# A method for the constrained interpolation of RLE-compressible chemical weather heatmaps

Victor Epitropou[1], Konstantinos Karatzas[1], Jaakko Kukkonen[2] Ari Karppinen[2]

## Abstract

Chemical Weather (CW) and other geospatial environmental information produced by numerical models are often published on-line in the form of heatmaps which have undergone several lossy processing and transformation steps, resulting in a relatively low visualization and recoverable data, compared to the model data which generated them,. In this paper, a method which is fine-tuned to the partial reconstruction of such chemical weather data starting from discrete-level heatmaps is presented, which relies on the augmentation of ordinary interpolation methods through the use of peak-limiting functions and other constraining methods.

## 1. Introduction

In recent years, there has been a noticeable increase in the amount of Chemical Weather (CW) and other geospatial environmental forecasting information published on-line in various forms, with the most important form being visualized pollution (Chemical Weather) forecasts maps in the form of heatmaps [1]. However, despite this availability of forecasts, which require large amount of data to generate, access the high-quality, numerical data behind them, for either research or service-creation purposes, is still relatively restricted.

Methods to recreate numerical data from heatmaps have been proposed and applied in [3] and [7], but they are limited by the lossy process used to create the heatmaps themselves, where a significant part of the initial information ends up being discarded. The data used to produce these heatmaps are commonly the outputs of numerical dispersion models simulating the variation of pollutant concentrations in time and space, such as the SILAM [6] integrated modelling for atmospheric composition, created and managed by the Finnish Meteorological Institute.

In this paper, a method which is fine-tuned to the partial reconstruction of chemical weather data in the form of discrete-level heatmaps is presented. This method relies on the augmentation of ordinary interpolation methods through the use of peak-limiting functions and other constraining methods which ensure that the interpolation's results will adhere to specific context-dependent boundaries and value distributions.

## 2. Materials and methods

In general, chemical weather data as produced by air quality (AQ) models, henceforth referred to forecasts, can be viewed as 2D signals extending into geographical space (longitude-latitude), when referred to a particular time instance. The amplitudes of these signals express the concentration for a given pollutant, usually in $\mu g/m^3$, ppm or ppb units. Forecasts often also have a time and an altitude

[1] Informatics Applications and Systems Group: vepitrop@isag.meng.auth.gr, kkara@eng.auth.gr , Dept. of Mechanical Engineering, Aristotle University, Thessaloniki, Greece

[2] Finnish Meteorological Institute, jaakko.kukkonen@fmi.fi , Ari.Karppinen@fmi.fi, Erik Palménin aukio 1, FI-00560 HELSINKI, Finland

dimension, effectively making them 3D or even 4D signals, but they can be treated as purely 2D signals for the purpose of data analysis, if a specific time instance and altitude are considered. The most typical heatmap type encountered in the domain of Chemical Weather Air Quality Forecasting [4] uses colour-coded coverages to represent the concentration value ranges of air pollutants such as CO, NO, $NO_2$, $SO_2$, $O_3$, $PM_{10}$ and $PM_{25}$.

In this paper, AQ forecast datasets provided by the Finnish Meteorological Institute and generated by the SILAM model [6] were used, covering the region of Europe between the bounding coordinates of (25° W, 30° N) and (45° E, 72° N), approximately, with an orthogonally projected 352 (lon.) x 220 (lat.) data grid.

For the constructive training of the interpolator functions, a dataset of model data from the SILAM model, spanning the entirety of 2013 was used, and including over 8000 forecasts . As a control set for performance evaluation, a set of 100 randomly chosen forecasts was used from those available.

For the visualization of heatmaps, other than those directly taken from AQ forecasting services, ad-hoc visualization software was developed, written in Java and Matlab.

## 2.1. AQ heatmaps

Due to the nature of the quantization process applied to AQ forecasts, the resulting heatmaps can be regarded as a particular class of non-uniformly sampled signals [2,5]. In addition, due to the specifics of CW forecasting itself, it is possible to make certain assumptions regarding the distribution of values in a CW forecast heatmap [3,7], if the quantization process and the value ranges of the quantization levels employed in the heatmap itself are known. An example of a CW forecast heatmap and its colour scale is given in Figure 1.

