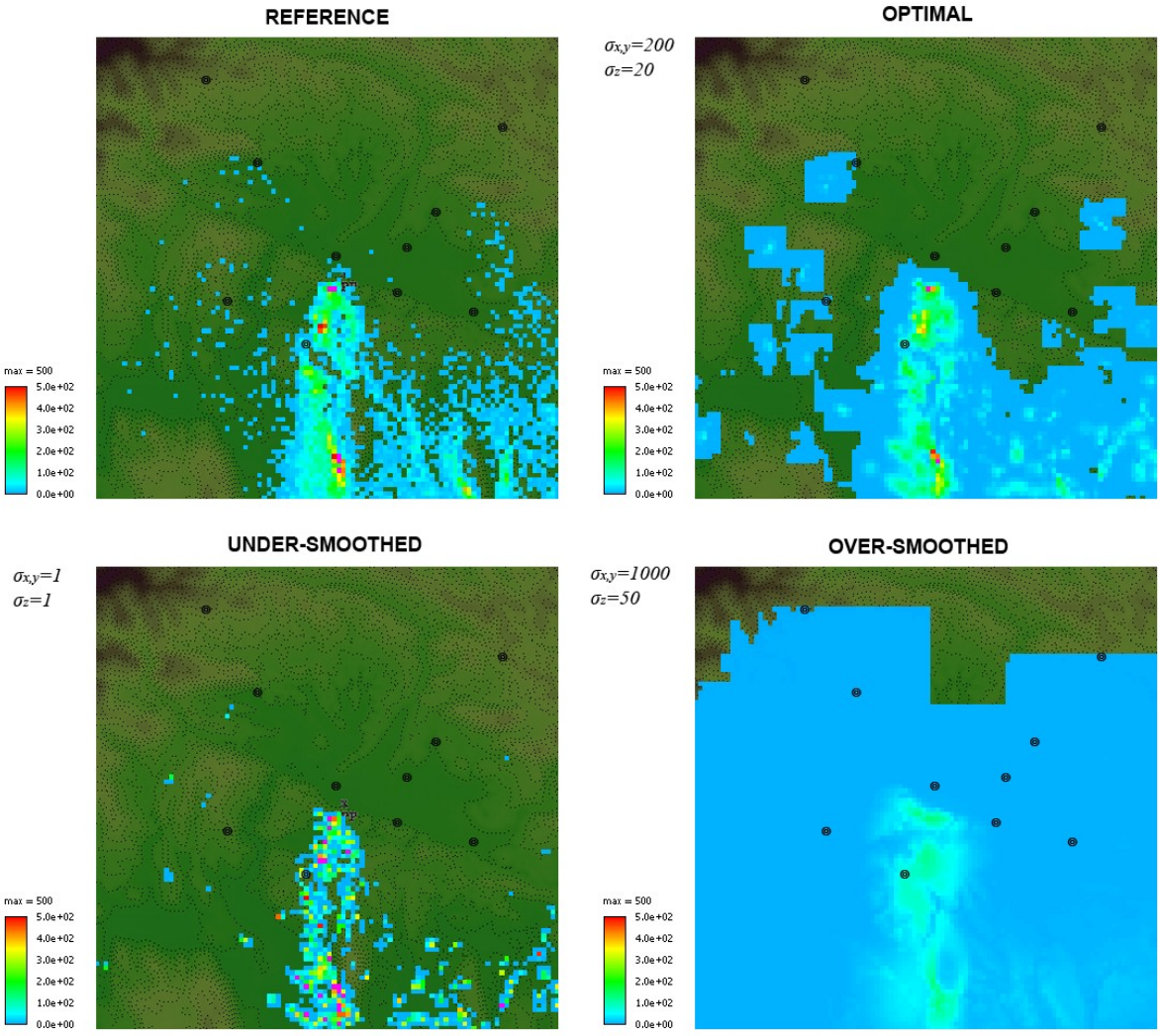


Figure 88: Examples of the density kernel concentration estimations for a good correlation at a simulated time interval 08:30

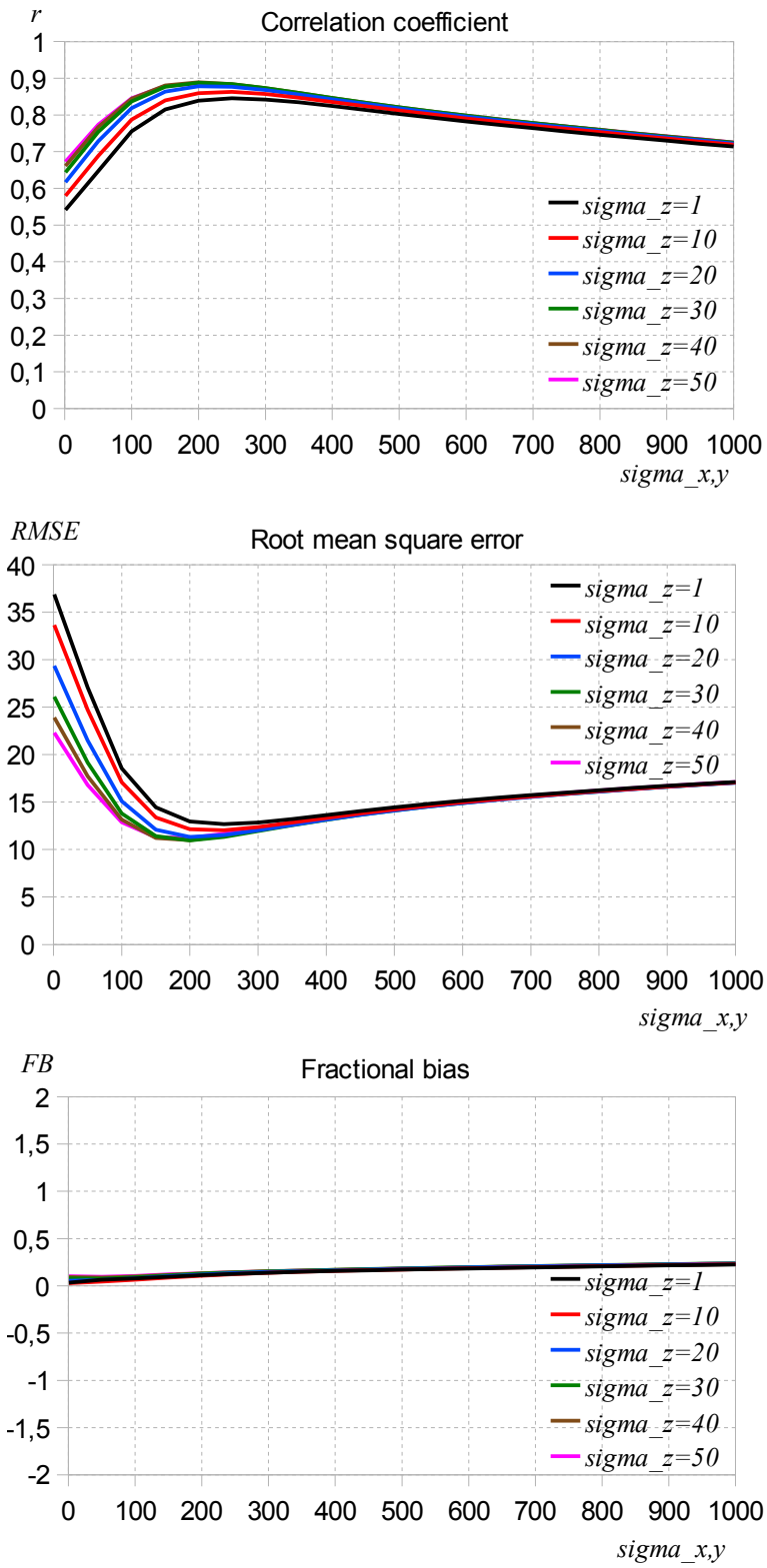


Figure 89: Comparisons of the density kernel concentration estimations for a good correlation at a simulated time interval 08:30

A very good correlation coefficient is obtained at the simulated time interval at 23:30 hour. In Figure 91 the results of the comparisons between the reference and "kernel density estimated" ground concentrations are presented. The dependence of the correlation is presented in the first graph of Figure 91, the second graph presents the root mean square error and the third graph the fractional bias. From the obtained results we can see that the trend of the results is the same as obtained in the previous case, only that different optimal values of $\sigma_{x,y}=150$ *and* $\sigma_z=20$ were determined. In Figure 90, besides reference ground concentration, three examples of "density kernel estimated" ground concentrations are also presented: the over-smoothed, the optimal and the under-smoothed.

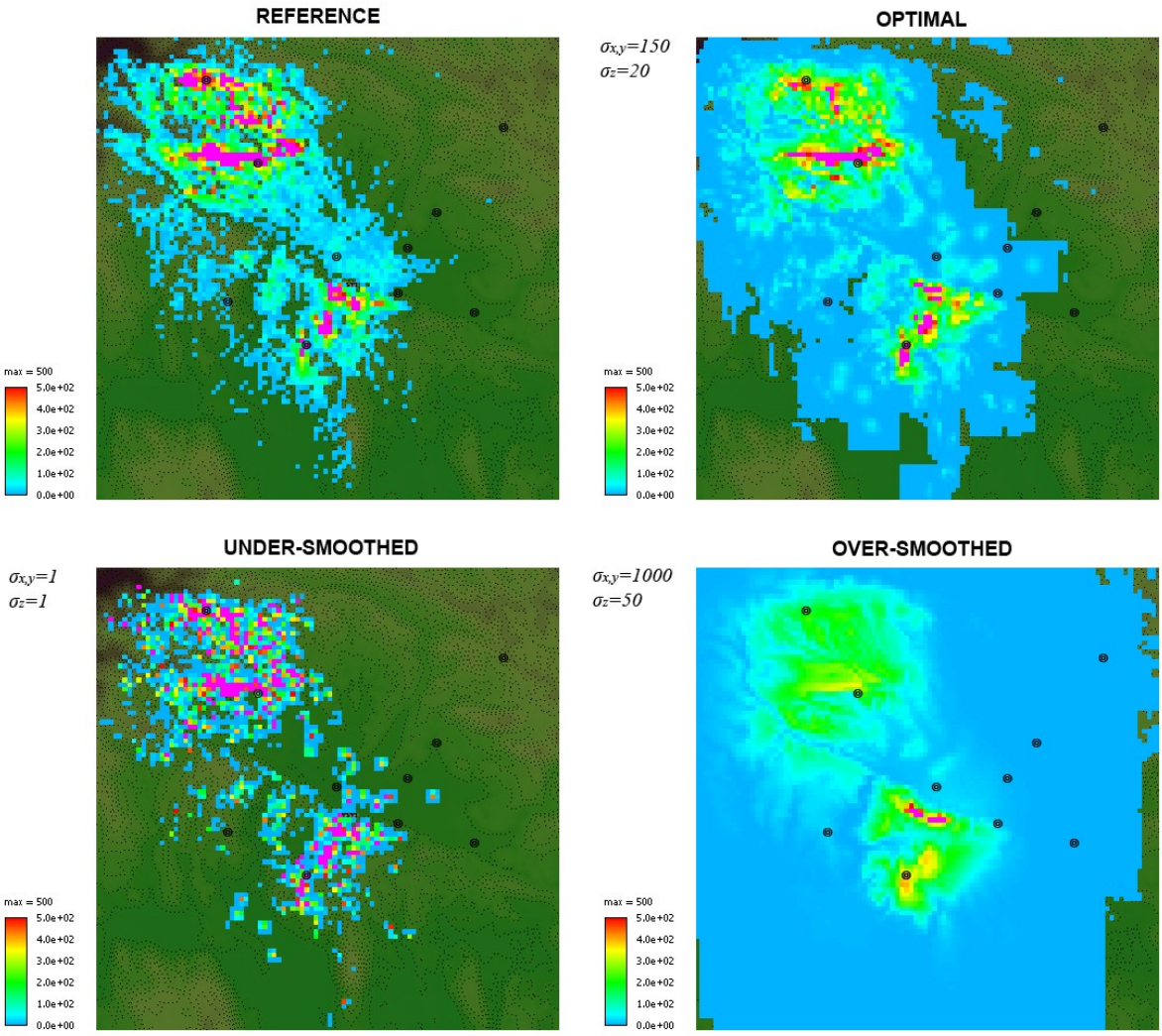


Figure 90: Examples of the density kernel concentration estimations for a very good correlation at a simulated time interval 23:30

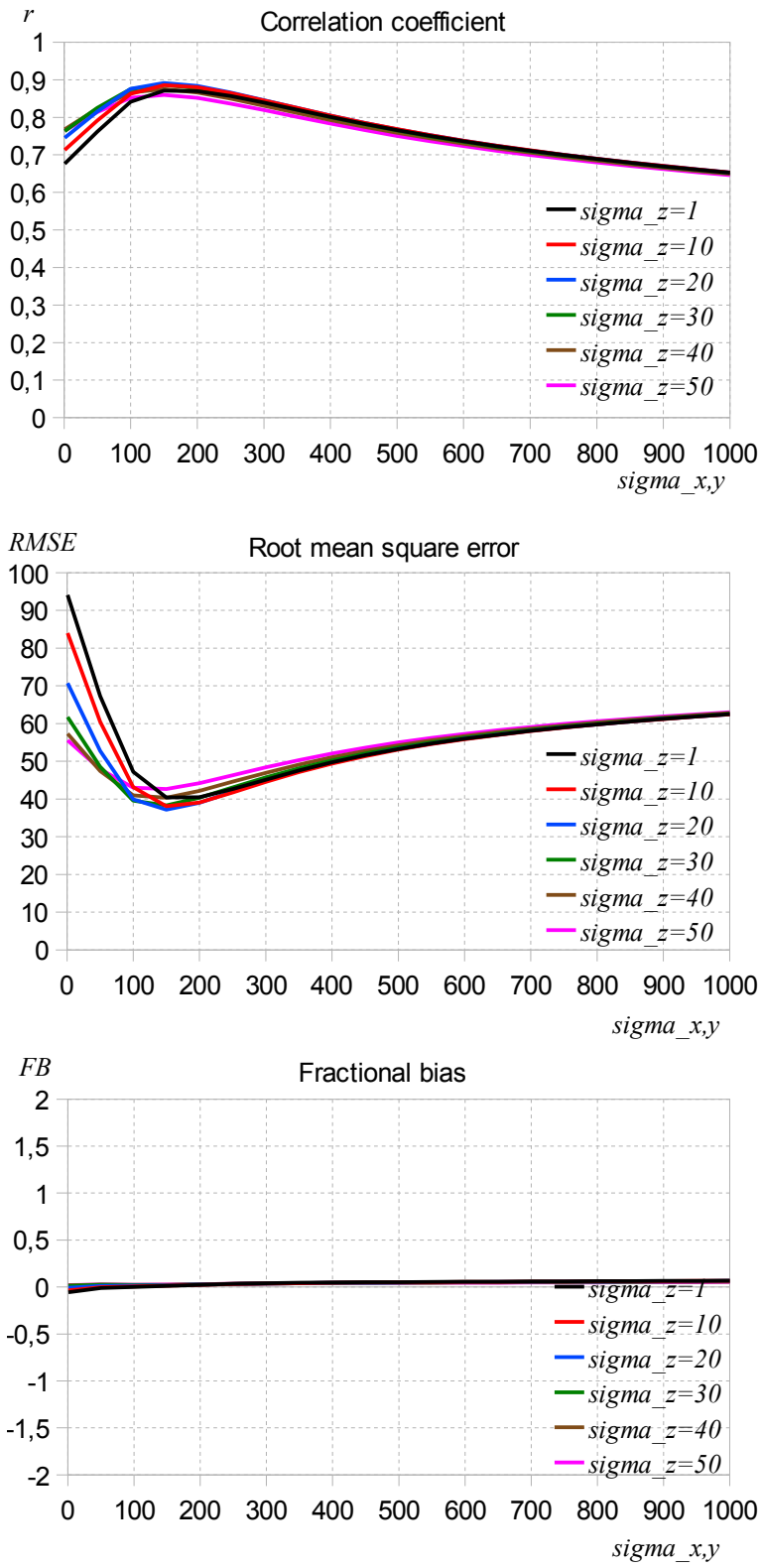


Figure 91: Comparisons of the density kernel concentration estimations for a very good correlation at a simulated time interval 23:30

After the three different optimal input values for the different correlation coefficients have been determined, for each optimal point one simulation is performed by using the improved AP computer model. In Figure 92 the comparisons of the simulation results are presented. The black curve represents the results of the simulation performed with the optimal value of *PDNC=0.25,* where the ground concentrations were estimated with the "box-counting" method, while the red, blue and green curves represent the simulations made with the kernel density concentration estimation method. All three graphs prove that the correlation coefficient and the root mean square error are significantly improved, while the fractional bias comparison showed that there is practically no underestimations in the results. The comparisons of the three different optimal points show that the final optimum point is taken from an example where the correlation with the original result is very good. The other two obtained optimal points that gave some slightly better or slightly worse results.

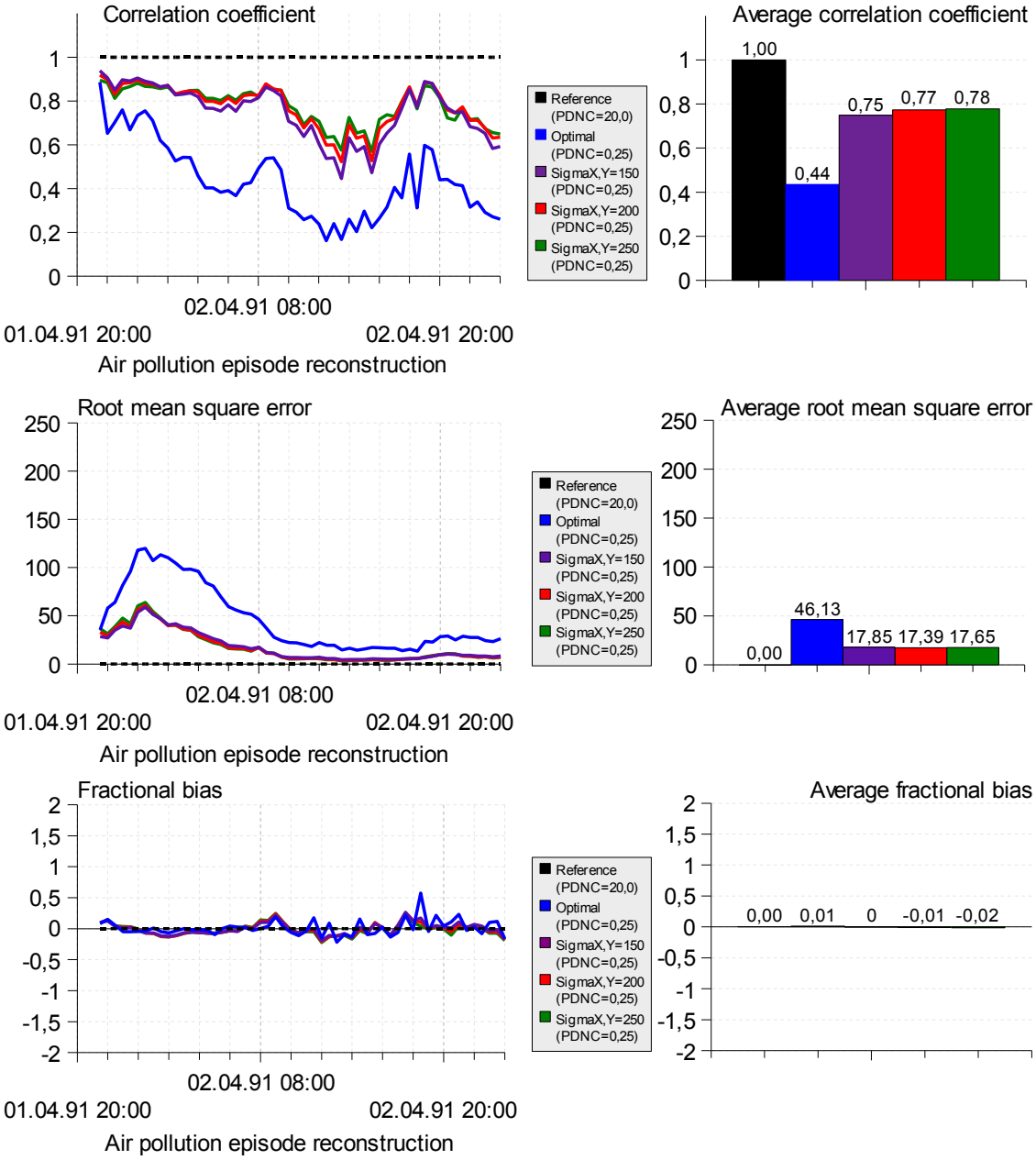


Figure 92: Final comparison of the results obtained with the original box-counting concentration estimation method and the results of the density kernel concentration estimation method with different parameters

4.2.4. Discussion

In this subsection a new, advanced kernel density cell concentration estimation is proposed to improve the results of the "box counting" ground-level concentration estimations. The contributed method is adopted according to a method proposed by De Haan, where the expansion from the point concentration estimations to the cell concentration estimations is made. An adaptation of the method is presented in detail in this subsection.

The Lagrangian particle-dispersion model estimates the ground-level concentrations by counting the particles in cell that has a rectangular volume. The term box counting is used in the literature to denote this concentration estimation method. In a study performed by De Haan[109], the kernel density estimation method was proposed as an alternative, where the focus was on point concentration estimations. But for regulatory purposes the ground-level concentrations for a certain area of interest must be reconstructed. An additional study is performed in this subsection to extend the focus from the point concentration estimations to the cell concentration estimation.

To validate the method and to determine the dependency of the quality of results according to different input parameters, several experimental simulations were performed and the results are compared to the reference ground concentration field. The simulations were performed on the Šaleška region field data set situation that lasted from 1st of April 1991 at 20:00 until 3rd of April 1991 at 00:00. Before the simulations were performed with the AP computer model based on the improved LPD computer model, several experiments were made to determine the optimal input parameters σ_x, σ_y and σ for the field data set domain. In the performed experiments three cases of optimal inputs into the kernel density concentration application were determined and presented: the over-smoothed, the optimal and the under-smoothed. After the three different optimal input values for the different correlation coefficients are determined, for each optimal point one simulation is performed with the improved AP computer model. The comparisons of the three different optimal points show that the final optimal point is set from the case where the correlation with the original result is very good. The other two obtained optimal points give some slightly better or worse results. The final comparisons prove that the correlation coefficient and the root mean square error are significantly improved, while the fractional bias comparison showed that there is practically no underestimations in the results, which is crucial for long-term evaluations of air pollution.

4.3. Lagrangian particle-dispersion control

4.3.1. Introduction

When the number of particles in a single air-pollution reconstruction episode exceeds the computational resources, the simulation process is interrupted and the reconstruction of the proceeding air-pollution episode begins without a consideration of the previous air-pollution situation. When such a problem occurs, the high air-pollution situation reconstruction can be lost. A clustering method is proposed to reduce the number of active particles that enter into the single air-pollution reconstruction episode as the initial state of the air pollution. The number of new, active particles that are released during a single air-pollution reconstruction episode can be calculated from the *PDNC* (particle density number coefficient) and the emission during the episode. To reconstruct a single air-pollution episode correctly the emission cannot be changed. To avoid an excess of computational resources in a single air-pollution reconstruction episode the *PDNC* should be controlled: when a high emission occurs, the *PDNC* should be decreased, and when a low emission occurs, the *PDNC* can be increased.

Proposed concept of LPD model control is presented on Figure 93. It is also compared to original concept where *PDNC* is not controlled during the simulation. The control is based on a model constructed of feedforward neural network and it's main idea is to change *PDNC* parameter during the simulation to avoid the excess of computational resources.

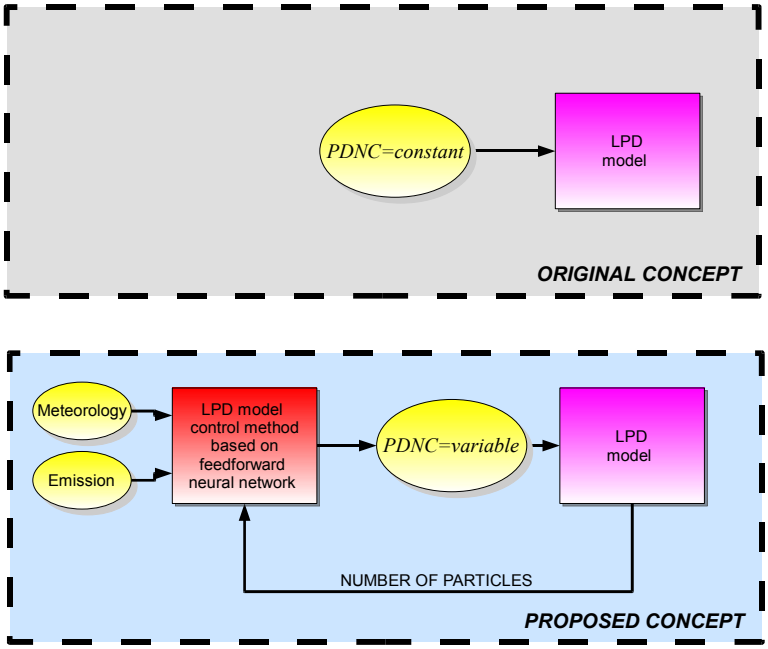


Figure 93: Simple illustration of proposed concept of LPD model control

Control of the *PDNC* can depend on the number of active particles from the reconstruction of a previous air-pollution episode reconstruction and the emission in the current air-pollution episode reconstruction. To improve the control of the *PDNC* an estimation of number of active particles that are lost during the proceeding air-pollution reconstruction episode can

also be used:

- if many active particles are lost out of the domain during the simulation a larger *PDNC* should be used because more new active particles can be released to simulate the same amount of emission,
- if a very small number of active particles are lost out of the domain, a smaller *PDNC* should be used because fewer new particles can be released to ensure that the computational resources will not be exceeded.

In special situations when the smallest possible *PDNC* is used, and it is still expected that the computational resources will be exceeded, the clustering must be activated to reduce the number of particles from a previous air pollution. This is very important for situations when extreme air pollution is expected. Such situations occur very often in calm meteorological conditions when the air pollution starts to accumulate in the domain for a longer time interval. In practice this is achieved by setting the N_{max} parameter of the clustering method to a value that is lower than the number of particles from the previous air-pollution episode reconstruction. To control the described *PDNC* and N_{max} clustering parameters a Lagrangian particle dispersion (LPD) control method is proposed, as presented in Figure 94. It consists of two main subsequent methods. In the first step the percentage of lost particles is predicted based on the meteorology, the emission and the situation of the air pollution at the end of the previous episode reconstruction. In the second step the *PDNC* and N_{max} clustering parameters are determined by a decision-making method. The development of both methods is presented in the following subsections.

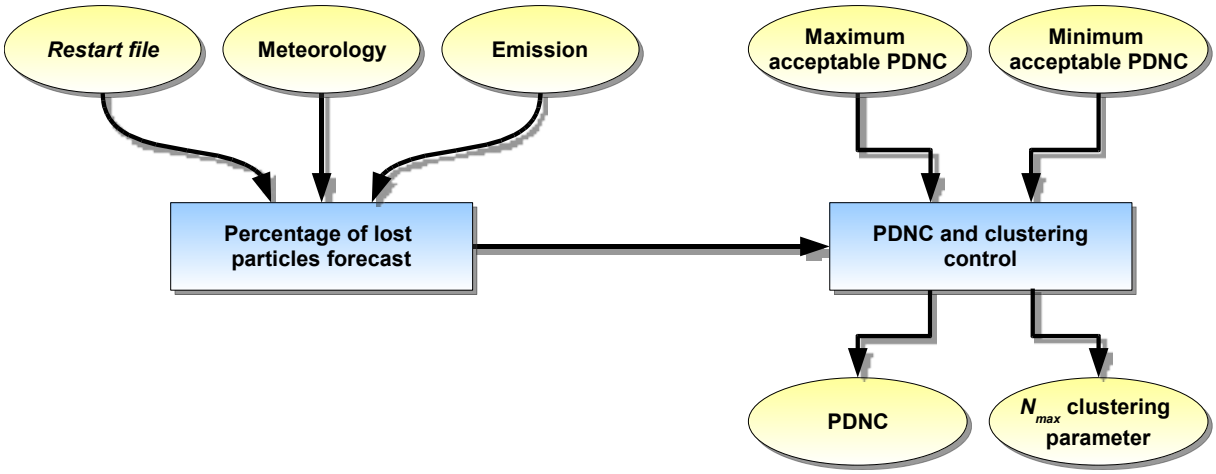


Figure 94: LPD model control method which consists of two main steps: forecast of percentage of lost particles and control of PDNC and clustering based on decision-making method

The improved LPD computer model scheme with the integrated LPD control module is presented in Figure 95 where also clustering module is included in scheme.

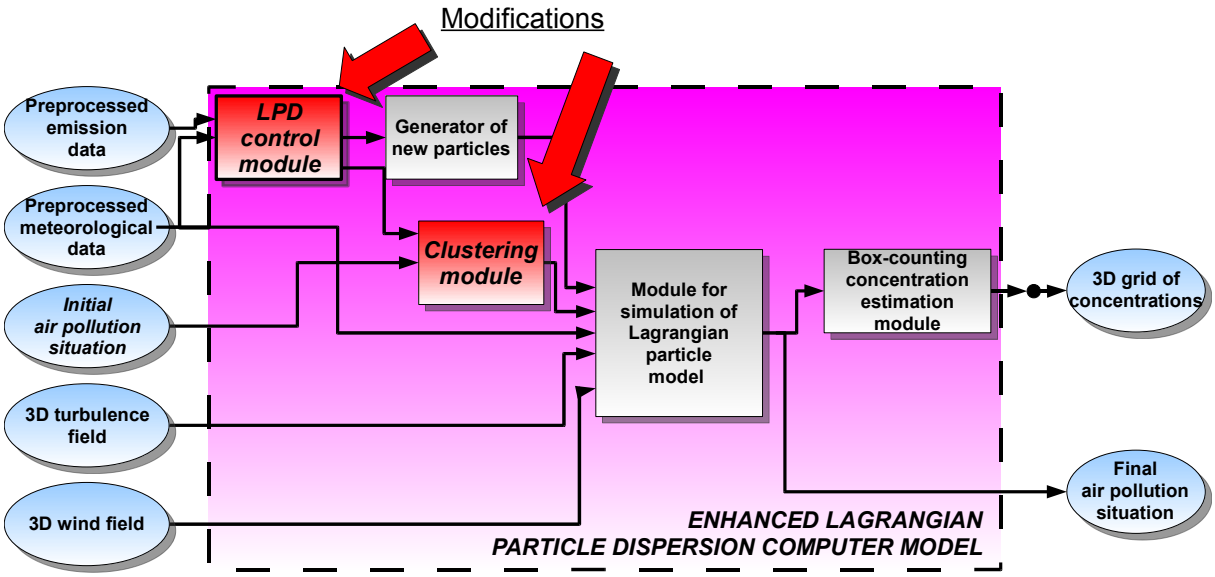


Figure 95: LPD computer model with integrated LPD control module and clustering module

4.3.2. Method for forecasting the percentage of lost particles

The estimation of the percentage of lost active particles during the current air-pollution reconstruction episode is determined with an artificial neural network. Artificial neural networks are selected because it was proven by Hornik et al. that multilayer feedforward networks are universal approximators[47].

They have become a useful and efficient tool in the past ten years for establishing forecasting models in the field of air pollution. Many authors reported the successful forecasting of air pollution using artificial neural networks in recent years. An overview of applications is given by Gardner and Dorling[48].

Formulation of the model's output

The output of the lost particles' number prediction method is the percentage of particles that is lost during the proceeding air-pollution reconstruction episode, as defined by equation (4.12). The number of particles that are emitted in one air-pollution episode reconstruction is eliminated from the output to ensure that the output represents only the percentage of lost particles that were already in the domain at the beginning. The loss of the particles that are emitted in the air-pollution episode reconstruction is not included in the model's output. The output is limited to values from 0% to 100%, where the value of 0% defines that no particles are lost, and the value of 100%, that all particles are lost.

$$p_l = \frac{N_{begin} - (N_{end} - N_{emi})}{N_{begin}} 100\% \quad (4.12)$$

N_{begin} ...number of particles at the begining of air pollution episode reconstruction
N_{end} ...number of particles at the end of air pollution episode reconstruction
N_{emi} ...number of emitted particles in one air pollution episode reconstruction
p_l ...percentage of lost particles

Selection of learning data for construction of model based on feedforward neural network

To achieve good generalising capabilities for the model a data selection must be performed to create the learning set used in the learning process and to create the testing set used for the test generalization of the created model. The learning set consists of the training and validation datasets. The validation dataset consists of a randomly selected 10% of the data from the original learning set and it is used during the learning process to periodically test the feedforward neural network performance using the validation (unknown) dataset to determine the generalization capabilities. The final network is the one that gives the smallest error on the validation set and not on the training set.

The selected data used for the learning process should represent all the typical situations that can occur. It must be ensured that the learning set contains all the essential information needed for predicting the percentage of lost particles, which appear during the simulation performed with the original LPD model.

Usually, a large learning dataset is available, which contains a large number of redundant data. When such a situation occurs the selection of suitable data should be performed to significantly reduce the amount of data that results in significant reduction of the computational cost. A selection of data for the learning can also improve the generalising capabilities of the networks and subsequently the quality of the predictions. Several advanced methods for data selection to reduce the large learning dataset were developed and described by different authors: Cachin[116] and Munro[117]presented pedagogical data selection strategies, where the data for the learning set are selected during the learning process, and Božnar[118] presented two different strategies where one is based on the meteorological knowledge and the other on the Kohonen neural networks clustering capabilities.

In the following paragraph an example of data selection for learning will be presented. The dataset obtained during the simulation performed by the model benchmark kit is relatively small, which shows that no advanced data selection is needed for reducing the learning set.

The lost particles' number prediction method is constructed from the data obtained from input and output parameters of the simulations that are performed on a Šaleška region field dataset air-pollution situation that lasted from 16th of March 1991 at 00:00 until 5th of April 1991 at 00:00, where all the available 913 data are taken. Due to the relatively small dataset all the available data are used for the learning and testing procedures. The emission is set to a constant value during the complete simulation for a better identification and a comparison of the episodes with high air pollution. Because the number of active particles is used for the

presentation of the complexity of the current air-pollution episode (a larger number of particles in the domain represents a more complex situation) the variability of the emission is removed to simplify the development of the method and the presentation of the results. In each air-pollution episode a reconstruction of 3780 new, active particles are released into the domain from the source of the air pollution. The number of active particles after each air-pollution episode reconstruction are presented on the graph in Figure 96. The results presented show that some episodes of high air pollution over the domain occurred when the number of active particles significantly increases. From the number of active particles the percentage of lost particles in each air-pollution episode (p_l) is determined according to equation (4.12). It is presented in Figure 97, which is consistent with Figure 96: in the episodes with a low air pollution when strong wind conditions are present, all the particles (100 %) are lost from the domain and in the episode with a high air pollution when the low wind conditions occurred, fewer particles (< 100%) are lost from the domain. In the most critical situations less than 10% of the particles are lost.

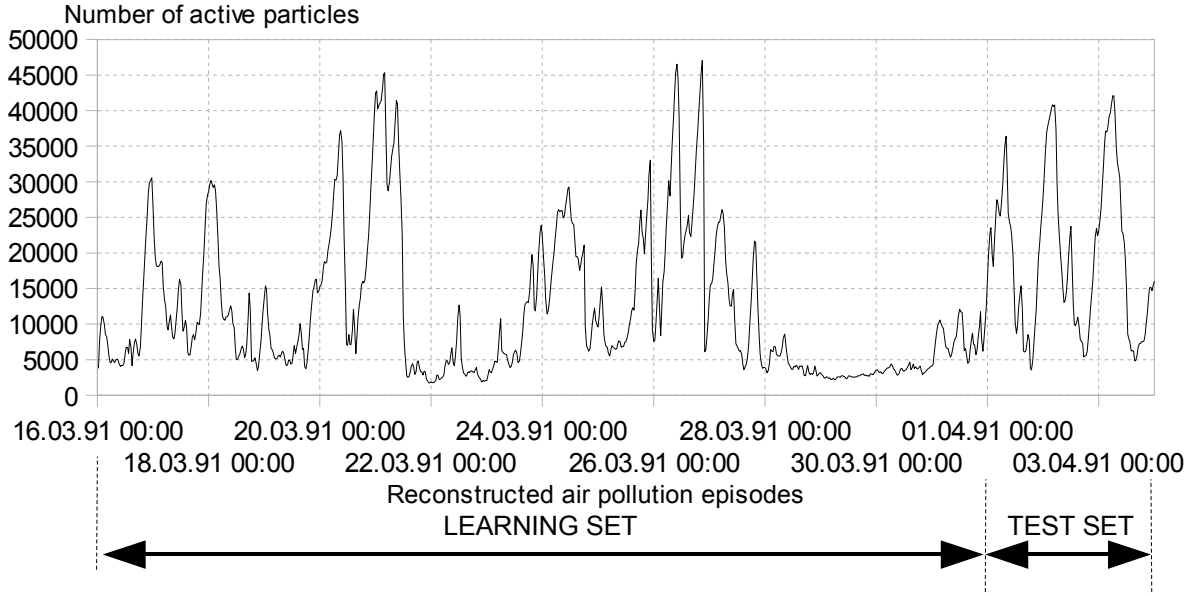


Figure 96: Number of active particles used to represent the complexity of the reconstructed air-pollution episodes

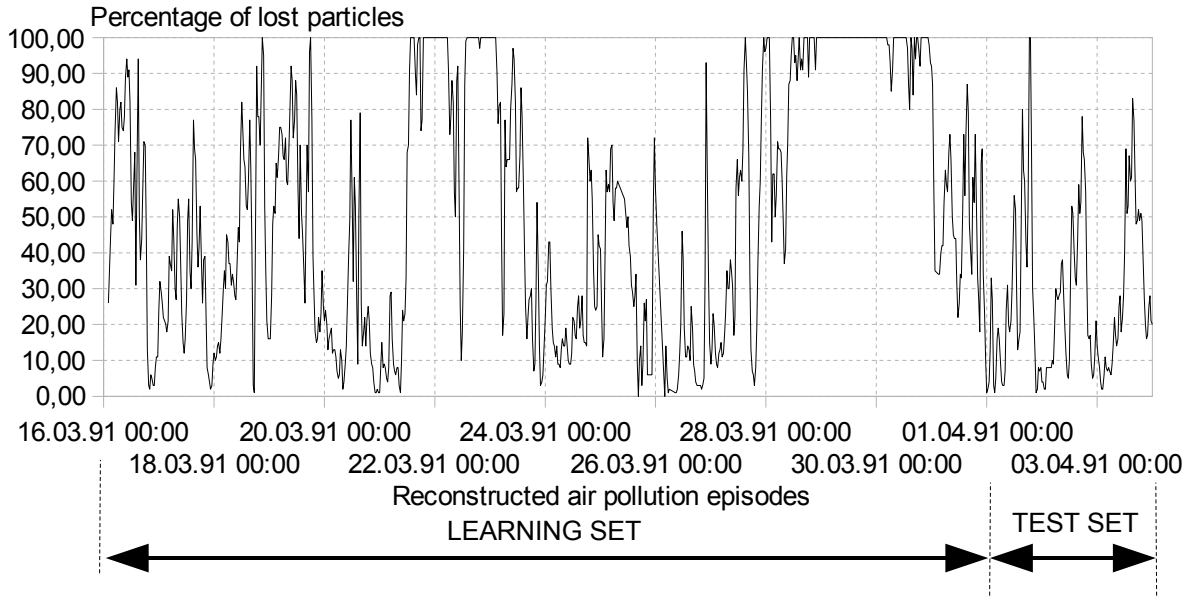


Figure 97: Percentage of lost particles during the proceeding air pollution reconstruction

For creating the neural network for the prediction of the percentage of lost particles the data records from the 1st of April 1991 at 00:00 until 5th of April 1991 at 00:00 have been excluded and used for the test set, while all the rest of the data have been used for the learning set, as presented at the bottom of Figure 97. The learning set has been divided randomly into a 10% set used for validation and the rest has been used for the training algorithm during the model's construction.

The inputs selection is a very important and sophisticated procedure within the development of the model[119]. Each air-pollution episode reconstruction is based on meteorological and emission data. The number of particles before and the air-pollution episode reconstruction depends in our case only on the meteorological data, since the emission has been set to a constant value. The meteorological data used for the reconstructions is presented in subsection *2.1.The Šaleška region field data set,* here the data from the automatic environmental measuring system is described in detail.

According to the procedures described in the literature[32,57] a heuristic determination of the inputs is made; the inputs are selected according to the modeller's knowledge of the phenomenon. In a process of input selection the optimal subset of the inputs is selected from the input space. This is achieved by eliminating those inputs that are redundant or do not have a relevant influence on the number of active particles[114].

The finally selected inputs are:

1. *the ground-level wind fluctuation,* defined as the ratio between the vector wind speed and the scalar wind speed, the value of 1.0 signifies no fluctuation, the values near 0.0 signify a strong fluctuation (where the measured wind data from the Veliki Vrh station are selected for the development of the method),

2. *the ground-level vector wind speed* (where the measured wind data from the Veliki Vrh station are used),
3. *the geostrophical vector wind speed* (where the measured horizontal vector wind speed from SODAR at a height of 350 m is used),
4. *the percentage of lost particles in the episode two steps before*
5. *the percentage of lost particles in the episode one step before.*

Construction of a model for forecasting the percentage of lost particles

After the inputs selection is made, the model is constructed using the training and validation sets of data. The finally constructed multilayer perceptron neural network (MPNN) consists of one hidden layer constructed of 10 neurons and one output layer constructed of 1 neuron. The neurons in the hidden layer are based on a tangent-sigmoid activation function and the neuron in the output layer is based on a linear activation function.

The MNPP is implemented using Matlab's Neural Network Toolbox[64] and it is trained using the Levenberg-Marquardt[70] method.

For each selected input a contribution factor is determined to have a rough measure of the importance of the input in predicting the network's output relative to the other inputs in the same network. The equation for calculating the contribution factor (4.13) is obtained from the definition[66]: "The contribution factor is the sum of the absolute values of the weights leading from a particular variable" to the first layer of hidden neurons. The higher value of the contribution factor signifies the higher contribution to the prediction, but it is very important to notice that the contribution factors for different networks cannot be compared. A special caution is given after the definition[66]: "The value of the contribution factor should not be considered as *gospel* when deciding whether to include a variable in the network. The neural nets are capable of finding data among the variables when none of the variables themselves are highly correlated to the answers. Obviously, if a certain variable is highly correlated with the answer, the variable will have a high contribution factor".

$$CF_i = \sum_{n=1}^{N} \left| (IW_{i,n}) \right| \tag{4.13}$$

CF ...contribution factor
IW ...input weight
i ...index of input feature
n ...index of neuron in first layer
N ...number of neurons in first layer

The obtained values of the contribution factors are presented in Figure 98. A comparison of the contribution factors shows that the percentage of lost particles in the last two episodes have the highest contribution to the model's output, while other input features have only a slightly smaller contribution.

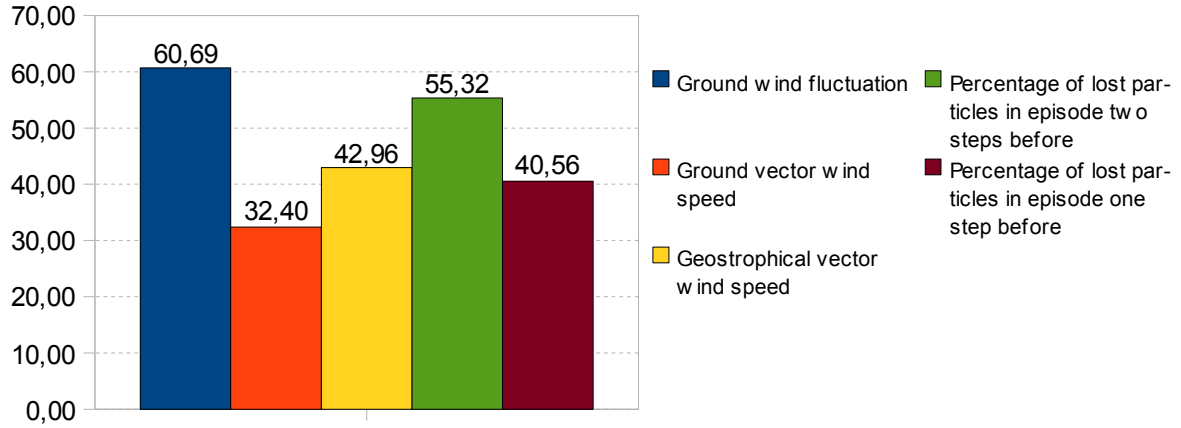


Figure 98: Contribution factors of the selected input features

Model validation

When a model has been constructed a simulation is performed on a learning set of data. The scatter plot between the measured and the predicted number of active particles is presented in Figure 99. The correlation coefficient between the measured and predicted values was *r=0.97,* the root mean square error was *RMSE=8.7* and the fractional bias was *FB=0*.

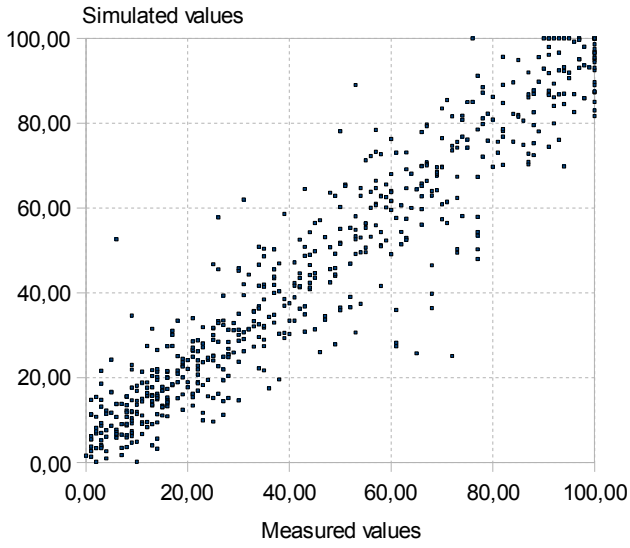


Figure 99: Scatter plot of the results of the simulation performed on the learning dataset

After the validation of the model on the learning data set another validation is performed on an independent validation set that is not used during the learning period. The scatter plot between the measured and the predicted number of active particles is presented in Figure 100, and in Figure 101 a comparison is made on a time scale. The correlation coefficient between the measured and predicted values is *r=0.95,* the root mean square error is *RMSE=7.33* and

the fractional bias is *FB=-0.01*. The correlation factor is lower in accordance with the correlation obtained in the learning process, which is the result of a relatively small learning dataset. Also, the root mean square error increases, but not significantly, while the fractional bias remains practically zero.