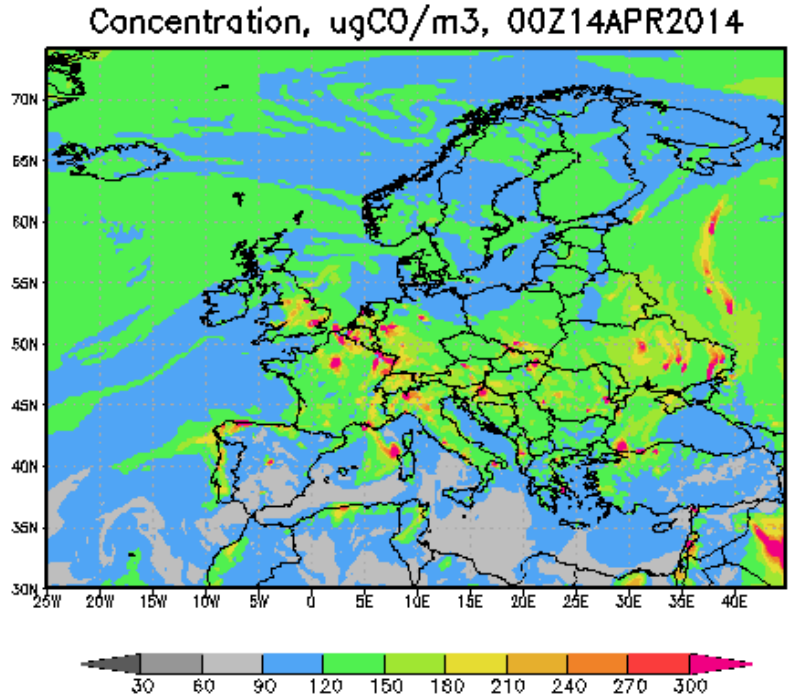


*Figure 1: A sample forecast heatmap, representing CO concentrations over Europe with its colour scale at the bottom (courtesy of FMI),*

Quantizing a CW forecast produces more or less continuous connected contoured regions of local maxima and minima in a quantized signal domain. It can be shown that the transition boundaries (isolines) between such regions coincide with points where the uncertainty of the true value of a

signal is minimized [3], thus they can be used as known threshold crossings or interpolation keypoints, regardless of the interpolation method used. In addition, even if the quantization process classifies ranges of continuous values into discrete bins or levels, using a many-to-one irreversible mapping, this means that the true values of points in the 2D space belonging to a particular classification bin, can only vary between a specific floor and ceiling value, with the values at the borders between different contours being known very precisely or even exactly, due to their transitory nature. This is generally true for most contour points, though small-scale inaccuracies might be introduced by the contouring algorithm, which might cause permanent loss or misclassification of certain small-scale details.

By using these elements, as well as knowledge about the value distributions inside the contours themselves, it is possible to produce an interpolation of higher quality and reliability for the reconstruction of CW data from heatmaps.

## 2.2. Numerical method and interpolation

The method presented in this paper uses a set of pre-trained constraining functions, with their domain defined in the longitude and latitude dimensions of a 2D signal representing CW data on a geographical regions (the base unit may also simply be grid points), and their co-domain defined on the pollutant's concentration value. The functions are produced in sets of triplets representing the maximum, mean and minimum values encountered within a longitudinal or latitudinal section of a contour.