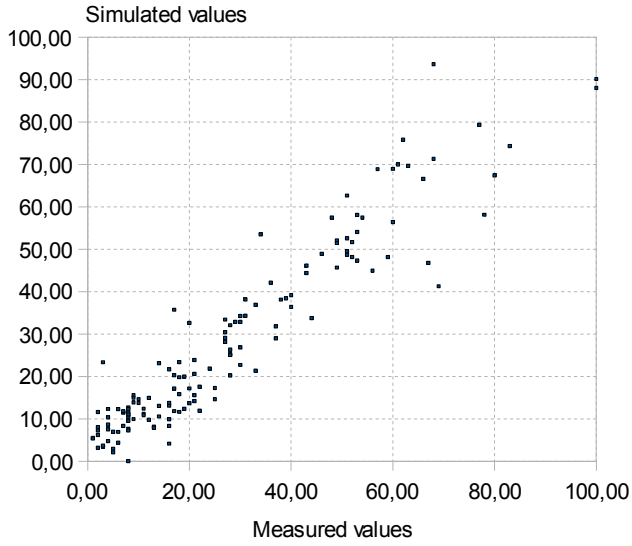


Figure 100: Scatter plot of results of the simulation performed on the validation dataset

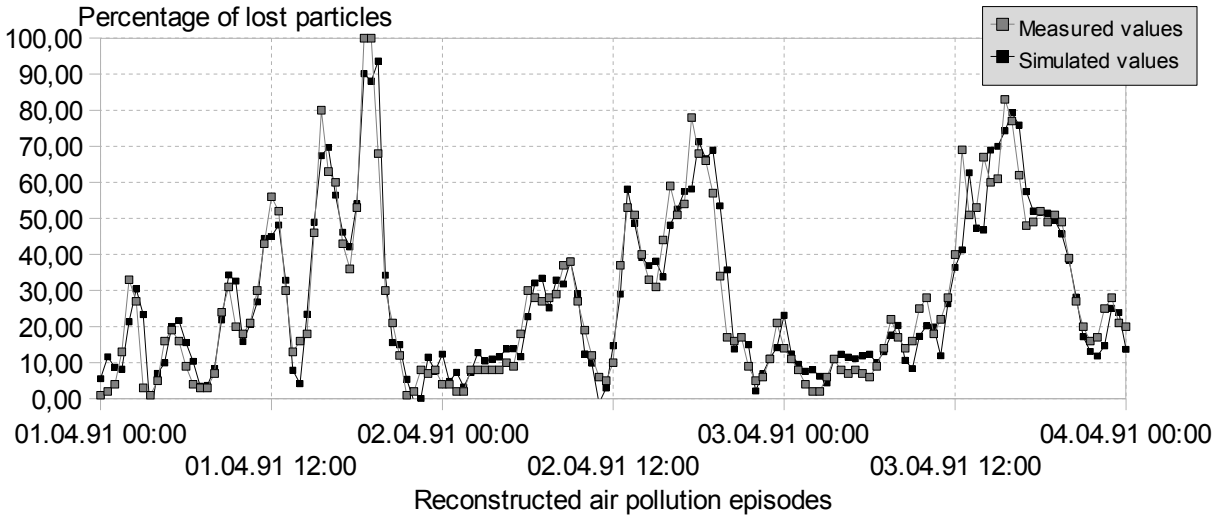


Figure 101: Time-scale comparison of the measured and predicted number of active particles

4.3.3. The *PDNC* and clustering control method

A control module is proposed to control the number of active particles during the current air-pollution episode reconstruction around the optimal range determined by the available computational resources. The inputs into the module are:

- the number of particles at the end of the previous air-pollution reconstruction episode (*Initial number of active particles)*,
- the predicted percentage of lost particles at the end of the previous air-pollution reconstruction episode provided by the *Particle number prediction method (Predicted percentage of lost initial active particles)*
- the number of particles that will be emitted during the proceeding air-pollution episode reconstruction (*Number of particles from the proceeding emission*).

The method provides two outputs:

- the first controls the c*lustering* by setting the N_{max} parameter
- the second controls the *PDNC* parameter of the LPDM.

The method is presented in Figure 102. In the first step of the procedure the optimal *PDNC* is determined to keep the predicted number of active particles in the range of available computational resources, as defined in equation (4.14).

$$PDNC = \frac{(N_{max} - N_{init} \cdot (1 - p_{lost}))}{N_{emi,learn}} \cdot PDNC_{ref} \tag{4.14}$$

$PDNC$ ...particle density number coefficient
N_{max} ...maximum allowed number of particles according to computational resources
N_{init} ...initial number of particles from previous air pollution episode reconstruction
p_{lost} ...predicted percentage of lost number of particles
$N_{emi,ref}$ ...number of particles from following emission if reference PDNC is used
$PDNC_{ref}$ ...reference PDNC

If the determined *PDNC* exceeds the maximum acceptable *PDNC*, the optimal *PDNC* is reduced to the maximum acceptable *PDNC*. The maximum acceptable *PDNC* is exceeded in strong wind conditions, when most of the particles from the previous air-pollution episode reconstruction is quickly lost out of the domain and almost all the computational resources become available only for the particles that are emitted in the current air-pollution episode reconstruction. The clustering of particles from the previous air-pollution episode reconstruction is unnecessary.

If the determined *PDNC* exceeds the minimum acceptable *PDNC*, the optimal *PDNC* is reduced to the minimum acceptable *PDNC*. The minimum acceptable *PDNC* is exceeded in low wind conditions, when most of the particles from the previous air-pollution episode remains in the domain during the current air-pollution episode reconstruction. To avoid exceeding the computational resources, the clustering must be activated to reduce the number of particles from the previous air pollution. The input parameter N_{max} into the clustering

procedure is set according to equation (4.15). When the clustering is finished the number of particles from the previous air-pollution episode reconstruction is decreased to the $N_{control}$ number of particles.

$$N_{max} = N_{control} - \frac{PDNC_{min}}{PDNC_{ref}} \cdot N_{emi,ref} \tag{4.15}$$

N_{max} ...limit parameter of the clustering method
$N_{control}$ ...maximum allowed number of particles according to computational resources
$N_{emi,ref}$ ...number of particles from following emission if reference PDNC is used
$PDNC_{min}$ ...minimum PDNC
$PDNC_{ref}$ ...reference PDNC

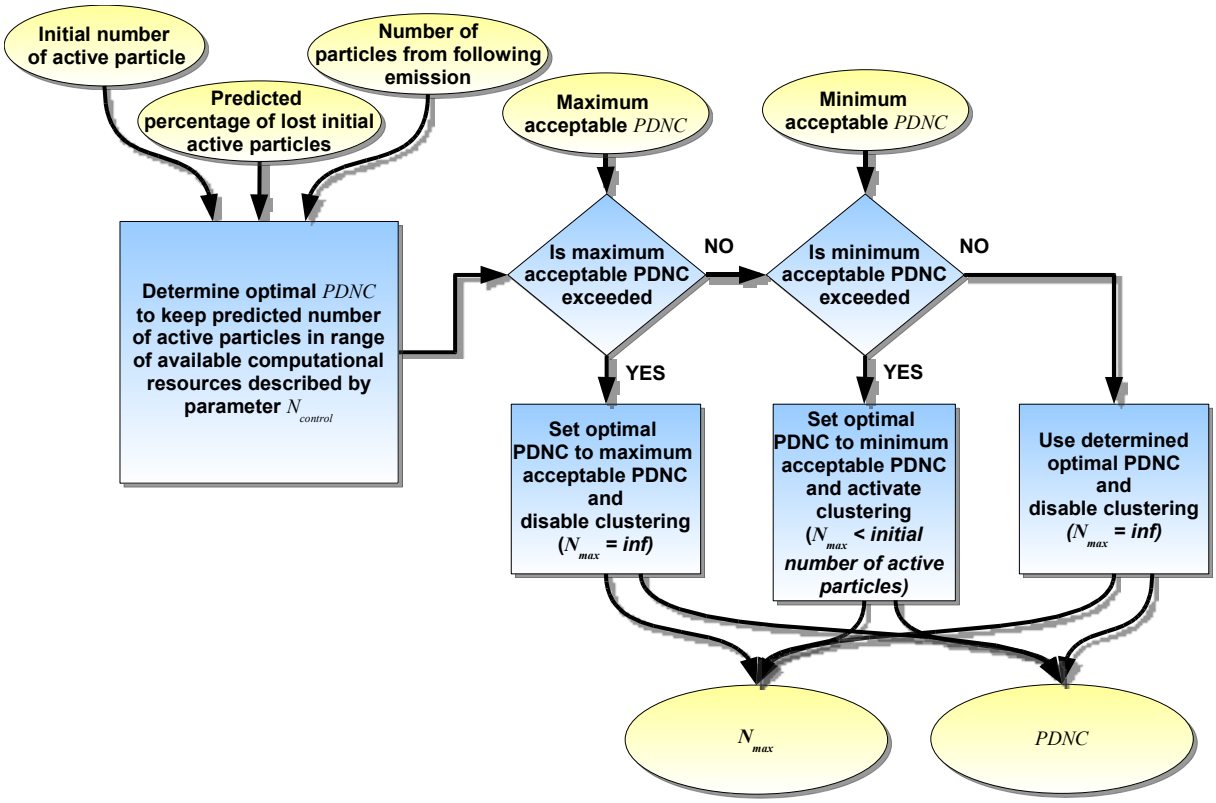


Figure 102: PDNC and clustering control method procedure where optimal PDNC and clustering parameters are determined according to initial number of particles, predicted percentage of lost particles and emission

4.3.4. Discussion

A method is proposed in this subsection to avoid situations where the number of active particles during some air-pollution episode reconstructions exceeds the computation resources. The result of such an overload is an interrupted simulation process and the reconstruction of the proceeding air-pollution episode begins with a corrupted, or in the worst case without a consideration of the previous, air-pollution situation. The high air-pollution situation episode reconstruction can be lost when such a problem occurs.

The contributed method controls the clustering parameter N_{max} and the input parameter *PDNC* of the LPD model. The rule for the *PDNC* control is the following: when high emission occurs the *PDNC* should be decreased, and when a low emission occurs the *PDNC* can be increased. In special situations, when the smallest possible *PDNC* is used and it is still expected that computational resources will be exceeded, the clustering must be activated to reduce the number of particles from the previous air pollution. In practice this is achieved by setting the N_{max} parameter of the clustering method to a value that is lower than the number of particles from the previous air-pollution episode reconstruction. To control the described *PDNC* and N_{max} clustering parameters a LPD control method is proposed. It consists of two main subsequent methods. In the first step the percentage of lost particles is predicted, based on the meteorology, the emission and the situation of the air pollution at the end of the previous episode reconstruction. In the second step the *PDNC* and N_{max} clustering parameters are determined by a decision-making method.

A method for the estimation of the percentage of lost active particles at the end of the current air-pollution reconstruction episode is constructed by an artificial neural network. It is proposed to use its predictions as the input to the control method that is developed to avoid exceeding the computational resources in a single air-pollution reconstruction episode. The prediction method is developed on the dataset obtained from the full duration of a simulation performed to reconstruct air-pollution episodes over the Šaleška valley in spring 1991. The data for the last four days of the measuring campaign are excluded and used for the test set, while all the rest of the data are used for the learning set. In the important and sophisticated procedure of input selection a heuristic determination of the inputs is made. The finally selected inputs were: *ground wind fluctuation, ground vector wind speed*, *geostrophical vector wind speed, percentage of lost particles in the episode two steps before* and the *percentage of lost particles in the episode one step before*. *The c*onstructed model is tested on an independent test set that is not used during the learning period. The correlation between the measured and predicted values in test simulation was weaker than during the learning simulation due to the relatively small set of available data. But the results are still very satisfactory for the simple control of the *PDNC* to ensure that the computational resources are not exceeded during the air-pollution episodes' reconstructions.

A decision-making method is proposed to control the number of active particles during the current air-pollution episode reconstruction around the optimal range determined by the available computational resources. The inputs into the module are: the number of active particles at the end of the previous air-pollution reconstruction episode, the predicted

percentage of lost particles at the end of the previous air-pollution reconstruction and the number of particles that are emitted during the proceeding air-pollution episode reconstruction. The module provides two outputs, where the first controls the N_{max} clustering parameter and the second controls the *PDNC* parameter of the LPD model. The complete method is integrated in the advanced AP computer model and validated in the following Section 5.

5. INTEGRATION OF THE PROPOSED IMPROVEMENTS IN THE COMPUTER MODEL

5.1. Enhanced Lagrangian particle-dispersion computer model

Three new methods are proposed in previous subsections to be used for the improvement of the air-pollution modelling methodology based on the Lagrangian particle-dispersion model: *4.1.Clustering method*, *4.2.Method for estimation of a cell concentration based on kernel density* and *4.3.Lagrangian particle-dispersion control*.

To achieve the best computational performances and to successfully avoid situations when the computational resources could be overloaded, all three new proposed methods are integrated into the new, enhanced Lagrangian particle-dispersion (ELPD) model, as presented in Figure 103. The improvements are emphasized in Figure 103 with a red colour. It is possible to use only one of the proposed methods, but it is suggested to use the contributed methods together to obtain the best possible results.

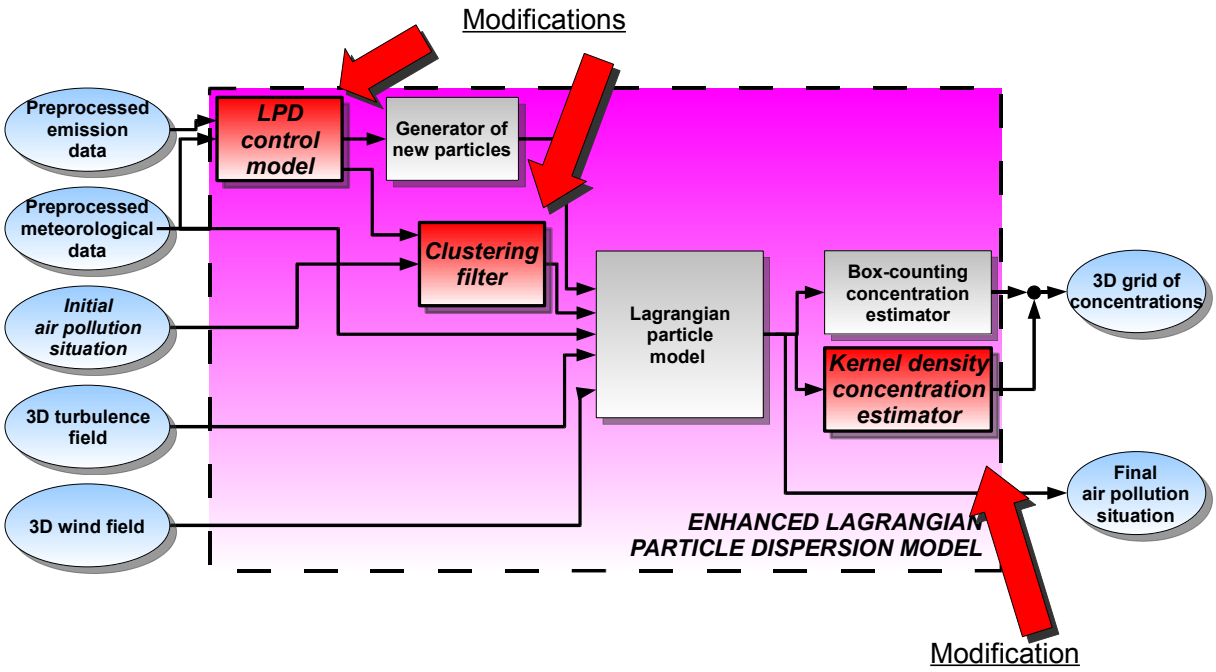


Figure 103: Enhanced Lagrangian particle-dispersion model structure

On the basis of the presented ELPD model in Figure 103 the existing LPD computer model is enhanced, as presented in Figure 104. The new modules of the ELPD computer model are presented in a red colour.

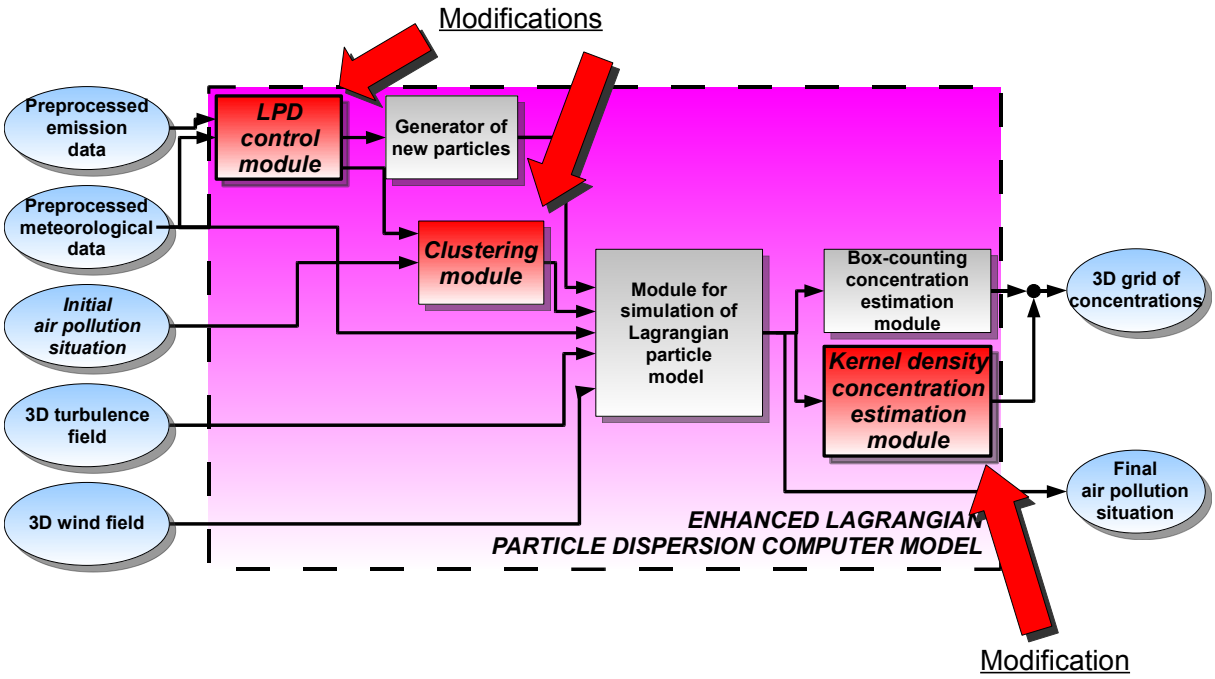


Figure 104: Enhanced Lagrangian particle-dispersion computer model

LPD control module

A control module is the centre of the integration and it is used to control the number of active particles during the current air-pollution episode reconstruction around the optimal range determined by the available computational resources, defined by the parameter $N_{control}$. The inputs into the module are the current meteorological conditions, the emission and the number of active particles at the end of the previous air-pollution reconstruction episode. The module provides two outputs, where the first controls the *Clustering method* by setting the N_{max} parameter and the second controls the *PDNC* of *Lagrangian particle model*. The module consists of two main subsequent algorithms. In the first step the percentage of lost particles is predicted with the use of an artificial neural network based on the meteorology, the emission and the situation of the air pollution at the end of the previous episode reconstruction. In the second step the clustering parameters are determined by a decision-making algorithm.

An algorithm for the estimation of the percentage of lost particles at the end of current air-pollution reconstruction episode is realised with an artificial neural network. To achieve the best possible results it is recommended to determine the parameters of the module for each complex terrain separately, because of their dependence on the domain size and the shape and position of the ground stations and the SODAR. The inputs into the algorithm are: *ground wind fluctuation, ground vector wind speed, geostrophical vector wind speed, percentage of lost particles in episode two steps before* and *percentage of lost particles in episode one step before.* The output of the model is the estimated percentage of lost active particles at the end of the current air-pollution episode reconstruction.

Clustering module

A clustering module is proposed to control the maximum number of particles in the domain where certain particles are joined according to some rules into new, heavier particles. The new, heavier particles are introduced into the domain and the old, lighter particles are removed. The properties of the new particles are composed of the properties of the old, lighter particles. To achieve satisfactory results with the clustering algorithm, four basic parameters must be set to the optimal values that were recommend in the subsection *4.1.Clustering method* where the method is also described in detail: $N_{sub}=2$, $N_{size}=5$, $m_{max}=0.1$ and $N_{max}=\{determined\ by\ LPD\ control\ module\}$. The input into the clustering module is a file containing the previous air-pollution situation reconstruction and the output is the modified file where the number of particles is decreased according to the instructions from the *Control method*. This modified file is used to describe the initial state of the air pollution over the domain and is used as the input into the existing LPD computer model.

Density Kernel Concentration Estimation module

The air-pollution modelling methodology based on the Lagrangian particle-dispersion model estimates the ground-level concentrations by counting the particles in a cell. A new, advanced density kernel concentration estimation method is proposed to improve the results of the "box counting" ground-level concentration estimations. Before the reconstructions are performed with the advanced AP computer model, several experiments must be made to determine the optimal input parameters σ_x, σ_y *and* σ_z. It is recommended that several air-pollution episode reconstructions are performed with the maximum acceptable *PDNC* and again with the minimum acceptable *PDNC*. When the results of both simulations are available a comparison is made to obtain correlation coefficients between the results. An air-pollution episode reconstruction where the correlation is very good should be taken and used to set the optimal values of the input parameters, as presented in subsection *4.2.Method for estimation of a cell concentration based on kernel density*. The optimal values are determined by using the density kernel concentration estimation method with different input parameters σ_x, σ_y *and* σ_z on a air-pollution episode reconstruction with the maximum acceptable *PDNC*. Each result is compared with an air-pollution episode reconstruction with the minimum acceptable *PDNC*. Those input parameters where the highest correlation is achieved are finally used as the optimal input parameters.

5.2. Air-pollution simulation with an enhanced Lagrangian particle-dispersion computer model

The air-pollution episode reconstruction based on the original LPD computer model consists of three main steps, as already presented in Figure 41. Due to new, integrated modules a single air-pollution episode reconstruction changed according to the presentation in Figure 105, where additional time is spent for the LPD control, clustering and the kernel density cell concentration estimations.

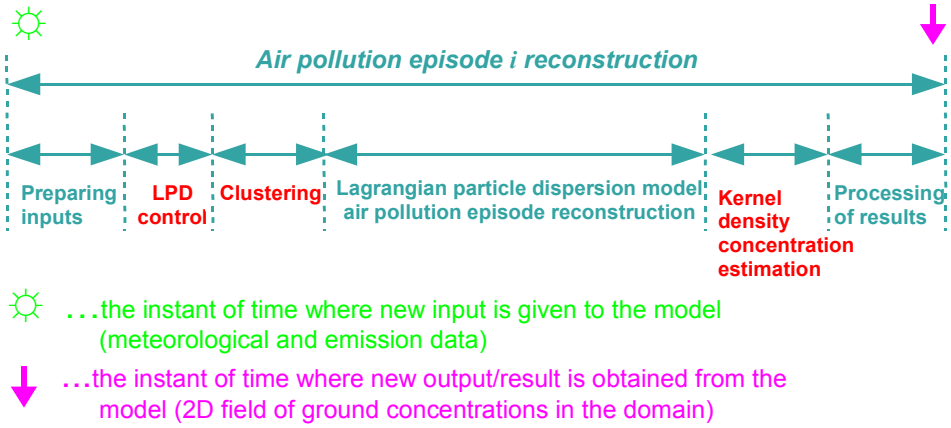


Figure 105: A single air-pollution episode reconstruction by the ELPD computer model

5.3. Validation of the enhanced air-pollution computer model

The advanced computer APM based on the enhanced LPD computer model presented in the previous subsection is validated in this subsection. All the parameters of the ELPD computer model are determined during the demonstrations of new methods in previous section *4.Proposed improvements in air pollution modelling methodology*. To demonstrate and validate the performances of the integrated methods, three simulation runs for the selected Šaleška region field data set air-pollution situation are performed and compared with each other:

- the reference simulation run is performed using the original LPD computer model where the maximum acceptable *PDNC=20* is used, usually in practice such a high *PDNC* parameter is not used because the computational complexity can sometimes fall outside the computational capabilities and also there is not enough time available (in the on-line system for each ½ hour the air-pollution episode reconstruction less than ½ hour is available and in the off-line system for the assessment of the air-pollution situation for a one-year long elaboration much less than one-year of time available),
- the optimal simulation run is performed by using the original LPD computer model where the optimal *PDNC=2 i*s used,
- the advanced simulation run that is performed with the improved ELPD computer model, where the *PDNC i*s controlled in the range from the minimum acceptable value of *0.04* and the maximum acceptable value of *10*.

In all the following presentations the results of the *reference simulation* are coloured in *black*, the results of the *optimal simulation* are coloured in *blue*, and the results of the *advanced simulation* are coloured in *red*.

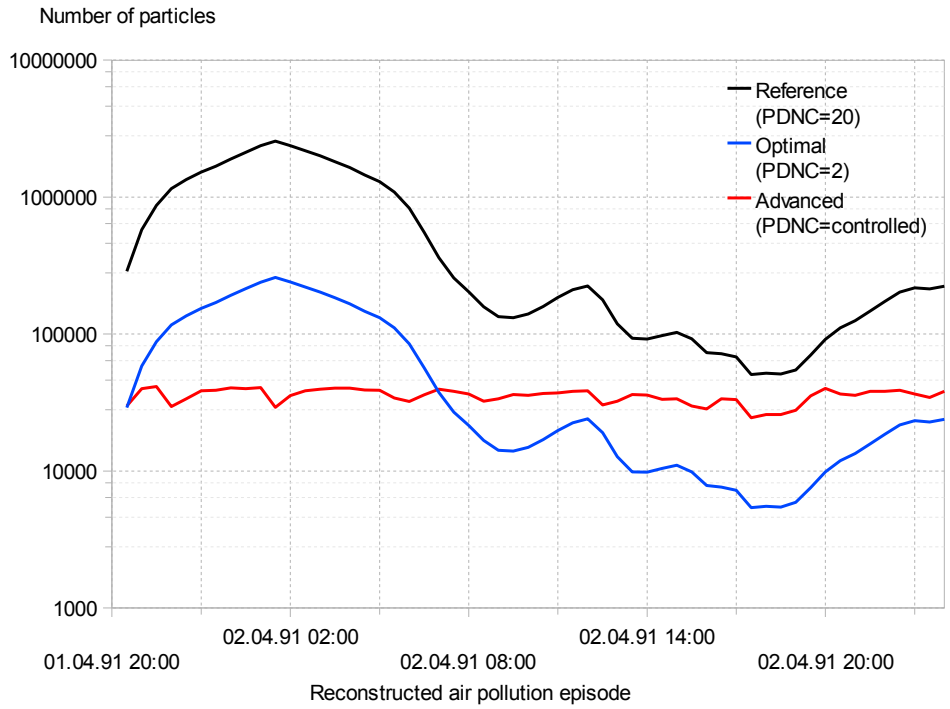


Figure 106: Number of active particles after each air pollution episode reconstruction

The number of particles at the end of the reconstruction of each air-pollution episode is presented in Figure 106. A comparison with the following Figure 107 shows a linear dependence between the time used and the number of active particles. The figure shows that the active particle number in the first and second simulation is strongly dependent on the current meteorological conditions, while in the third controlled simulation it is practically constant.

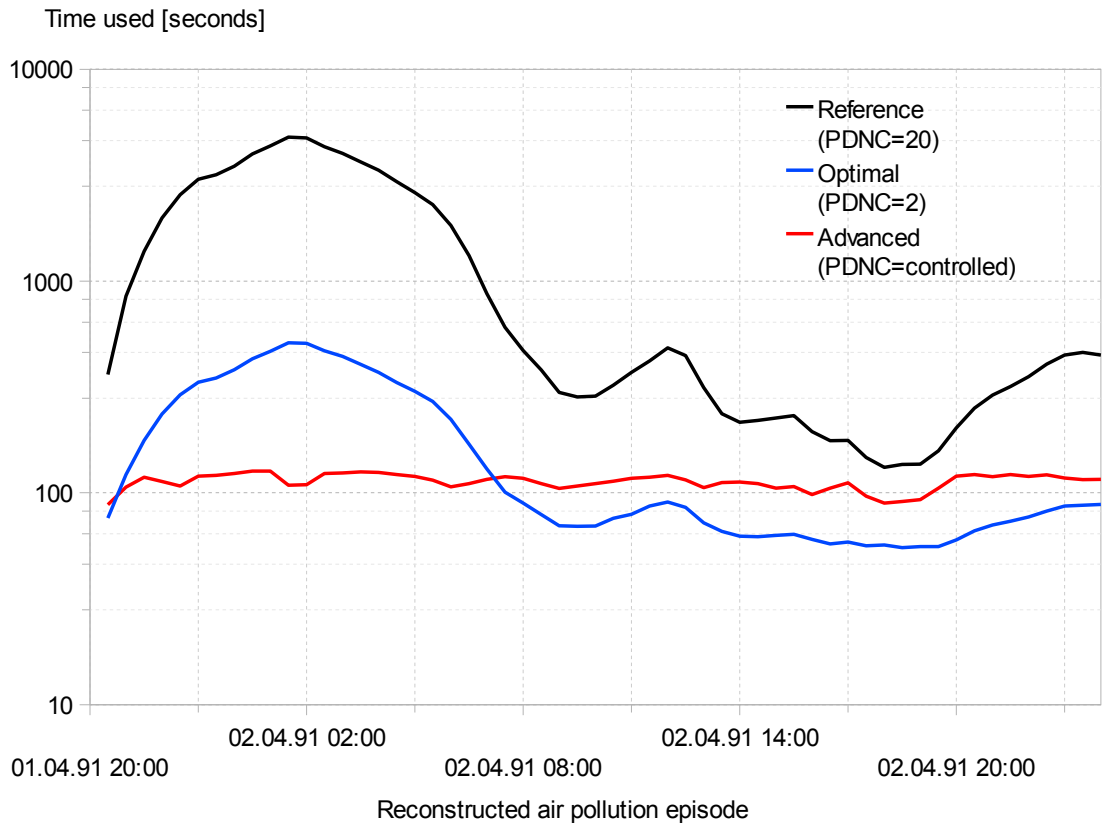


Figure 107: Comparison of time used for reconstructions of the air-pollution episodes

The results of the time used for each simulation run are presented in Figure 107. The figure shows that the time used in the first and second simulation runs strongly depends on the current meteorological conditions, where the low-wind meteorological conditions occur in the middle of the selected time period and last until the end of the simulation. While in the third controlled case where the ELPD computer model is used, the time used for each air-pollution episode reconstruction is practically constant during the whole simulation run.

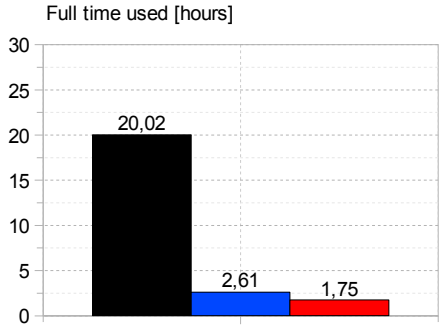


Figure 108: Comparison of full time used

A comparison of the full time used for each simulation run is presented in Figure 108. The results show that the full time used within the advanced simulation run is decreased by one order of magnitude compared to the reference simulation and by approximately 30% compared to the optimal simulation run.

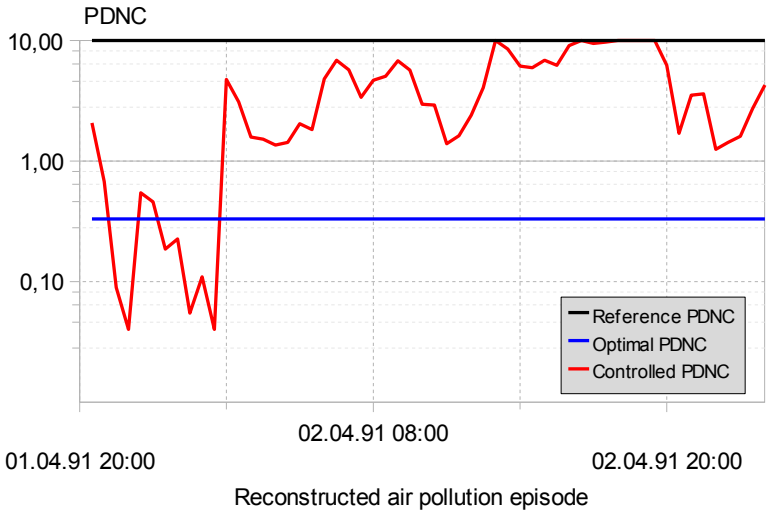


Figure 109: Comparison of PDNC during the evaluation simulation

Figure 109 shows the *PDNC* of each air-pollution episode reconstruction for all the simulation runs. In the case of the reference simulation run and the optimal simulation run it was set to a constant value, while in the advanced simulation run it was determined by the control module. At the beginning a smaller value of the *PDNC* is used due to the high emission and the calm meteorological conditions. And when the emission decreases and stronger wind conditions appear in the second part of simulation run, higher *PDNC* values are used.

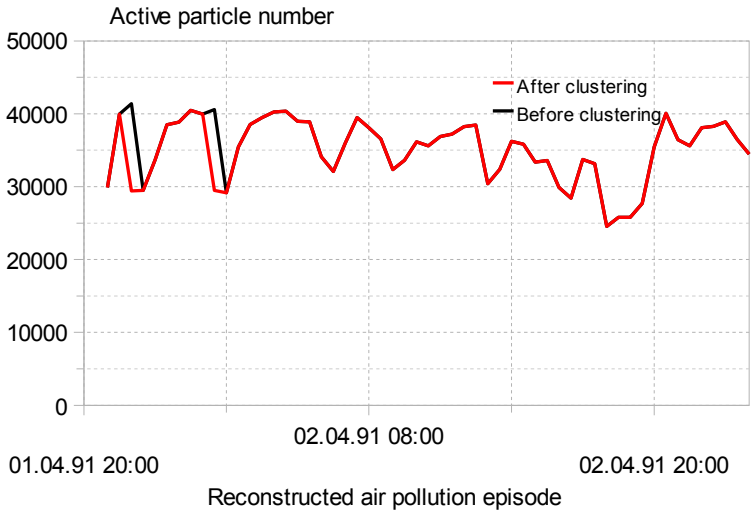


Figure 110: Comparison of the number of active particles before and after clustering

In Figure 110 the number of particles at the end of each air-pollution episode reconstruction and the number of particles after clustering are also presented. Where the values differ the clustering is activated. The results show that clustering is activated only in a few situations and when it is activated the number of reduced particles is relatively small according to the number of all the particles. But the final influence of the clustering is crucial when there is no other way to prevent the abnormal increase in the number of particles. Such a situation can occur in a calm meteorological situation when the air pollution accumulates in the domain.

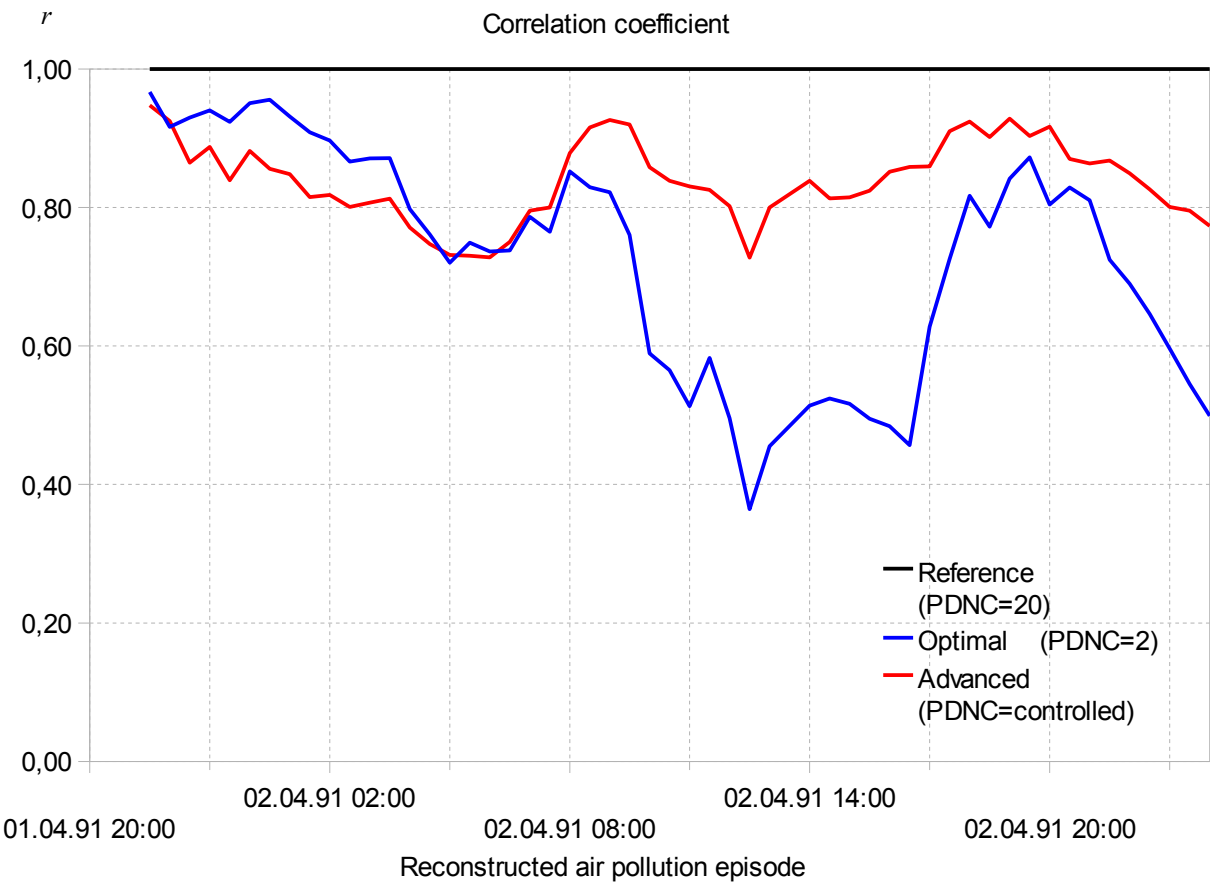


Figure 111: Correlation coefficient comparison between the results of the reference simulation and the optimal and advanced simulation results

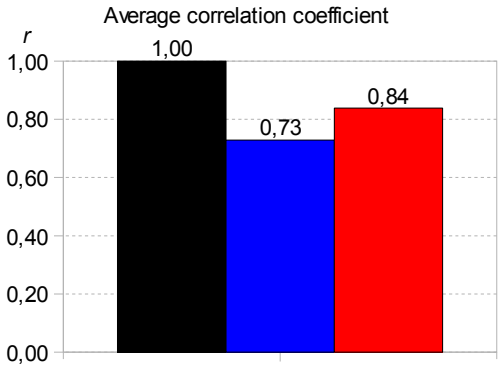


Figure 112: Average correlation-coefficient comparison

Figure 111 shows the correlation coefficient between each optimal and advanced simulation result and the reference simulation results are presented. The results of the comparisons show a significant improvement in the correlation coefficient when the ELPD computer model is used, which can also be observed in Figure 112, where the average correlation coefficients for all the simulation episode results are presented.

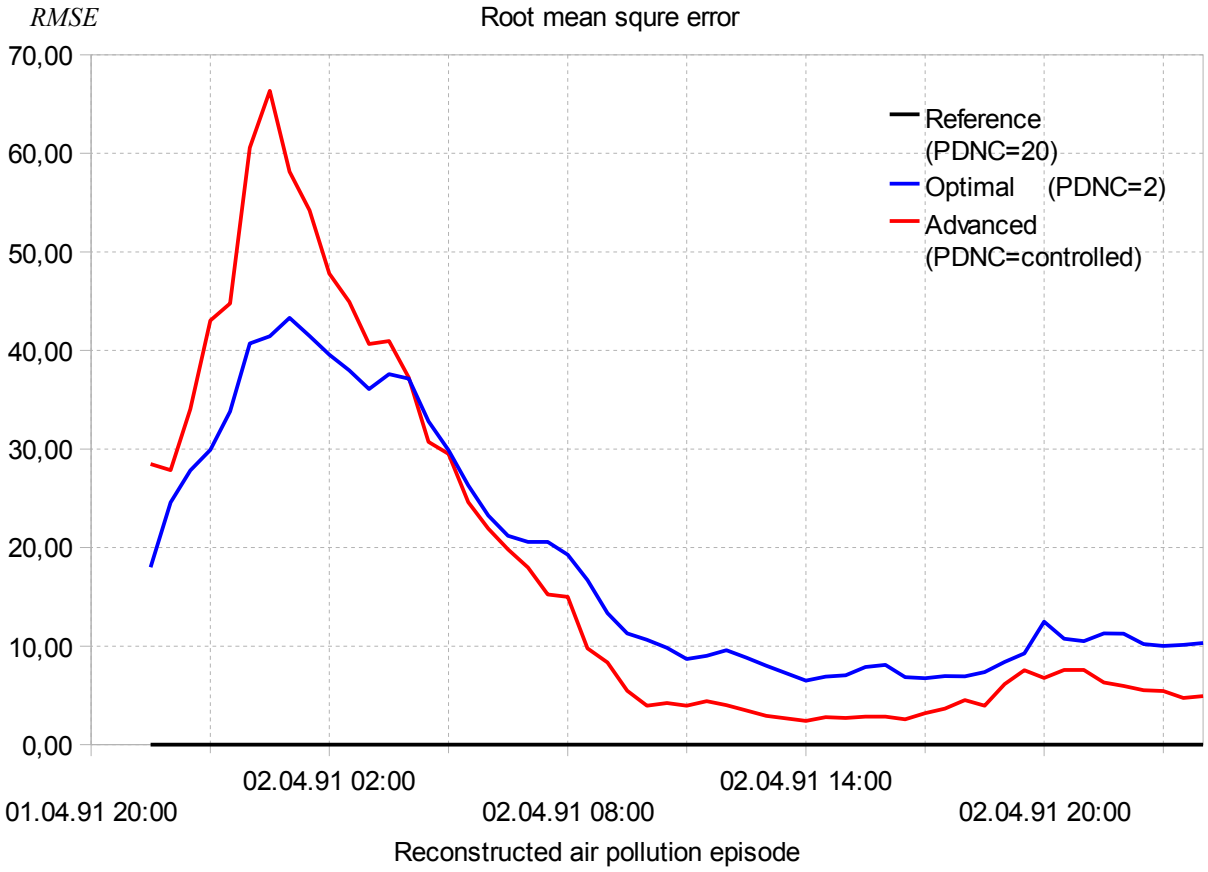


Figure 113: Root mean square error comparison between the results of the reference simulation and the optimal and advanced simulation results

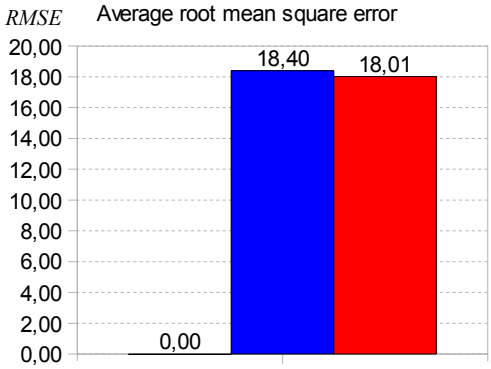


Figure 114: Average root mean square error comparison

Figure 113 presents the root mean square error between each optimal and advanced simulation result and the reference simulation results. The results of the comparisons show that during high air pollution a lower value of the root mean square error is achieved with the optimal simulation. The difference is compensated in the second part of the simulation when the root mean square error in the advanced simulation is 50% lower than in the optimal simulation. The same results are obtained in Figure 114, where the average root mean square error of the advanced simulation run result is approximately the same as the optimal

simulation results.

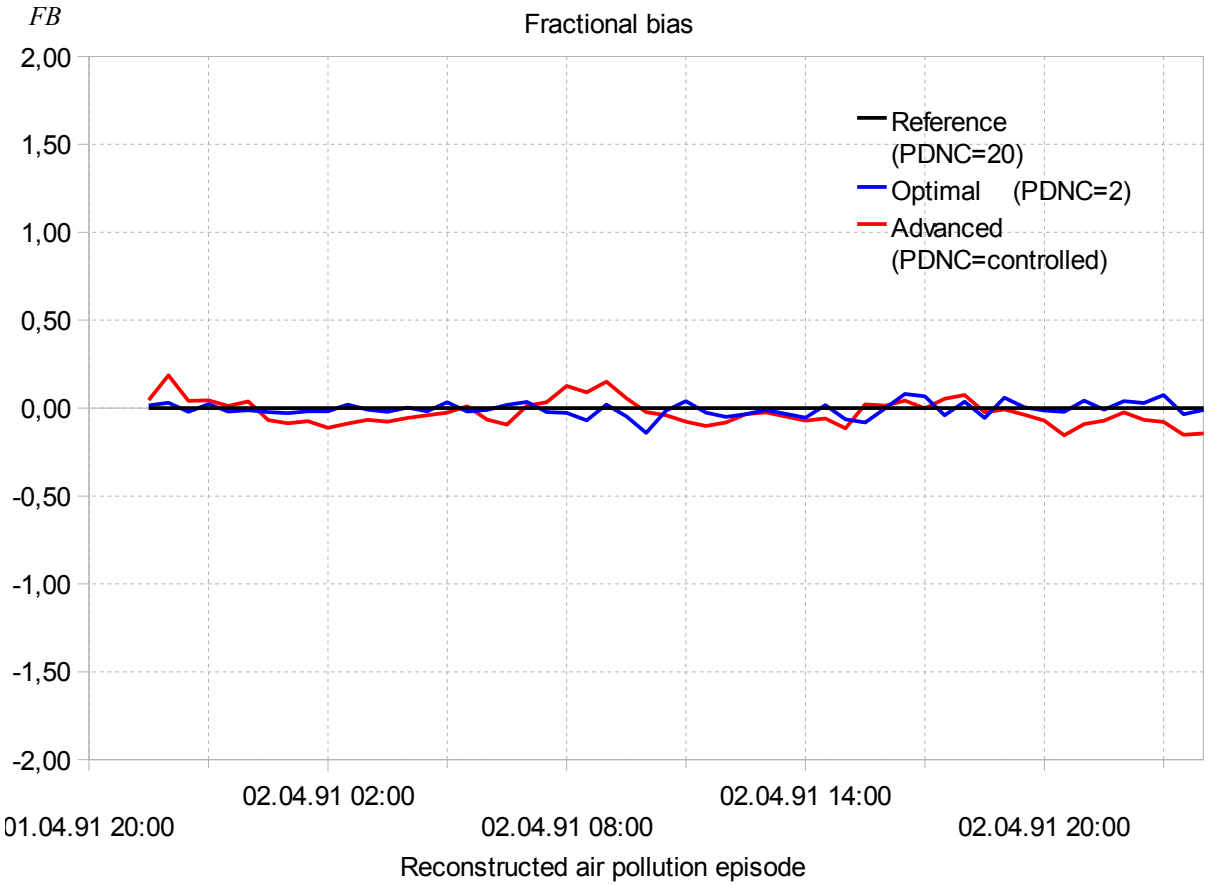


Figure 115: Fractional bias comparison between the results of the reference simulation and the optimal and advanced simulation results

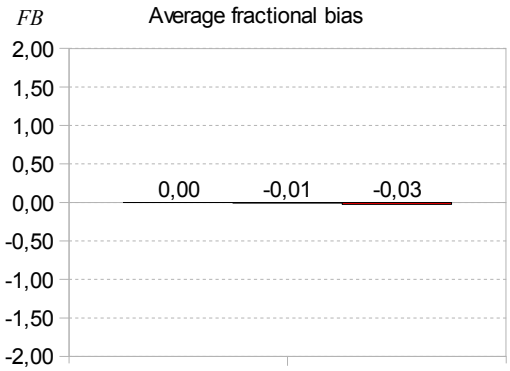


Figure 116: Average fractional bias comparison

Comparison of the fractional bias presented in Figure 115 shows that the results of the simulation with the ELPD computer model are very slightly underestimated in the first part of the simulation, where the limitation process occurred. According to Figure 116 the results of the average values show that there is practically no under or over estimations of concentrations.