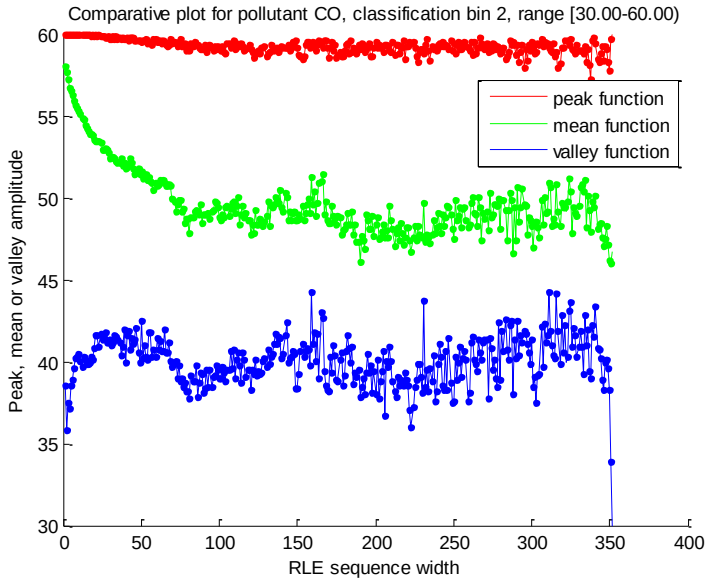


*Figure 2: The structure of constraining function for a specific pollutant (CO) and a particular classification bin (range of 30 to 60 μg/m$^3$) trained on the longitudinal dimension, width expressed in grid units.*

Each quantization level employed in the heatmap has its own function triplet, as shown in Figure 2. These functions allow for the explicit exposure of any relationships that might exist between the spatial dimensions of a quantized contour and the peak, average or minimum value that it might contain. The functions are constructed by statistically analysing several CW datasets (for the time being, of a single dataset provider), and expose that a non-trivial relationship often exists. With this knowledge, it's possible to construct a contour width-dependent interpolator function.

The quantized heatmap is converted to a Run-Length Encoding (RLE) compressed representation, which segments the image in horizontal and vertical scan lines, each containing several RLE sequences, as shown in Table 1. RLE sequences themselves are classed into the following categories, depending on their position and relationship with other RLE sequences: 1) initial/ending maxima or minima 2) local minima or maxima 3) transitional ascending or descending.

The endpoints $P_{start}$ and $P_{end}$ of each RLE sequence are used as constrained interpolation keypoints (any interpolation method used must cross them), as their value is known with the best possible (for a heatmap) accuracy, a property which derives from their construction as part of isolines. An example of how a 1D section of a 2D curve is quantized is shown in Figure 3 and Table 1.

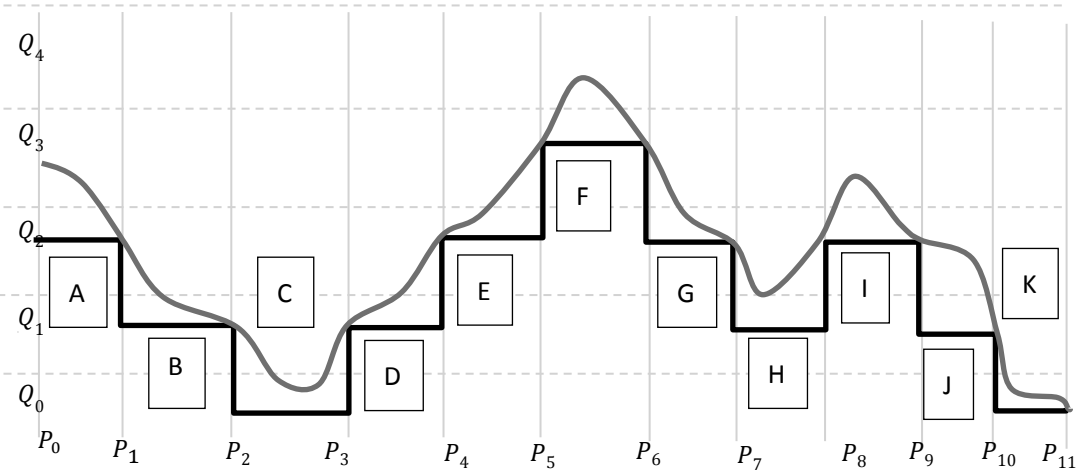


*Figure 3: An example of how a curve is quantized into a finite number of levels and split into RLE sequences, each labelled from A to K.*