5.4. Discussion

To avoid situations where computational resources could be overloaded and to achieve the best computational performance the integration of three new contributed methods into the existing air-pollution modelling methodology based on the Lagrangian particle dispersion is proposed: *LPD control method, clustering method* and *density kernel concentration estimation method.* Based on the proposed advanced air-pollution modelling methodology all the presented methods are realised and integrated into the new enhanced Lagrangian particle dispersion (ELPD) computer model.

The performance of the advanced AP computer model based on the ELPD computer model with integrated new modules are demonstrated on a selected Šaleška region field data set air-pollution situation. The parameters of the ELPD computer model that are used are those already determined during the demonstrations of the new contributed methods in previous sections. Three simulations for the selected situation are performed and compared with each other: the reference simulation with the original LPD computer model, where the maximum acceptable *PDNC* was used, the optimal simulation with the original LPD computer model, where the optimal *PDNC* is used, and the advanced simulation with the ELPD computer model, where the *PDNC* is controlled.

The results of the comparisons show that the time used and the particle number in the reference and optimal simulations strongly depended on the current meteorological conditions and the emission rate, while in the advanced simulation it is practically constant. A significant improvement of the correlation coefficients is achieved by using the ELPD computer model. Also, the fractional bias comparison showed that there is practically no underestimations in the results of the advanced simulation, which is crucial for any long-term evaluations of air pollution. The obtained results show that the use of the ELPD computer model is recommended, not only in situations where the computational resources are constrained, but also in general to optimally take advantage of the available computational resources.

6. VALIDATION OF THE ENHANCED LAGRANGIAN PARTICLE-DISPERSION COMPUTER MODEL

6.1. Introduction

All the proposed methods in the previous sections are validated on a complex terrain field data set air-pollution situation that is based on an experimental measuring campaign performed over the Šaleška region. To evaluate the performance of the new, advanced AP computer model with the integrated new methods and to confirm that the proposed methods can be generally applied in various complex terrain systems, an evaluation was performed on a field data set based on an experimental measuring campaign performed over the Zasavje region, which is introduced in subsection *2.2.Zasavje region field data set*. For the evaluation a typical complex terrain air-pollution situation is selected, which lasted from the 14th of October 2005 at 00:00 until the 16th of October 2005 at 00:00. The selected air-pollution situation is described in detail in subsection *2.2.5.Situation selection from the Zasavje region field data set.*

6.2. Determination of the parameters for the lost particle number prediction method

The parameters of the method for estimating the percentage of lost particles at the end of the current air-pollution reconstruction episode are readjusted. The new parameters of the method for each complex terrain must be determined because the domain size and shape can be very different. To achieve the best possible results it is recommended to determine the method parameters for each complex terrain separately.

The optimal parameters of the method are determined from the data obtained from the input and output parameters of the simulations that are performed on the Zasavje region field data set air-pollution situation that lasted from the 1st of September 2005 at 00:00 until the 1st of October 2006 at 00:00, where 16177 data are available. A relatively large set of data is fully used for the neural network construction. The emission is a constant value during the complete simulation, where 3780 particles are emitted in each air-pollution episode from a single source. The number of active particles after each air-pollution episode reconstruction is presented in the graph in Figure 117. The presented results show that some episodes of the strong air pollution over the domain occurred when the number of active particles strongly increased. From the number of active particles the percentage of lost particles in each air-pollution episode (*output*) is determined according to equation (4.12) defined in subsection *4.3.2.Method for forecasting the percentage of lost particles*. It is presented in Figure 118, which is consistent with Figure 117: in the episodes with low air pollution, when strong wind conditions are present, all the particles (100 %) are lost from the domain and in the episode with a high air pollution, when low wind conditions occurred, fewer particles (< 100%) are lost from the domain. In the most critical situations, fewer than 10% of the particles are lost.

Logs from the 1st of September 2005 at 00:00 until the 15th of October 2005 at 00:00 are excluded for the determination of the optimal parameters and used for the test set, while all the rest of the data are used for the learning set. The learning set is divided randomly into a 10% set used for validation, and the rest are used for the training algorithm during the model

construction.

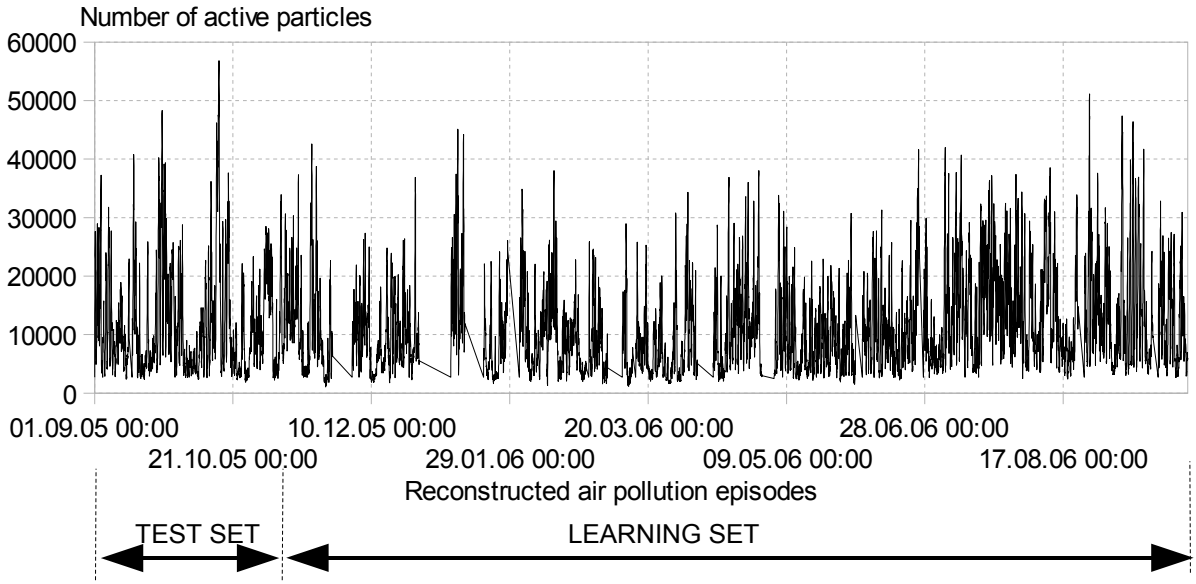


Figure 117: Number of active particles used to represent the complexity of reconstructed air-pollution episodes

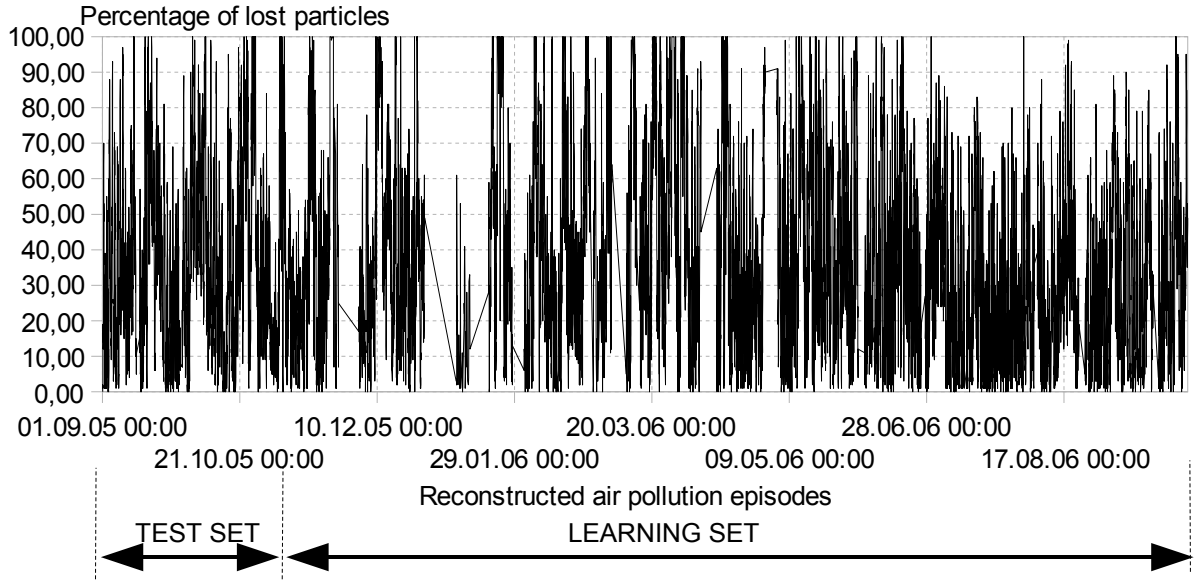


Figure 118: Percentage of lost particles during the proceeding air-pollution reconstruction

The inputs are modified only according to the different available measurements, but generally they remain the same:

1. *ground wind fluctuation* (measured wind data from Kovk station),
2. *ground vector wind speed* (measured wind data from Kovk station),
3. *geostrophical vector wind speed* (measured horizontal vector wind speed from SODAR at a height of 350 m),

4. *percentage of lost particles in the episode two steps before*
5. *percentage of lost particles in the episode one step before.*

After the inputs are selected, the model parameters are determined using the training and validation datasets. Finally, the constructed multilayer perceptron neural network (MPNN) consists of one hidden layer, constructed of 10 neurons, and one output layer, constructed of 1 neuron. The neurons in the hidden layer are based on a tangent-sigmoid activation function and the neuron in the output layer is based on a linear activation function.

The MNPP is implemented using Matlab's Neural Network Toolbox[64] and it is trained using the Levenberg-Marquardt[70] method.

For each selected input a contribution factor is determined to have a rough measure of the importance of the input in predicting the network's output relative to the other input features in the same network. For calculating the contribution factor an equation (4.13) was used from subsection *4.3.2.Method for forecasting the percentage of lost particles*.

The obtained values of the contribution factors are presented in Figure 119. A comparison of the contribution factors shows that the geostrophical vector wind speed has the highest contribution to the model's output. The second-highest contribution is from the percentage of lost particles one step before, while the other input features contribute only slightly less to the model's output.

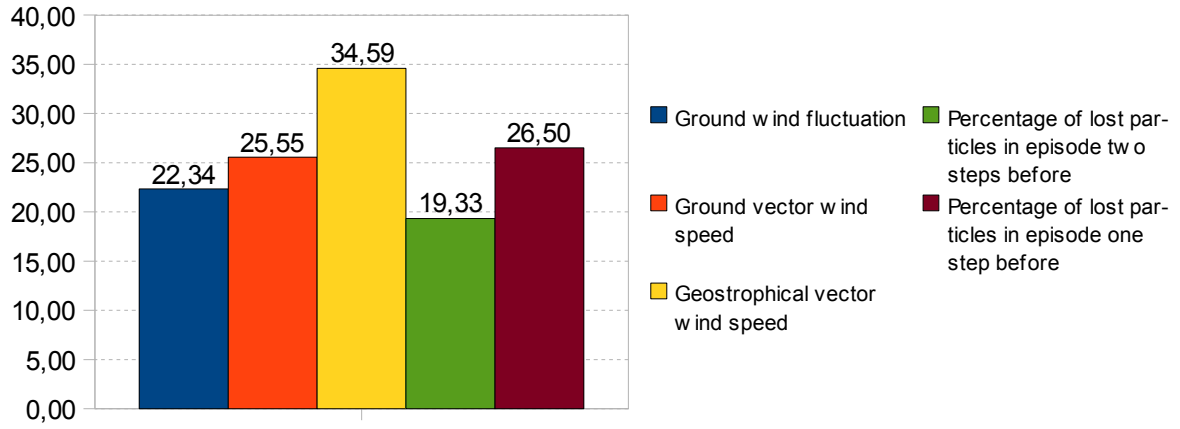


Figure 119: Contribution factors between the selected input features and the percentage of lost particles at the end of the air-pollution episode

When the parameters are determined, a simulation is performed on the learning set of data. The scatter plot between the measured and predicted number of active particles is presented in Figure 120. The correlation coefficient between the measured and predicted values is *r=0.94,* the root mean square error is *RMSE=8.98* and the fractional bias is *FB=0.00*.

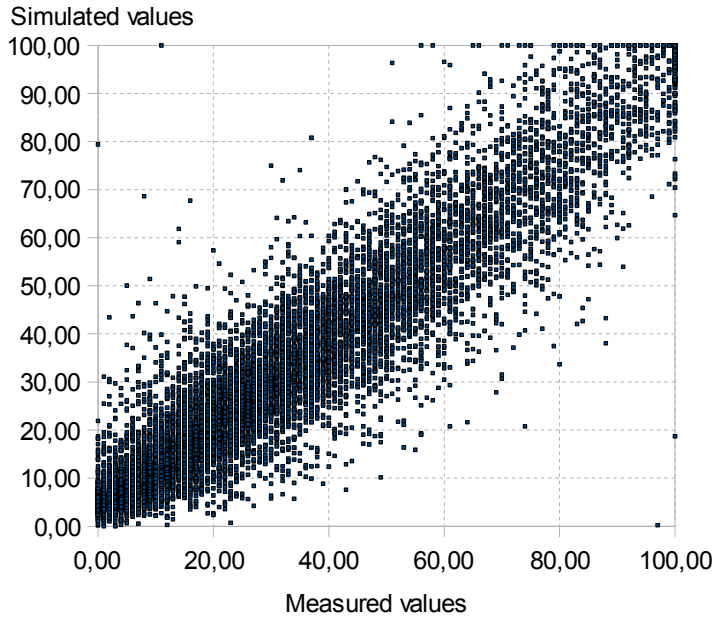


Figure 120: Scatter plot of the results of the simulation performed on a training dataset

Another simulation is also performed on an independent validation data set that is not used during the learning process to evaluate the model. The scatter plot between the measured and predicted number of active particles is presented in Figure 121. The correlation coefficient between the measured and predicted values is $r=0.95$, the root mean square error is $RMSE=8.07$ and the fractional bias is $FB=0.02$. The obtained results are similar to the results obtained during the development of the method, which is as expected.

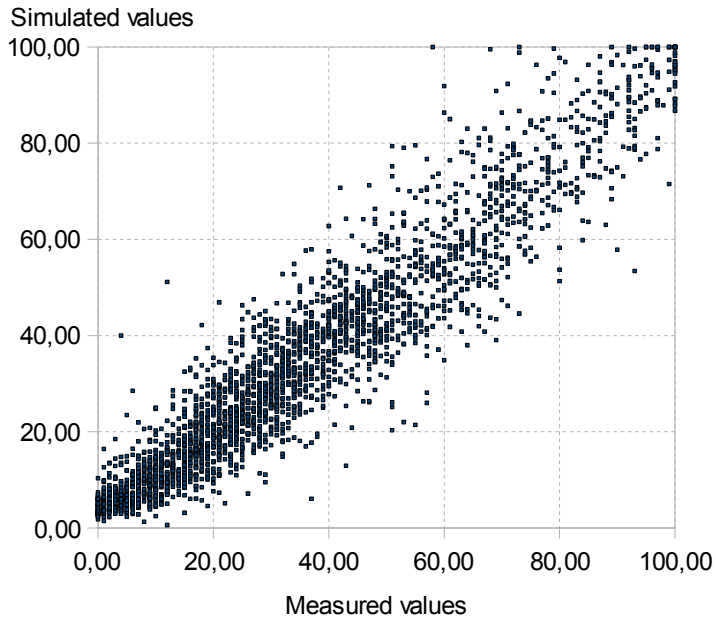


Figure 121: Scatter plot of the results of the simulation performed on a testing dataset

In Figure 122 a special comparison is made on a time scale for the selected period of time from the 14th of October 2005 at 00:00 until the 16th of October at 00:00.

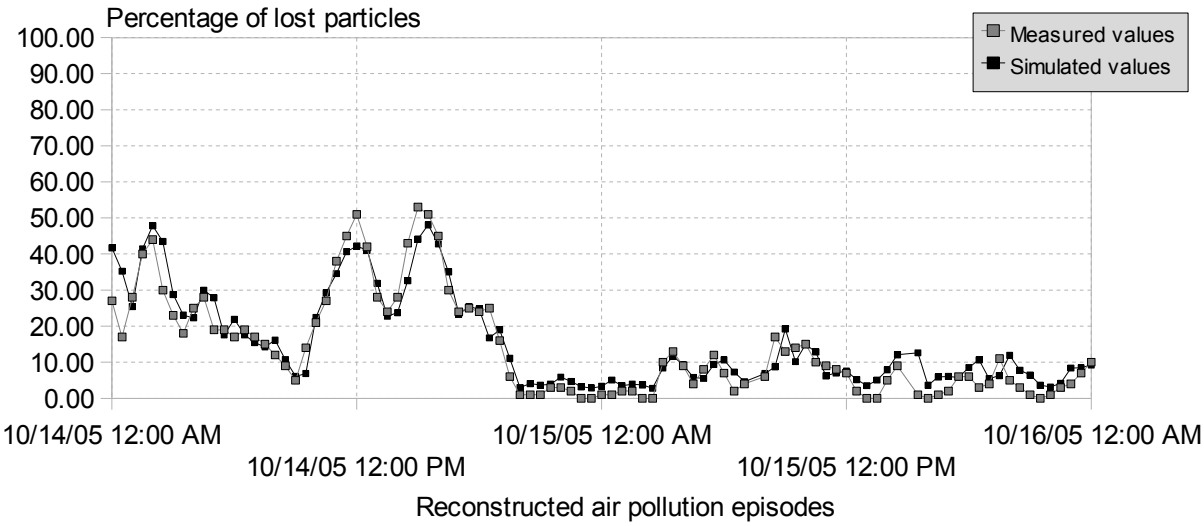


Figure 122: Time scale comparison of measured and predicted number of number of active particles for time interval from 14th October 2005 00:00 till 16th October 2005 00:00

6.3. Determination of the parameters for the density kernel concentration estimation method

The kernel density concentration estimation method represents a very important method for an advanced air-pollution modelling methodology based on LPD. It is integrated to improve the results of the "box counting" ground-level concentration estimations. To determine the input parameters of the density kernel concentration estimation module several simulation runs are performed and the results are compared to the reference ground concentration field, as defined in detail in subsection *3.5.Evaluation methods.* The simulations are performed on the Zasavje region field data set air-pollution situation that lasted from the 14th of October 2005 at 00:30 until the 16th of October 2005 at 00:00.

The first simulation is performed to generate the reference ground concentration fields for all the simulated time intervals in the validation situation. To accomplish this task, the particle number density coefficient is set to the maximum possible value of *PDNC=20,* according to the available computational resources. The results of the time used for the first simulation run are presented in Figure 126 (black curve). After that, the second simulation is performed where the particle number density is decreased to a value of *PDNC=0.33*. A comparison of the results for the first (black curve) and the second (blue curve) simulation run are also presented in Figure 126, showing that significantly less computational time is used for the second simulation run. Unfortunately, the consequence of this profit is a very poor correlation, presented in Figure 130 (blue curve), the large root mean square error is presented in Figure 132 (blue curve) and the scattered fractional bias is presented in Figure 134 (blue

curve).

Several experiments were performed to determine the optimal input parameters σ_x, σ_y and σ_z for the validation field data set air-pollution situation. In all the experiments the horizontal parameters σ_x and σ_y are set to be the same. In all presentations of the results both parameters are presented as a single independent parameter $\sigma_{x,y}$, while the second independent parameter is vertical σ_z. The conclusions of subsection *4.2.Method for estimation of a cell concentration based on kernel density* show that the optimal inputs of σ_x, σ_y and σ_z can be obtained from a single air-pollution episode reconstruction, where the correlation with the reference reconstruction is very good. According to the obtained comparison, the results of the correlation coefficient presented in Figure 130 a good correlation coefficient is obtained at a simulated time interval at the air-pollution episode of the 15th of October 2005 at 04:30 hour. In the results of the comparisons between the reference and kernel density the estimated ground concentrations are presented. The dependence of the correlation is presented on the first graph of Figure 124; the second graph presents the root mean square error and the third graph the fractional bias. From the obtained results we can see that the trend of the results is the same as those obtained in the previous case, only that different optimal values of $\sigma_{x,y}=150$ *and* $\sigma_z=20$ are determined. In Figure 123 three examples of the "density kernel estimated" ground concentrations are also presented: over-smoothed, optimal and under-smoothed.

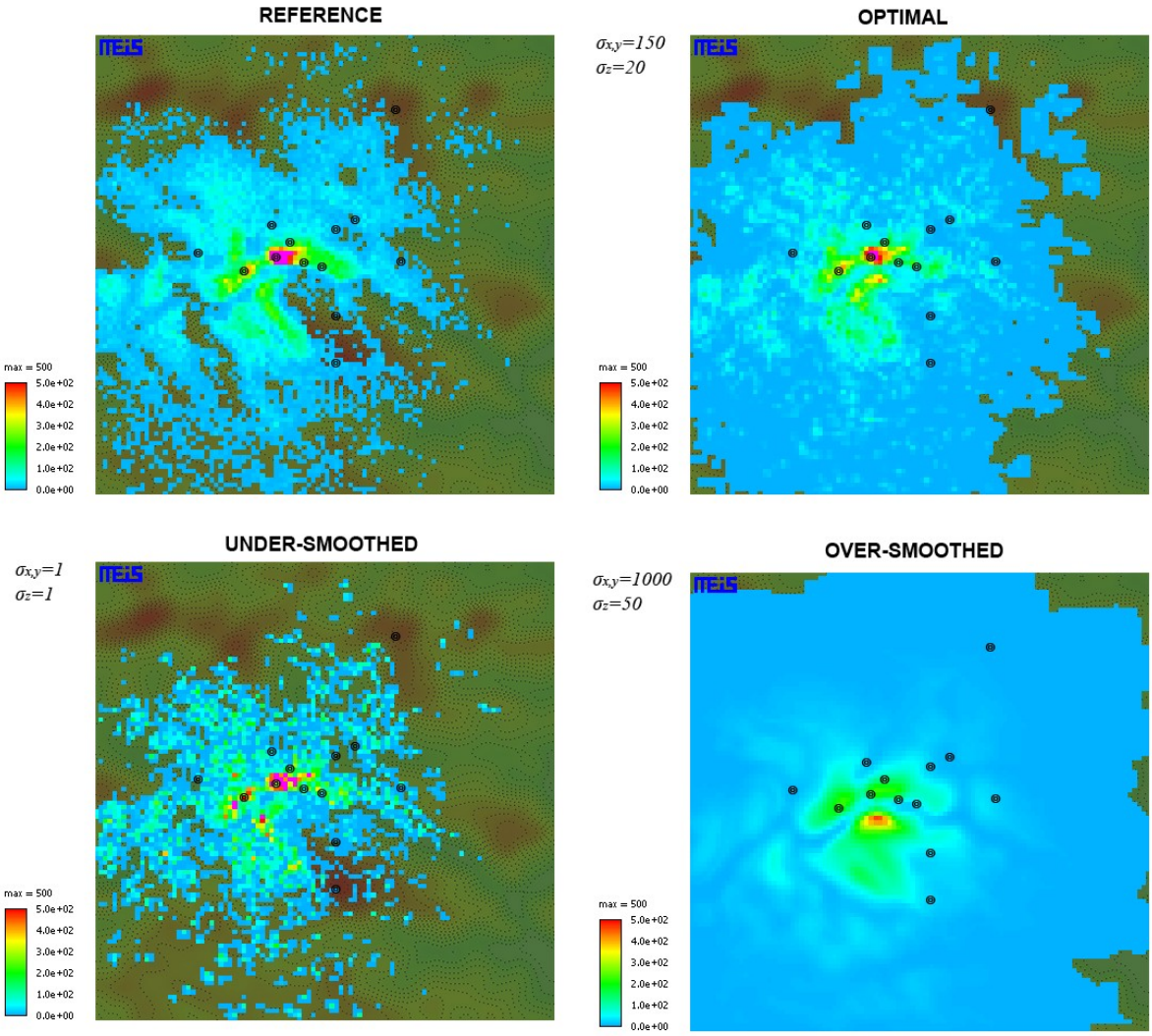


Figure 123: Examples of density kernel concentration estimations at simulated time interval n 15th of October 2005 at 04:30

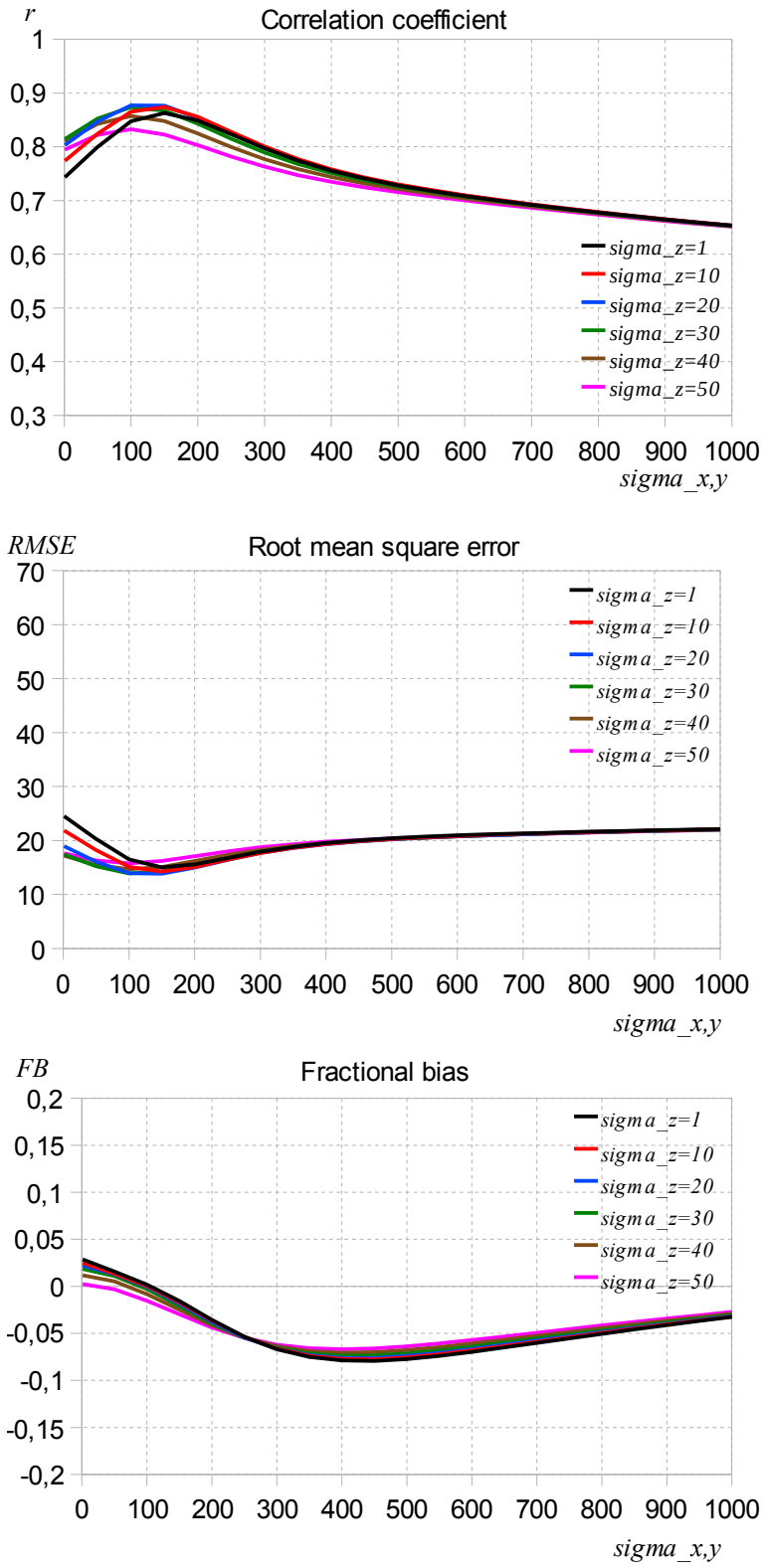


Figure 124: Comparisons of the density kernel concentration estimations for the good correlation for the air-pollution episode reconstruction on 15th of October 2005 at 04:30

6.4. Evaluation of results

When all the parameters of the advanced AP computer model are determined, three simulation runs for the selected situation are performed and compared with each other:

- the reference simulation run is performed using the original LPD computer model, where the maximum acceptable *PDNC=10* is used,
- the optimal simulation run is performed using the original LPD computer model, where the optimal *PDNC=0.33* is used, which is determined from the point of view of the time spent for the simulation. The optimal value has been selected according to the time that is spent to reconstruct the air pollution over the area of interest for a time period of one year. For the selected optimal point where 1 hours is spent for the reconstruction of one day, approximately 15 days are spent for the reconstruction of one year, which is still acceptable in practice.
- the advanced simulation run is performed using the new ELPD computer model, where the *PDNC* is controlled in the range from the minimum acceptable value of *0.04* and the maximum acceptable value of *2.0*.

In all the following presentations the results of the *reference simulation run* are coloured in *black*, the results of the *optimal simulation run* are coloured in *blue,* and the results of the *advanced simulation run* are coloured in *red*.

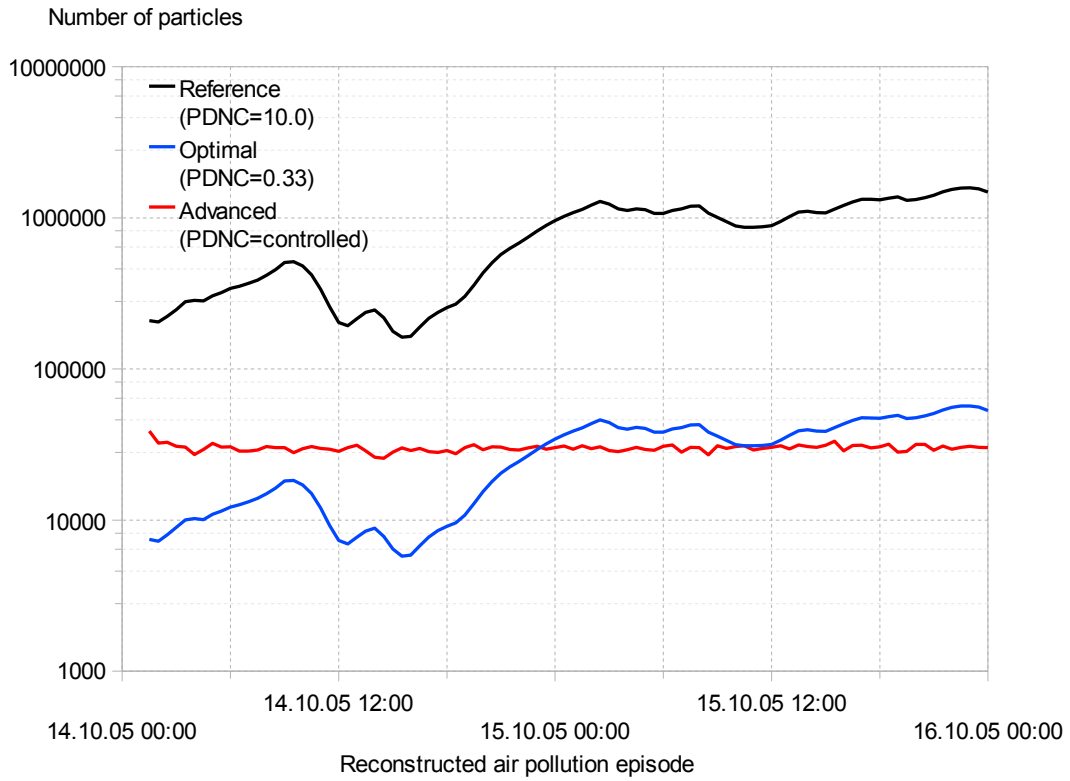


Figure 125: Number of active particles after each air pollution episode reconstruction

The number of particles at the end of the reconstruction of each air-pollution episode is presented in Figure 125. A comparison with the following Figure 126 shows a linear dependence between the time used and the number of active particles. Again, the figure shows that the active particle number in the first and second simulation runs strongly depended on the current meteorological conditions, while in the third controlled simulation run it is practically constant.

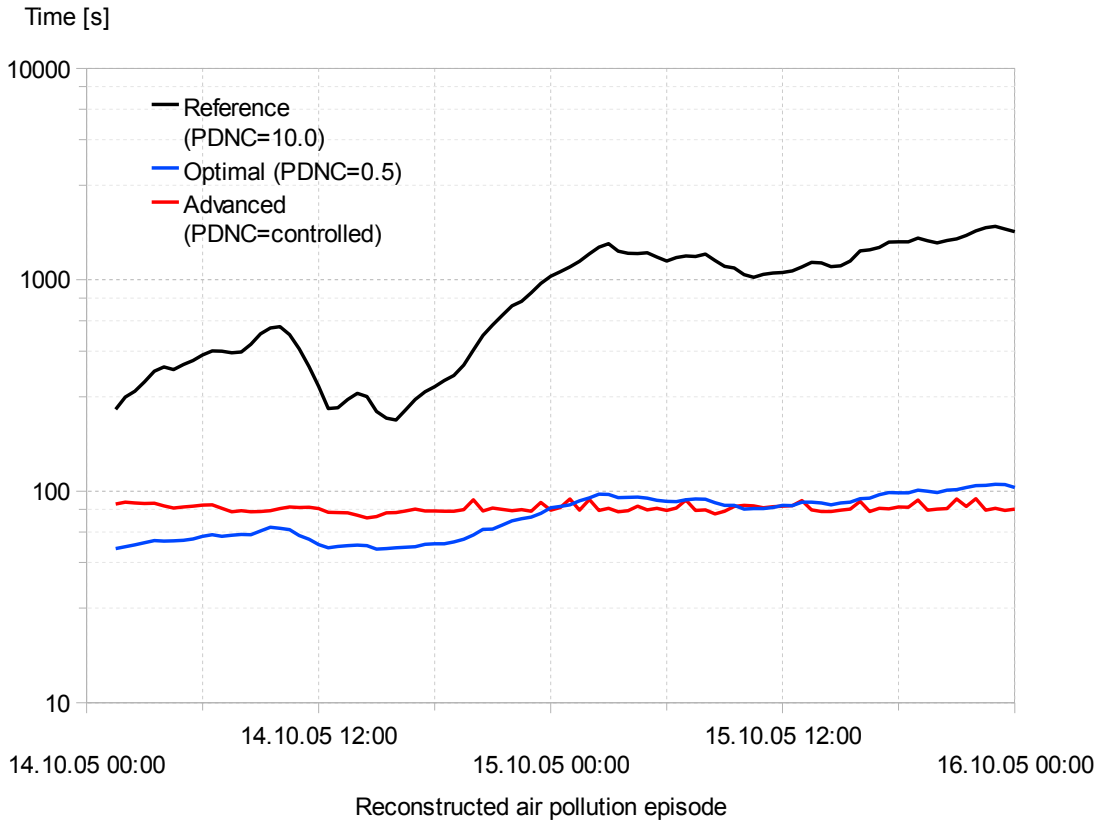


Figure 126: Comparison of the time used for the reconstructions of air-pollution episodes

The results of the time used for each simulation are presented in Figure 126. The figure shows that the time used in the first and second simulations strongly depends on the current meteorological conditions, where low-wind meteorological conditions occurred in the middle of the selected time period and lasted until the end of the simulation. While in the third controlled case, where the advanced AP computer model is used, the time used for each air-pollution episode reconstruction was practically constant during the whole simulation.

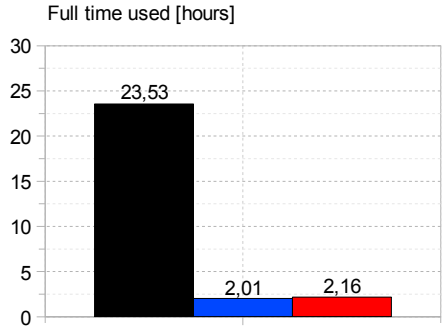


Figure 127: Comparison of full time used

A comparison of the full time used for each simulation run is presented in Figure 127. The results show that the full time used within the advanced simulation run is decreased by one order of magnitude compared to the reference simulation and slightly higher compared to the optimal simulation run.

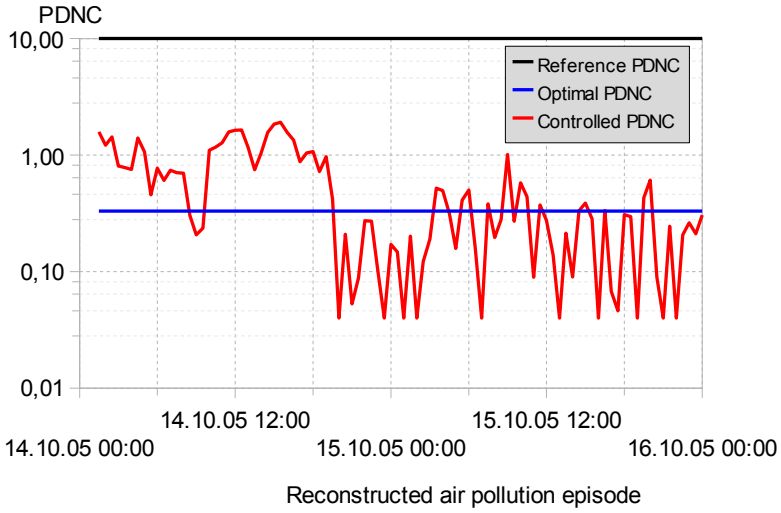


Figure 128: Comparison of the PDNC during the evaluation simulation

Figure 128 shows the *PDNC* of each air-pollution episode reconstruction for all simulations. In the case of the reference simulation and the optimal simulation it is set to a constant value, while in the advanced simulation it is determined by the control module. At the beginning of the advanced simulation run relatively high values of the *PDNC* are used due to the strong wind conditions, but when calm wind conditions occur, relatively small values of the *PDNC* are used to maintain the number of particles in the domain around the controlled value.

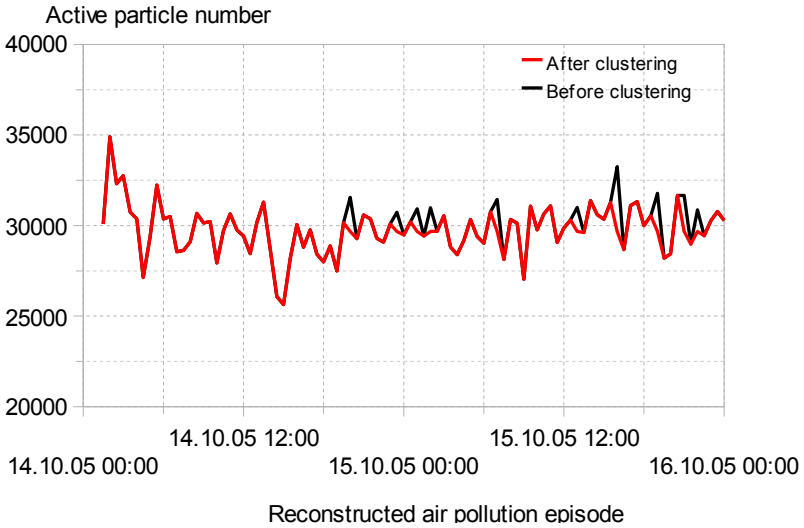


Figure 129: Comparison of number of active particles before and after clustering

In Figure 129 the number of particles at the end of each air-pollution episode reconstruction and the number of particles after clustering are presented. Where the values differ the clustering is activated. The result shows that clustering is activated only in a few situations and when it is activated the number of reduced particles was relatively small, according to the number of all the particles.

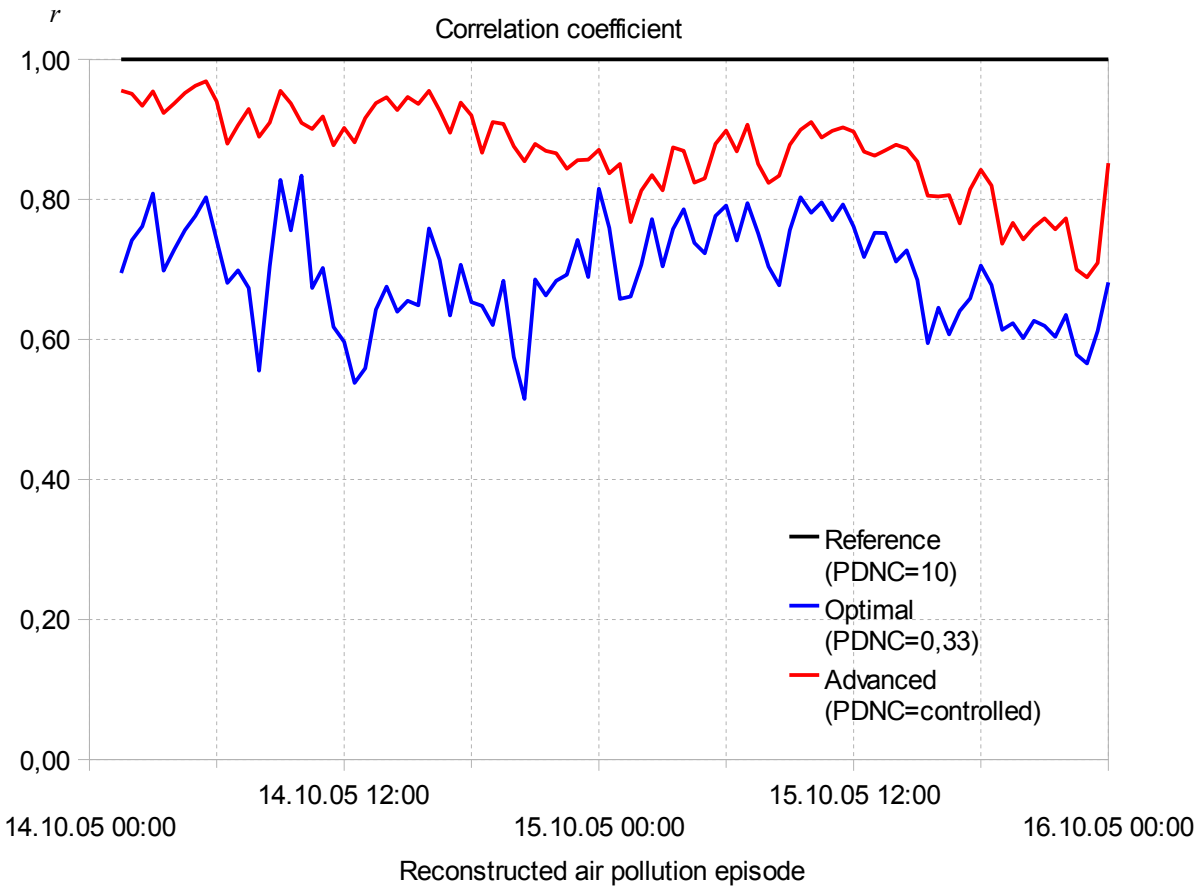


Figure 130: Correlation coefficient comparison between the results of the reference simulation and the optimal and advanced simulation results

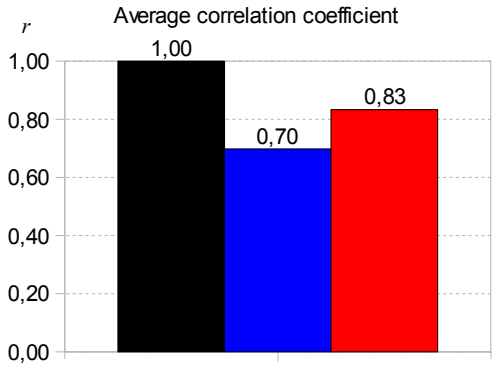


Figure 131: Average correlation coefficient comparison

In Figure 130 a correlation coefficient between each simulation and the reference simulation results are presented. The results of the comparisons show a significant improvement in the correlation coefficient when the ELPD computer model is used. The same conclusion is obtained in Figure 131, where the average correlation coefficients for all the simulation results are presented.

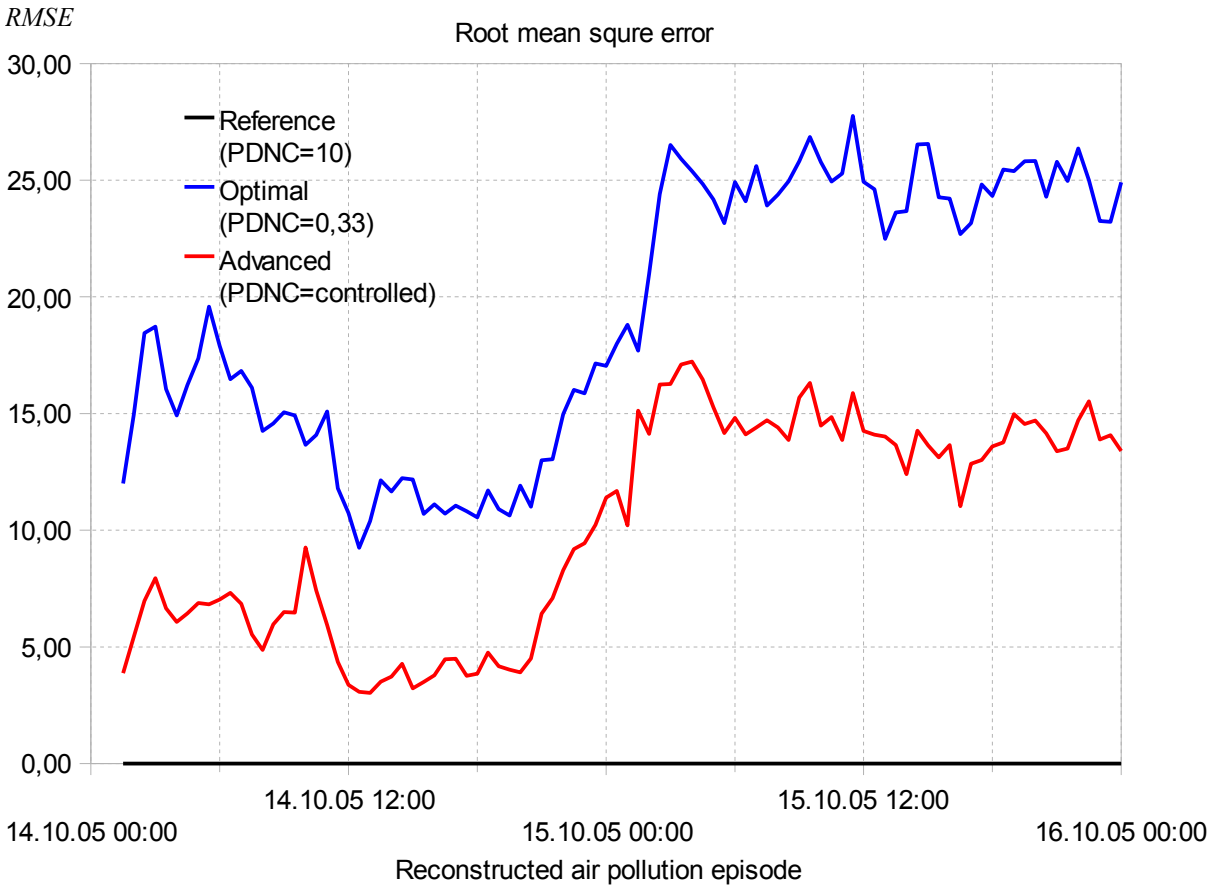


Figure 132: Root mean square error comparison between the results of the reference simulation and the optimal and advanced simulation results

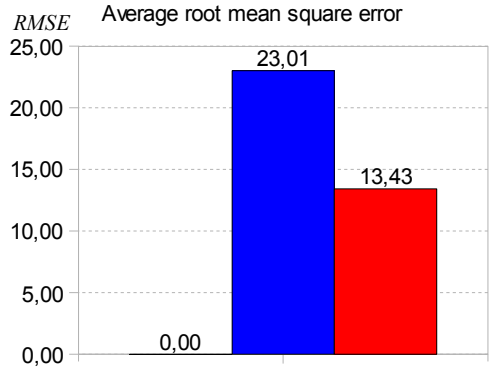


Figure 133: Average root mean square error comparison

In Figure 132 a root mean square error between each simulation and the reference simulation are presented. The results of the comparisons show a significant decrease in the root mean square error when the new ELPD computer model is used. The same results are obtained in Figure 133, where the average root mean square error of the controlled simulation results is reduced by almost 50% compared to the optimal simulation results.

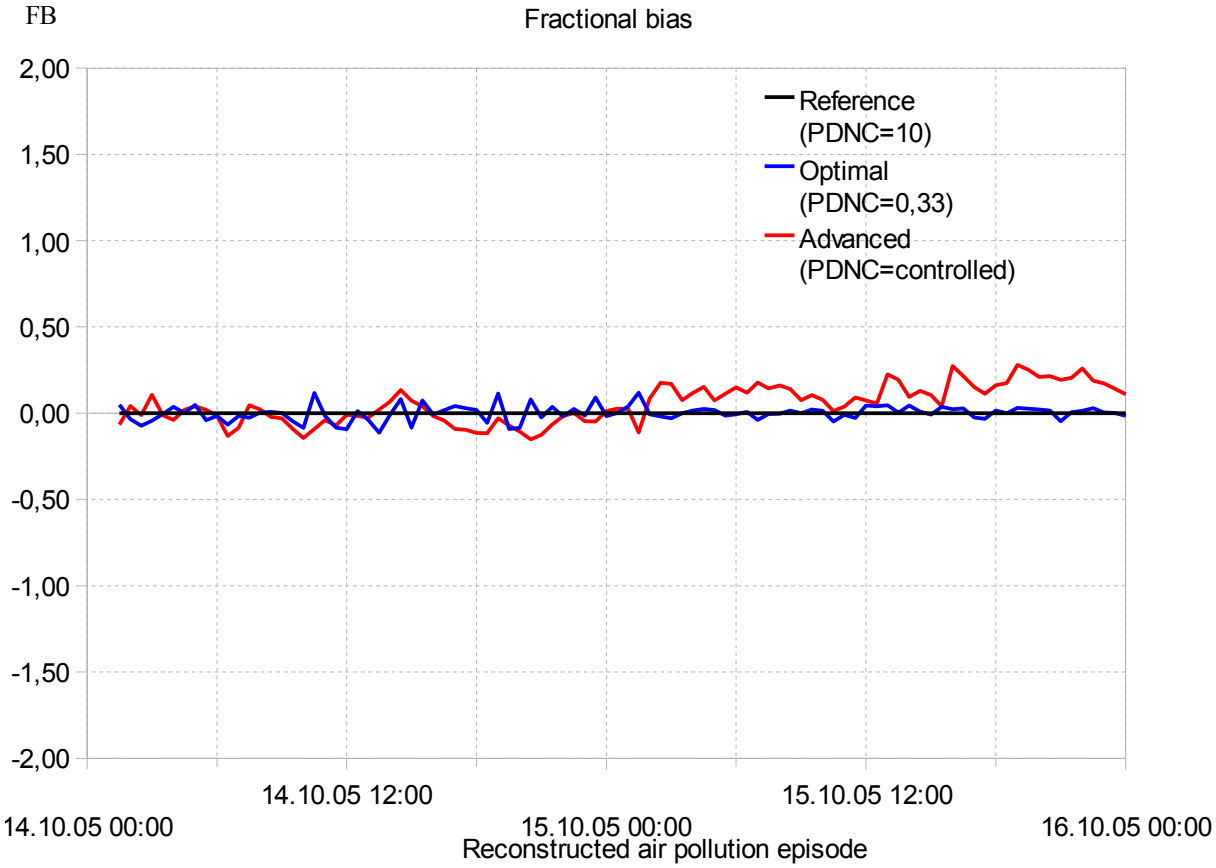


Figure 134: Fractional bias comparison between the results of the reference simulation and the optimal and advanced simulation results

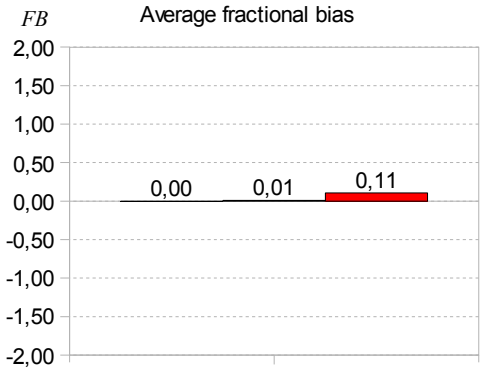


Figure 135: Average fractional bias comparison

According to Figure 135 the fractional bias of the advanced simulation results can be practically neglected for all the results. The comparison of the fractional bias presented in Figure 134 shows that the results of the simulation with the new ELPD computer model are very slightly underestimated in the second part of the simulation, where the limitation process occurred.

6.5. Discussion

The validation of an advanced AP computer model based on an enhanced LPD is contributed in this section to confirm that the proposed new methods in this dissertation can be generally applied in various complex terrain systems. The performance of the new, advanced, enhanced LPD computer model with integrated new modules is evaluated on the Zasavje region field data set air-pollution situation. Before the evaluation is performed the parameters of the *particle number prediction module* and the *density kernel concentration estimation module* are determined to achieve the best possible results. It is recommended to determine the parameters for each complex terrain separately, because the domain size and shape can be very different.

When all the parameters of the enhanced LPD computer model are determined, three simulation runs for the selected situation are performed and compared to each other: the reference simulation by the original LPD computer model, where the maximum acceptable *PDNC* is used, the optimal simulation by the original LPD computer model, where the optimal *PDNC* is used, and the advanced simulation by the new ELPD computer model, where the *PDNC* is controlled.

The results of the comparisons prove that the methods proposed can be generally applied in various complex terrain systems. It is shown that the time used and the active particle number in the reference and optimal simulations is strongly dependent on the current meteorological conditions, while in the advanced simulation it is practically constant. A significant improvement of the correlation coefficient and the root mean square error is achieved by using the ELPD computer model, while the fractional bias comparison showed that there is practically no underestimations in the results of the advanced simulation.

The obtained results show that the use of an advanced air-pollution modeling methodology is recommended, not only in situations where computational resources are constrained, but also in general to optimally take advantage of the available computational resources. The computational complexity of all the air-pollution episode reconstructions is being balanced in the advanced air-pollution modeling methodology based on the ELPD model because approximately the same computational time is spent for each air-pollution episode reconstruction.

7. CONCLUSIONS AND RECOMMENDATIONS

The topic of this dissertation is an air-pollution modelling methodology based on the Lagrangian particle dispersion (LPD). The limit capacities, the properties and the performance of the air-pollution (AP) computer model are determined and evaluated on a complex terrain. The used AP computer model is designed, based on the presented air-pollution modelling methodology. Several methods are suggested to improve the computational performance, based on the results of the evaluation of the AP computer model.

The new methods are developed in a manner such that the original methods in the air-pollution modelling methodology are not modified at all. The parameters, methods and the structure of the original AP model are preserved in their original form and no adjustments are made to the well developed air-pollution model based on LPD. The methods that determine and optimize the reconstruction of the computationally expensive air-pollution situations are proposed and integrated into the existing air-pollution modelling methodology. The computational performance is improved in such a manner that the available computational capabilities are optimally exploited.

The dissertation begins with an introduction, where the research problems, goals and working hypotheses are given. To evaluate achieved results and to confirm the given hypotheses several evaluation methods are proposed in section 3.

In section 3 the sixth hypothesis is confirmed, saying that the quality of the results depends on the number of particles used in the air-pollution episode reconstruction. The quality of the results does not change significantly when the number of particles remains above a certain particle number threshold. This particle number threshold depends on the size of the domain and the size of the cells in the domain. When the number of particles falls below the threshold the quality of the results starts to decrease drastically.

The dependency of the performance of the AP computer model according to the particle number density coefficient (*PDNC*) is determined by making an experimental simulation where several simulation runs are performed on Šaleška region field data set. The presented results show that a very weak dependence of time used on *PDNC* is observed when a small number of particles are used for the reconstruction, but when the number of particles significantly increases, a strong linear dependency can be determined. This means that the time used for a simulation run linearly depends only on the number of particles.

To define the minimum acceptable *PDNC* that is necessary to achieve a good air-pollution reconstruction the evaluation of each result with a different number of used particles is evaluated with the developed evaluation methods. The results of the c*orrelation coefficient* comparison show an exponential increase to some particular number of particles used in the simulation. When this point is reached, the increase in the number of particles has almost no influence on the quality of the results. So, in practice, the use of too large a number of particles yields an unnecessary consumption of computational power and a wasting of computational time. It is finally concluded that in practice the results that correlate with the original result above a factor of 0.8 are completely acceptable. The results of the *root mean*

square error show an exponential decrease to some constant value of 0. Again, it is estimated that the results do not improve significantly after a certain threshold value is reached. The results of the *fractional bias* comparison are almost completely free of bias according to the definition. From the value of the fractional bias it cannot be exactly estimated when the results are acceptable in practice, because there does not exist a certain value when the results are not any more acceptable. But the fractional bias comparison can be used to approximately estimate whether the result is in accordance with the original or is it completely under or over estimated.

The first presented method is clustering, which contributed to a decrease in the computational cost by decreasing the number of active particles in the simulations. The presentation of the clustering method is concluded with the recommendations about the optimal setting of the four basic parameters used to achieve satisfactory results: N_{sub}, N_{size}, m_{max} and N_{max}. According to the finally acquired results it is concluded in section 4 that the hierarchical clustering method with additional parameters can be used in practice only for the limitation of a very large number of particles, when the number of particles exceeds normal values when extreme situations occur, like failure of the desulphurization plant, when emissions increase by an order of magnitude, or when very stable meteorological situations occur, where low winds are present and the air pollution starts to accumulate in the domain. A strong limitation on the number of particles in the reconstructions during typical situations with the clustering method is not recommended, because the quality of the results becomes very poor. To preserve the good quality of the results only a slight limitation is recommended. A comparison of the original results and the results obtained with the clustering method show that the results obtained with the clustering method can become bubbled and less smooth than the original, which is the same effect that occurs when not enough particles are used in the reconstructions.

In section 4 the first hypothesis that computational expenses would be reduced with the clustering method is confirmed, because the used computational complexity can be reduced by at least 50%.

The second hypothesis that the use of the clustering method will have a minor influence on the quality of the results is only partially confirmed. The clustering method has a minor influence on the quality of results only when the final number of particles after clustering in the domain remains above a certain particle number threshold.

The cell concentration estimation method based on the kernel density is proposed in section 4 to substitute the box counting concentration estimation method and to improve the poor quality of the results when a smaller number of particles is used in the simulations. To evaluate the performance of the contributed method and the dependence on the input parameters, several simulations are performed and the results are compared to the reference ground concentration field. Several experiments are made to determine the optimal input parameters σ_x, σ_y and σ_z for the Šaleška region field data set. In the performed experiments three cases of optimal inputs into the density kernel concentration application are determined and presented: over-smoothed, optimal and under-smoothed. After the three different optimal

input values for the different correlation coefficients are determined, for each optimal point one simulation run is performed with the improved AP computer model. The comparisons of the three different optimal points show that the final optimal point is set from the case where the correlation with the original result is very good. The other two obtained optimal points gave some slightly better or worse results.

The final comparisons prove that the correlation coefficient and the root mean square error significantly improved. With this final comparison the eight hypothesis that the poor quality of the simulation results in the situations when a relatively small number of particles are used can be improved by using the kernel density concentration estimation method is confirmed.

The presented final comparison also confirms the ninth hypothesis, that the kernel density concentration estimation method can always be used to improve the quality of the results.

To control the *PDNC* parameter and clustering parameter N_{max} of the LPD computer model a third method is proposed in this thesis. The rule for the *PDNC* control is the following: when a high emission occurs the *PDNC* should be decreased, and when a low emission occurs the *PDNC* can be increased. In special situations when the smallest possible *PDNC* is used, and it is still expected that the computational resources will be exceeded, the clustering must be activated to reduce the number of particles from previous air pollution. This is very important for situations when extreme air pollution is expected. Such a common situation occurs in calm meteorological conditions when the air pollution starts to accumulate in the domain for a longer time interval.

The method consists of two main subsequent methods. In the first step the percentage of lost particles is predicted with the use of an artificial neural network, based on the the meteorology, the emission and the situation of the air pollution at the end of previous episode reconstruction. In the second step the clustering parameters are determined by a decision-making method.

The third and the fourth hypotheses are confirmed in section 5 and section 6. The third hypothesis is saying that the algorithm to determine the situations when the particle number should be reduced is based on a black-box modelling technique. And the fourth hypothesis is saying that the algorithm to determine the situations when the particle number should be reduced is efficient and reliable.

In section 5 the seventh hypothesis is confirmed: controlling the number of particles in the simulation is actually preserving the quality of results at a constant level. The results of comparisons show that the time used and the active particle number in the reference and optimal simulation runs strongly depends on the current meteorological conditions and the emission rate, while in the advanced simulation run they are practically constant. A significant improvement of the correlation coefficient and the root mean square error is achieved by using the ELPD computer model, while the fractional bias comparison showed that the results are not over or under estimated. A final comparison of the results of the verification show that time used for the advanced simulation run is reduced by 30% compared to the optimal simulation run, where also the correlation coefficient of the results is increased by 15%

compared to the optimal simulation results.

The performance of the new, advanced AP computer model is validated on the Zasavje region field data set in section 6. The validation is performed to confirm that the proposed methods can be generally applied in various complex terrains. When all the parameters of the ELPD computer model are determined according to field data set properties, three simulations for the selected situation are performed and compared to each other.

The results of the comparisons show that the time used and the active particle number in the reference and optimal simulations strongly depend on the current meteorological conditions, while in the advanced simulation it is practically constant. Again, a significant improvement of correlation coefficient and the root mean square error is achieved by using the improved AP computer model, while the fractional bias comparison show that the results are only very slightly underestimated. The obtained results show that the use of the improved AP computer model is recommended, not only in situations where the computational resources are constrained, but also in general to optimally take advantage of the available computational resources.

The seventh hypothesis is again confirmed in section 6: controlling the number of particles in the simulation preserves the quality of the results at a constant level. The final comparison of the results of the validation show that the time used for the advanced simulation run is approximately the same as for the optimal simulation, but the correlation coefficient of the advanced simulation results is increased by 15% ,and root mean square error is decreased by 50%, compared to the optimal simulation results. The presented results do not include the computational improvements proposed by Schwere et al.,[10] where the computational complexity could be reduced by an additional 50%.

In this dissertation the following original contributions are proposed:

- *The application of the clustering method for a reduction of the computational cost* is contributed to a decrease of the computational cost by decreasing the number of particles in the simulations. The clustering method is proposed with recommendations about the optimal setting of the four basic parameters: N_{sub}, N_{size}, m_{max} and N_{max}. It is recommended that the hierarchical clustering method with additional parameters can be used in practice only for the limitation of a very large number of particles. The strong limitation of the number of particles in reconstructions during typical situations with the clustering method is not recommended ,because the quality of results becomes poorer.
- *The cell concentration kernel density estimation method adaption* is contributed to substitute the box counting concentration estimation method and to improving the poor quality of the results when a smaller number of particles is used in the simulations. The final comparisons of the results prove that the correlation coefficient and the root mean square error can be significantly improved, while the fractional bias comparison showed that there is practically no underestimations in the results of the advanced simulation, which is crucial for long-term evaluations of the air pollution..
- *The Lagrangian particle dispersion control method based on artificial neural networks,* where two parameters are controlled: *PDNC* and N_{max}. The control rule is the following: when a high emission occurs the *PDNC* should be decreased and when a low emission occurs the *PDNC* can be increased. In special situations, when the smallest possible *PDNC* is used and it is still expected that computational resources will be exceeded, the clustering is activated to reduce the number of particles from the previous air pollution by the parameter N_{max}. The method consists of two main subsequent methods: the prediction of the percentage of lost particles used by the artificial neural network and a decision making method.
- *The integration of the contributed improvements into the advanced, enhanced Lagrangian particle dispersion model*, where the mutual use of the contributed methods is proposed to obtain the best possible results within the given computational resources. The contributed integration is realised in ELPD computer model, which is validated on the Šaleška region and the Zasavje region field data sets. The obtained results proved that the use of an advanced air-pollution modelling methodology is recommended, not only in situations where the computational resources are constrained, but also in general to optimally take advantage of the available computational resources. The computational complexity of all the air pollution episode reconstructions is being balanced because approximately the same computational time is spent for each air-pollution episode reconstruction.

8. REFERENCES

[1] The Council of the european communities 1984. Council Directive 84/360/EEC of 28 June 1984 on the combating of air pollution from industrial plants. Official Journal L 188 , 16/07/1984: 20-25

[2] Government of the Republic of Slovenia 1994. Decree on the emission of substances into the atmosphere from stationary sources of pollution. Official Journal RS, No. 73/1994: 4173-4180

[3] Božnar M. Z., Mlakar P, Grašič B., Popović D., Tinarelli G. 2007. Modelling of air pollution dispersion from blocks 5 and 6 of thermal power plant Šoštanj using Lagrangian particle model, Results of annual statistical elaboration, Elaborated period: July 2006 – June 2007, Progress report. MEIS environmental consultion d.o.o., Slovenia, MEIS environmental consultion d.o.o., Slovenia

[4] Božnar M. Z., Mlakar P, Grašič B., Popović D., Tinarelli G. 2008. Elaborat o rezultatih modeliranja vpliva blokov 5 in 6 TE Šoštanj na obmejno območje Avstrije, Obdobje meteoroloških podatkov: januar–december 2005 (in Slovene). MEIS environmental consultion d.o.o., Slovenia, MEIS environmental consultion d.o.o., Slovenia

[5] Government of the Republic of Slovenia 2006. Environmental Protection Act (official consolidated text). Official Journal RS, No. 39/2006: 4151-4189

[6] Breznik B., Božnar M, Mlakar P., Tinarelli G. 2004. Dose projection using dispersion models. International Journal of Environment and Pollution, Vol. 20: 278-285

[7] Bosanquet, C.H. and Pearson, J.L. 1936. The spread of smoke and gases from chimney. Transactions of the Faraday Society Vol. 32: 1249-1263

[8] Sutton O. G. 1947. The problem of diffusion in the lower atmosphere. Quarterly Journal of the Royal Meteorological Society Vol. 73: 257-263

[9] Wilson J. D., Sawford B. L. 1996. Review of Lagrangian stochastic models for trajectories in the turbulent atmosphere. Boundary-Layer Meteorology Vol. 78: 191-210

[10] Schwere S., Stohl A., Rotach M. W. 2002. Practical considerations to speed up Lagrangian stochastic particle models. Computers & Geosciences Vol. 28: 143-154

[11] Tinarelli G., Anfossi D., Bider M., Ferrero E., Castelli S. T. 2000. A New High Performance Version of the Lagrangian Particle Dispersion Model SPRAY, Some Case Studies. NATO challenges of modern society Vol. 23A: 499-508

[12] Graff A. 2002. The new German regulatory model – a Lagrangian particle dispersion model. 8th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, October 14-17, Sofia, Bulgaria: 153-158

[13] Mikkelsen T., Desiato F. 1993. Atmospheric Dispersion Models and Pre-processing of Meteorological Data for Real-time Application. Proceedings of the Third International Workshop on Real-time Computing of the Environmental Consequences of an Accidental Release to the Atmosphere from a Nuclear Installation, Schloss Elmau, Bavaria, October 25-30 1992. Journal of Radiation Protection Dosimetry Vol 50 Nos 2-4: 205-218

[14] Tinarelli G., Anfossi D., Brusasca G., Ferrero E., Giostra U., Morselli M. G., Moussafir J., Tampieri F., Trombetti F. 1994. Lagrangian particle simulation of tracer dispersion in the lee of a schematic two-dimensional hill. Journal of Applied Meteorology Vol. 33 No. 6: 744-756

[15] Ferrero E., Anfossi D., Brusasca G., Tinarelli G., Alessandrini S., Trini Castelli S. 1996. Simulation of atmospheric dispersion in convective boundary layer: comparison between different Lagrangian particle models. 4th Workshop on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Oostende, 6-9 May 1996, 67-74. International Journal of Environment and Pollution, Vol 8, Nos. 3-6: 315-323

[16] Thykier-Nielsen S., Santabarbara J. M., Mikkelsen T 1993. Dispersion scenarios over complex terrain. Radiation Protection Dosimetry, Vol. 50, Nos 2-5: 249-255

[17] Cox R.M., Sontowski J., Fry R.N., Dougherty C.M., Smith T.J. 1998. Wind and Diffusion Modeling for Complex Terrain. Journal of Applied Meteorology, Vol. 37, Issue 10: 996–1009

[18] Hirtl M., Baumann-Stanzer K., Kaiser A., Petz E., Rau G., 2007. Evaluation Of Three Dispersion Models For The Trbovlje Power Plant, Slovenia. 11th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, July 2007, Cambridge, UK: 21-25

[19] Ferrero E., Anfossi D., Tinarelli G., Trini Castelli S 2001. Lagrangian Particle Simulation of an EPA Wind Tunnel Tracer Experiment in a Schematic Two-Dimensional Valley. Air Pollution Modelling and its Applications XIV (2001), S.E. Gryning and F.A. Schiermeier eds., Kluwer Academic / Plenum Press, New York: 717-718

[20] Anfossi D., Desiato F., Tinareli G., Brusacsa G., Ferrero E., Sacchetti D. 1998. TRANSALP 1989 Experimental Campaign - part II: Simulation of a tracer experiment with Lagrangian particle models. Atmospheric Environment Vol. 32: 1157-1166

[21] Brusasca G., Tinarelli G., Anfossi D., 1992. Particle model simulation of diffusion in low windspeed stable conditions. Atmospheric Environment Vol. 26A: 707-723

[22] Anfossi D., Ferrero E., Brusasca G., Marzorati A., Tinarelli G., 1993. A simple way of computing buoyant plume rise in Lagrangian stochastic dispersion models. Atmospheric Environment Vol. 27A: 1443-1451

[23] Thomson D. J. 1984. Random walk modelling of diffusion in inhomogeneous turbulence. Quarterly Journal of the Royal Meteorological Society, Vol. 110, No. 466: 1107-1120

[24] Thomson D. J. 1987. Criteria for the selection of stochastic models of particle trajectories in turbulent flows. Journal of Fluid Mechanics. Vol. 180: 529-556

[25] Grašič B., Božnar M. Z., Mlakar P. 2007. Re-evaluation of the Lagrangian particle modelling system on an experimental campaign in complex terrain. Il Nuovo Cimento C, Vol. 30, No. 6: 19-

[26] Cushman-Roisin B. 2006. Environmental transport and fate. Hanover, USA, Darthmouth College, Thayer school of engineering

[27] U.S. Environmental Protection Agency. Support Center for Regulatory Atmospheric Modeling, Dispersion Modeling. http://www.epa.gov/scram001/dispersionindex.htm (2006)

[28] ENVIRON International Corporation, 101 Rowland Way, Suite 220 Novato, California 94945-5010. CAMx – Comprehensive Air Qualitiy Model with Extensions, Users' guide, version 4.30. http://www.camx.com/files/CAMxUsersGuide_v4.30.pdf (2006)

[29] Chaxel E., Brulfert G., Chemel C., Chollet J.-P. 2004. Evaluation of local ozone production of Chamonix valley (France) during a regional smog episode. 27th NATO/CCMS International Technical Meeting on Air Pollution Modelling and its Application, 24-29 October 2004, University of Calgary, Alberta, Canada: 49-56

[30] Calvert S., Englund H. M. 1984. Handbook of air pollution technology. New York, USA, John Wiley & Sons, Ltd

[31] Božnar M. 1997. Izbira učnih vzorcev za model napovedovanja onesnaženja ozračja na osnovi nevronskih mrež, Doktorska disertacija (in Slovene). Ljubljana, Univerza v Ljubljani, Fakulteta za elektrotehniko

[32] Mlakar P. 1997. Izbira značilk vzorcev za model kratkoročnega napovedovanja onesnaženja ozračja, Doktorska disertacija (in Slovene). Ljubljana, Univerza v Ljubljani, Fakulteta za elektrotehniko

[33] Schulze R. H. 1991. Practical guide to atmospheric dispersion modelling. Dallas, Texas, USA, Trinity Consultants

[34] European Topic Centre on Air and Climate Change. Model Documentation System. http://air-climate.eionet.eu.int/databases/MDS/index_html (2006)

[35] Harmancioglu N. B., Singh V. P., Alpaslan M. N. 1998. Environmental Data Management, Water Science and Techology Library, Vol 27. Netherlands, Kluwer Academic Publishers

[36] Blumen W., Banta R. M., Berri G., Carruthers D. J., Dalu G. A., Durran D. R., Egger J., Garratt J. R., Hanna S. R., Hunt J. C. R., Meroney R. N., Miller W., Neff W. D., Nicolini M., Paegle J., Pielke R. A., Smith R. B., Strimaitis D. G., Vukicevic T., Whiteman C. D. 1990. Atmospheric processes over complex terrain, Meteorological monographs, Volume 23, Number 45. Boston, USA, American Meteorological Society

[37] Božnar M., Brusasca G., Cavicchioli C., Faggian P., Finardi S., Mlakar M., Morselli M. G., Sozzi R., Tinarelli G. 1994. Application of advanced and tradicional diffusion models to an experimental campaign in complex terrain. 2nd International Conference on Air Pollution, Barcelona, 1994, Air Pollution II, Vol. 1, Computer Simulation, editors J. M. Baldsano et al., Southampton, Boston, Computational Mechanics Publ.: 159-166

[38] Božnar M., Brusasca G., Cavicchioli C., Faggian P., Finardi S., Minella M., Mlakar M., Morselli M. G., Sozzi R. 1994. Model Evaluation and Application of Advanced and Traditional Gaussian Models on the Experimental Šoštanj (Slovenia, 1991) Campaign. Workshop on Intercomparison of Advanced Practial Short-Range Atmospheric Dispersion Model, Manno, 1993, Proceedings of the Workshop, editor C. Cuvelier, European Commission: 112-121

[39] U.S. Environmental Protection Agency 2000. Meteorological Monitoring Guidance for Regulatory Modeling Applications, Document EPA-454/R-99-005. USA, Office of Air Quality Planning and Standards, Research Triangle Park, NC

[40] de Baas A. F., van Dop H., Nieuwstadt F. T. 1986. An application of the Langevin equation for inhomogeneours conditions to dispersion in a convection boundary layer. Quarterly Journal of the Royal Meteorological Society Vol. 112: 165-180

[41] Sawford B. L. 1984. The basis for, and some limitations, of the Langevin equation in atmospheric relative dispersion modelling. Atmospheric Environment Vol. 11: 2405-2411

[42] Siljamo P., Sofiev M., Ranta H. 2004. An approach to simulation of long-range atmospheric transport of natural allergens: an example of birch pollen. 27th NATO/CCMS International Technical Meeting on Air Pollution Modelling and its Application, 24-29 October 2004, University of Calgary, Alberta, Canada: 395-402

[43] Sjoberg J., Zhang Q., Ljung L., Benveniste A., Delyon B., Glorennec P.-Y., Hjalmarsson H., Juditsky A. 1995. Nonlinear black-box modeling in system identification: a unified overview. Automatica Vol. 31(12): 1691-1724

[44] Box George E. P., Jenkins Gwilym M., 1976. Time series analysis forecasting and control. San Francisco, USA, Holden-Day

[45] Lütkepohl H. 1991. Introduction to Multiple Time Series Analysis. Berlin, Germany, Springer-Verlag

[46] Juditsky A., Hjalmarsson H., Benveniste A, Delyon B., Ljung L., Sjöberg J., Zhang Q. 1995. Nonlinear black-box models in system identification: Mathematical foundations. Automatica Vol. 31(12): 1725 – 1750

[47] Hornik K., Stinchcombe M., White H. 1989. Multilayer feedforward networks are universal approximators. Neural Networks Vol. 2, Issue 5: 359 – 366

[48] Gardner, M.W., Dorling, S.R. 1998. Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. Atmospheric Environment Vol. 32: 2627-2636

[49] Grašič B., Mlakar P., Božnar M. Z. 2006. Ozone prediction based on neural networks and Gaussian processes. Il Nuovo Cimento C, Vol. 29, Issue 6: 651-661

[50] Mlakar P., Božnar M. 1996. Analysis of winds and SO2 concentrations in complex terrain. Air pollution IV : monitoring, simulation and control. Southampton; Boston: Computational Mechanics Publications: 455-464

[51] Božnar M., Lesjak M., Mlakar P. 1993. A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. Atmospheric Environment Vol. 27 B: 221-230

[52] Kocijan J. 2007. Modeliranje dinamičnih sistemov z umetnimi nevronskimi mrežami in sorodnimi metodami (in Slovene). Nova Gorica, Slovenia, Univerza v Novi Gorici

[53] Agarwal M. 1997. A systematic classification of neural-network-based control. IEEE Control Systems Magazine Vol. 17, No. 2: 75-93

[54] Nørgaard M, Ravn O., Poulsen N. K., Hansen L. K. 2000. Neural Networks for Modelling and Control of Dynamic Systems: A Practitioner's Handbook. London, GB, Springer

[55] Engelbrecht A. P. 2002. Computational Intelligence: An Introduction. Chichester, England, John Wiley & Sons, Ltd

[56] Rojas R. 1996. Neural Networks: A Systematic Introduction. Berlin, Germany, Springer-Verlag

[57] Mlakar P., Božnar M. 1997. Feature determination for air pollution forecasting models. International Conference on Air Pollution – Proceedings: 577-586

[58] Abdul-Wahab S. A., Al-Alawi S.M. 2002. Assessment and Prediction of Tropospheric Ozone Concentration Levels using Artificial Neural Netwoks. Environmental Modelling & Software Vol. 17(3): 219-228

[59] Ballester E. B., Valls G. C., Carrasco-Rodriguez J. L., Olivas E. S., S. del Valle-Tascon, 2002. Effective 1-day ahead prediction of hourly surface ozone concentrations in eastern Spain using linear models and neural networks. Ecological Modelling Vol. 156: 27-41

[60] Comrie A. C. 1997. Comparing neural networks and regression models for ozone forcasting. Journal of Air & Wast Management Association Vol. 47: 653-663

[61] Finzi G., Nunnari G. 2004. Air quality forecast and alarm systems. Air Quality Modelling-Theories, Methodologies, Computational Techniques and Available Databases and Software, Vol. II, Ch. 16, EnviroComp Institute and Air & Waste Management Association copublishers: 653-663

[62] Gardner, M.W., Dorling, S.R. 1996. Neural network modelling of the influence of local meteorology on surface ozone concentrations. Proceedings 1st International Conference on GeoComputation, University of Leeds: 359-370

[63] Yi J., Prybutok R. 1996. A neural network model forcasting for prediction of daily maximum ozone concentration in an industrialized urban area. A neural network model forcasting for prediction of daily maximum ozone concentration in an industrialized urban area: 349-357

[64] The MathWorks Inc. 2002. Neural Network Toolbox, Matlab version 6.5 documentation. Natick, Massachusetts, USA,

[65] Lawrence J. 1991. Introduction to Neural Networks. Grass Valley, USA, California Scientific Software

[66] Ward Systems Group, Inc. 1993. NeuroShell 2. Frederick, Maryland, USA, User's manual

[67] Rumelhart D. E., Hinton G. E., Williams R. J. 1989. Learning internal representations by error propagation, Parallel distributed processing: Explorations in the Microstructure of Cognition, Vol. 1. Cambridge, MA, MIT Press

[68] Hagan, M.T., Demuth H.B., Beale M.H. 1996. Neural Network Design. Boston, USA, PWS Publishing

[69] Marquardt D. 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters.

SIAM Journal on Applied Mathematics Vol. 11: 431–441

[70] Hagan M.T., and M. Menhaj 1994. Training feed-forward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks, Vol. 5, No. 6: 989-993

[71] Jain A. K., Murty M. N., Flynn P. J. 1999. Data Clustering: A review. ACM Computing Surveys, Vol. 31, No. 3: 264-323

[72] Jain A. K., Dubes R. C. 1988. Algorithms for Clustering Data. Upper Saddle River, NJ.,

[73] MacQueen J. B. 1967. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press: 281-297

[74] Kohonen T. 1989. Self-organization and associative memory 3rd edition. Berlin, Germany, Springer-Verlag

[75] Garten, J. F., Schemm, C. E., & Croucher, A. R. 2003. Modeling the transport and dispersion of airborne contaminants: A review of techniques and approaches. Johns Hopkins APL Technical Digest (Applied Physics Laboratory) Vol. 24(4): 368-375

[76] Holmes, N. S., & Morawska, L. 2006. A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. Atmospheric Environment Vol. 40: 5902-5928

[77] Vardoulakis, S., Fisher, B. E. A., Pericleous, K., & Gonzalez-Flesca, N. 2003. Modelling air quality in street canyons: A review. Atmospheric Environment Vol. 37: 155-182

[78] Rizza U., Mangia C., Tirabassi T. 1996. Validation of an operational advanced Gaussian model with Copenhagen and Kincaid datasets. 4th Workshop on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Oostende, 6-9 May 1996, 67-74. International Journal of Environment and Pollution, Vol 8, Nos. 3-6: 41-48

[79] Kaasik M. 1988. Validation of the AEROPOL model against the Kincaid data set. 10th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, October 17-20 2005, Sissi, Crete: 327-331

[80] Ferrero E., Anfossi D., Brusasca G., Tinarelli G. 1995. Lagrangian Particle Model: Evaluation against Tracer Data. Int. J. Environment and Pollution: 360-374

[81] Olesen H.R 1996. Toward the establishment of a common framework for model evaluation. Air Pollution Modeling and Its Application XI, Edited by S-E. Gryning and F. Schiermeier, Plenum Press, New York: 519-528

[82] Elisei G., Bistacchi S., Bocchiola G., Brusasca G., Marcacci P., Marzorati A., Morselli M. G., Tinarelli G., Catenacci G., Corio V., Daino G., Era A., Finardi S., Foggi G., Negri A., Piazza G., Villa R., Lesjak M., Božnar M., Mlakar P., Slavic F. 1992. Experimental Campaign for the Environmental Impact Evaluation of Šoštanj Thermal Power Plant, Progress report. CISE Technologie innovative, Milano, Italia, ENEL S.p.A, CRAM-Servizio Ambiente, Milano, Italy, C.I.S.E. Tecnologie Innovative S.p.A, Milano, Italy, Institute Jozef Stefan, Ljubljana

[83] Souto, M. J., Souto, J. A., Peěrez-Munuzuri, V., Casares, J. J., Bermuědez, J. L. 2006. A comparison of operational lagrangian particle and adaptive puff models for plume dispersion forecasting. Atmospheric Environment Vol. 35: 2349-2360

[84] Brusasca, G., Carboni, G., Finardi, S., Sanavio, D., Tinarelli, G. and Toppetti, A. 2001. Comparison of a Gaussian (ISC3) and a Lagrangian Particle Model (SPRAY) for regulatory application in flat and complex terrain sites representative of typical Italian landscape. Proceedings of the 7th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Belgirate, Italy:

[85] Hurley P., Physick W. 1993. A skewed homogeneous Lagrangian particle model for convective conditions. Atmospheric Environment Vol. 27 A: 619-624

[86] Ryall D. B., Maryon R. H. 1998. Validation of the UK Met. Office's NAME model against

ETEX dataset. Atmospheric Environment Vol. 24: 4265-4276

[87] Ermak D. L., Nasstrom J. S., 2000. A Lagrangian stochastic diffusion method inhomogeneous turbulence. Atmospheric Environment Vol. 34: 1059-1068

[88] Melheim J. A. 2005. Cluster integration method in Lagrangian particle dynamics. Computer Physics Communications Vol. 171: 155-161

[89] Krasnopolsky V. M., Chevallier F. 2003. Some neural network applications in environmental sciences. Part II: advancing computational efficiency of environmental numerical models. Neural Networks Vol. 16: 335-348

[90] Krasnopolsky V. M., Chevallier F. 2005. New approach to Calculation of Atmospheric Model Physics: Accurate and Fast Neural Network Emulation of Longwave Radiation in a Climate model. Monthly Weather Review, Vol. 133, American Meteorological Society: 1370-1383

[91] Wikipedia, the free encyclopedia. Šaleška valley. http://sl.wikipedia.org/wiki/%C5%A0ale %C5%A1ka_dolina (2006)

[92] Campaign home page. Experimental Campaign for the Environmental Impact Evaluation of Šoštanj Thermal Power Plant. http://193.77.212.133/tes-campaign91/indexe.html (2007)

[93] Wikipedia, the free encyclopedia. Zasavje. http://sl.wikipedia.org/wiki/Zasavje (2006)

[94] Božnar M., Grašič B., Tinnareli G. 2006. Thermal power plant Trbovlje air pollution impact modelling in complex terrain. V: Sixth Annual Meeting of the European Meteorological Society (EMS)[and] Sixth European Conference on Applied Climatology (ECAC) : Ljubljana, Slovenia, 4-8 September 2006, (EMS annual meeting abstracts, volume 3). Ljubljana: European Meteorological Society: Agencija RS za okolje:

[95] Baumann-Stanzer K., Kaiser A., Rau G., Petz E., Hirtl M. 2006. Europe's highest stack in complex terrain – evaluation of dispersion modelling for the Trbovlje power plant, Slovenia. V: Sixth Annual Meeting of the European Meteorological Society (EMS)[and] Sixth European Conference on Applied Climatography (ECAC) : Ljubljana, Slovenia, 4-8 September 2006, (EMS annual meeting abstracts, volume 3). Ljubljana: European Meteorological Society: Agencija RS za okolje:

[96] Builtjes P. J. H. 2001. Major Twentieth Century Milestones in Air Pollution Modelling and Its Application . Air Pollution Modelling and its Applications XIV (2001), S.E. Gryning and F.A. Schiermeier eds., Kluwer Academic / Plenum Press, New York: 3-16

[97] Karba R. 1991. Modeliranje procesov (in Slovene). Ljubljana, Slovenia, Univerza v Ljubljani, Fakulteta za elektrotehniko

[98] The Council of the european communities. Council Directive 84/360/EEC of 28 June 1984 on the combating of air pollution from industrial plants. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31984L0360:SL:NOT (1984)

[99] Republika Slovenija. Decree on the emission of substances into the atmosphere from stationary sources of pollution (Ur.l. RS, št. 31/2007). http://zakonodaja.gov.si/rpsi/r06/predpis_URED4056.html (2007)

[100] Finardi S., Brusasca G., Calori G., Nanni A., Tinarelli G., Agnesod G., Pession G., Zublena M 2002. Integrated air quality assessment of an alpine region: evaluation of the Mont Blanc tunnel re-opening effects. 8th Conference on Harmonization within Atmospheric Dispersion Modeling for Regulatory Purposes. Sofia, 14-17 October 2002: 404-408

[101] Anfossi D. 1985. Analysis of plume rise data from five TVA Steam Plants. Journal of Applied Meteorology: Vol. 24, No. 11: 1225-1236

[102] Geai P. 1985. Methode d'Interpolation et de Reconstitution Tridimensionelle d'un Champ de Vent: le Code d'Analyse Objective MINERVE. E.D.F., 6 Quai Waitier, 78400, Chatou, France, Technical Report DER/HE/34-87.03, EDF

[103] Tinarelli G. 2001. SPRAY 3.0 - General Description and User's Guide. AriaNet Srl,

Milano, Italy, User's manual

[104] Paine R.J. 1988. User's guide to the CTDM meteorological preprocessor (METPRO) program. USA, US-EPA report EPA/600/8-88/004

[105] Hanna S. R. 1982. Application in air pollution modelling in "Atmospheric Turbulence and Air Pollution Modelling". Ed. by S.T.M. Nieuwstadt and H.Van Dop, D.Reitell Pub. Comp.: 275-310

[106] Pielke R.A., Cotton W.R., Walko R.L., Tremback C.J., Lyons W.A., Grasso L.D., Nicholls M.E., Moran M.D., Wesley D.A., Lee T.J., Copeland J.H. 1992. A comprehensive meteorological modeling system—RAMS. Meteorology and Atmospheric Physics Vol. 49: 69-91

[107] Irwin J. S. 2000. Statistical Evaluation of Centerline Concentration Estimates by Atmospheric Dispersion Models. International Journal of Environment and Pollution, Vol. 14, Nos. 1-6: 28-38

[108] Gutfreund P., Liu C., Nicholson B., Roberts E. 1983. COMPLEX I and II model performance evaluation in Nevada and New Mexico. Journal of Air Pollution Control Association Vol. 33: 846-871

[109] De Haan, P. 1999. On the use of density kernels for concentration estimations within particle and puff dispersion models. Atmospheric Environment Vol. 33: 2007-2022

[110] Rodgers J. L., Nicewander W. A 1988. Thirteen ways to look at the correlation coefficient. The American Statistician 42: 59-66

[111] Graham D. I., Rana A. Moyeed 2002. How many particles for my Lagrangian simulations?. Powder Technology Vol. 125: 179-186

[112] Trolltech ASA, Norway. Qt – Trolltech. http://trolltech.com/products/qt (2007)

[113] Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo. Open source Clustering software. http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/index.html (2007)

[114] Devijver J. K. 1990. Pattern recongnition: A Statistical Approach. , Englewood Cliffs, Prentice-Hall, Inc.

[115] Free Software Foundation, Inc., 51 Franklin St - Fifth Floor, Boston, MA 02110-1301 USA. GSL - GNU Scientific Library. http://www.gnu.org/software/gsl/ (2007)

[116] Cachin C. 1994. Pedagogical Pattern Selection Strategies. Neural Networks Vol. 7, Issue 1: 175-181

[117] Munro P. W. 1992. Repeat until bored: A pattern selection strategy. Advances in neural information processing systems 4, San Mateo, CA, Morgan Kaufmann Publishers: 1001-1008

[118] Božnar M. 1997. Pattern selection strategies for a neural network - based short term air pollution prediction model. International Conference on Air Pollution – Proceedings Intelligent Information Systems IIS'97, Grand Bahama Island, Bahamas, December 8-10, 1997. Proceedings. Los Alamitos, California [etc.]: IEEE Computer Society: 340-344

[119] Belue L. M., Bauer K. W. Jr. 1995. Determing input features for multilayer perceptrons. Neurocomputing Vol 7, 2: 111-121

ACKNOWLEDGEMENTS

I would like to sincerely thank my research advisor Dr. Primož Mlakar for his advice, guidance and direction in the subject of air pollution dispersion modelling and artificial neural networks.

I would like to address special thanks to my advisers Prof. Dr. Juš Kocijan and Dr. Marija Zlata Božnar who guided me through the all necessary steps to arrive at the final version of this thesis.

I would like to express to them my gratitude for introducing me to this very interesting research topic and for the many helpful dialogues and comments I have received regarding the air pollution dispersion modelling and artificial neural network modelling but mostly for the ideas that occurred during our dialogues.

In addition, I would also like to express my gratitude to Dr. Gianni Tinarelli and Dr. Giuseppe Brusasca from *AriaNet s.r.l.* from Milano, Italy for their support, explanations and adjustments of the air pollution model *AriaIndustrie* and *Lagrangian* particle dispersion model *Spray*, for their helpful discussions, comments and mostly for their ideas on the improvements to the Lagrangian particle dispersion model that occurred during our dialogues.

I would like to thank the AMES d.o.o company and the MEIS d.o.o. company and their research groups for motivating me to get involved into this interesting research field and to finish this dissertation.

I am also deeply grateful to my colleagues from MEIS d.o.o. who helped me prepare and process the field data sets to create model validation kits, to successfully perform all the simulations and to take most of a burden of the administrative work.

I would like to show my gratitude to the Slovenia Research Agency for providing funds to support this research through the Young Researchers Program, Grant No. 3211-05-000552.

Finally, I am deeply indebted to my wife Tatjana and daughter Klavdija for their support and endless patience during my writing, to my parents Jožica and Martin for their faith in me and allowing me to be as ambitious as I wanted.

APPENDIX A – SHORT AIR POLLUTION MODELLING GLOSSARY

air pollution

Air pollution applies to any chemical, physical or biological agent that modifies the natural composition of the atmosphere. It is the contamination of air by the discharge of harmful substances. Air pollution can cause health problems and it can also damage the environment and property.

air pollution dispersion

Air pollution dispersion is a spreading of air pollution in the atmosphere caused by many atmospheric factors including wind direction and wind speed, type of terrain and heating effects.

air pollution episode

Air-pollution situation is split into several equally long episodes of air pollution for evaluation purposes.

air pollution model

Air pollution model is a mathematical model for air pollution dispersion reconstructions. Air pollution model is the final result of the modelling process and is used for computer air pollution model design.

air pollution modelling

Air pollution modelling is defined as an attempt to describe a functional relation between emissions and occurring concentrations in the surrounding. It can give us a relatively complete and consistent description which also includes an analysis of the causes (emissions sources) which lead to measured concentrations.

air pollution modelling methodology

Air pollution modelling methodology describes the methods used in air pollution modelling and its application.

air pollution simulation

Air pollution simulation is used to reconstruct the air-pollution situation over selected area of interest for the selected period of time.

air pollution situation

Air pollution that usually lasts for some defined period of time. It is determined by average concentrations over area of interest.

area of interest

see *domain*

computer air pollution model

When an air pollution model is created a computer air pollution model is designed to be used in the computer simulation.

domain

Defines the region where air pollution occurs. It can be local (up to 30 km), local-to-regional (30-300 km), regional-to-continental (300-3000 km) or global (hemispheric to global scale).

field data set

Field data set consists of all available emission data, meteorological data and topographical data for selected area of interest.

Lagrangian particle dispersion model

Lagrangian particle dispersion model is a three dimensional model designed to simulate the airborne pollutant dispersion.

measuring campaign

The term describes an air-pollution measuring experiment that is performed over selected area of interest for longer time interval. During the measuring campaign measurements of meteorology, emission and air-pollution concentrations in the ambient air are collected at different locations.