| RLE Sequence ID | $P_{start}$ | $P_{end}$ | Quantization level | Classification |
|---|---|---|---|---|
| A | $P_0$ | $P_1$ | $Q_2$ | Initial Maximum |
| B | $P_1$ | $P_2$ | $Q_1$ | Descending |
| C | $P_2$ | $P_3$ | $Q_0$ | Local Minimum |
| D | $P_3$ | $P_4$ | $Q_1$ | Ascending |
| E | $P_4$ | $P_5$ | $Q_2$ | Ascending |
| F | $P_5$ | $P_6$ | $Q_3$ | Local Maximum |
| G | $P_6$ | $P_7$ | $Q_2$ | Descending |
| H | $P_7$ | $P_8$ | $Q_1$ | Local Minimum |
| I | $P_8$ | $P_9$ | $Q_2$ | Local Maximum |
| J | $P_9$ | $P_{10}$ | $Q_1$ | Descending |
| K | $P_{10}$ | $P_{11}$ | $Q_0$ | Ending Minimum |

*Table 1: The RLE Sequences generated by the curve in Figure 3 and their properties.*

Each RLE sequence $k$ is used to compute the amplitude and standard deviation of a Gaussian (normal distribution) interpolation kernel with a general formula of:

$$G_k(x) = A(Q_k, w_k)e^{-\frac{(x-\mu_k)}{{\sigma_k}^2}}, \text{ with } x \in [P_{start}(k), P_{end}(k)] \quad (1)$$

where $A(k, w_k)$ is the peak amplitude that corresponds to a RLE sequence of a given width $w_k = P_{end}(k) - P_{start}(k)$ and quantization level $Q_k$, according to the computed constraining functions, and $\mu_k$ is the k[th] RLE sequence's midpoint. The value of ${\sigma_k}^2$ is computed as:

$${\sigma_k}^2 = -\frac{\left(\frac{w_k}{2}\right)^2}{2\ln\left(\frac{y_{cross}(Q_k)}{A(Q_k,w)}\right)} \quad (2)$$

where $y_{cross}(Q_k)$ is a function which returns the amplitude at the crossing point for a given quantization level (for the example of Figure 3, that would be 30). This way, keypoint crossing by each RLE sequence's interpolation kernel is guaranteed, a property which is retained even when combining vertical and horizontal interpolation passes. Local minima, transitional and initial/ending sequences use a variation of formula (1), which computes inverted Gaussian distributions or only a portion of a Gaussian distribution's curve, as appropriate, but always respecting the constrained crossing and floor/ceiling conditions. The various sequences and their interpolated values thus computed, are subsequently spliced together to form a continuous reconstruction for a given scanline, as is shown in Figure 4.

This interpolation procedure is applied on both the horizontal and vertical dimensions of the quantized heatmap and then the two separate 2D interpolations obtained from each pass are averaged, yielding the final 2D interpolation. It has been determined experimentally, that averaging one horizontal and one vertical interpolation pass gives the best results. It's possible to use additional passes e.g. in reverse for each direction, but those don't always bring benefits, while wasting computational time.

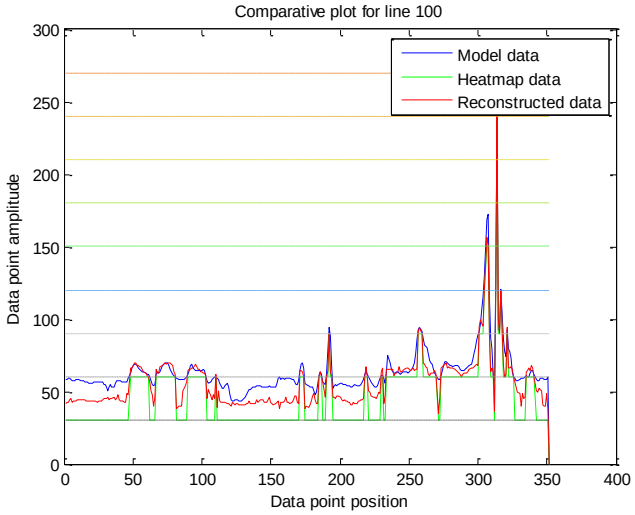


*Figure 4: Comparison between model data, heatmap data (quantized) and reconstructed data for a specific latitudinal line (line 100) of a reference CO forecast.*

## 3. Results

In the conducted tests, it was determined that for sufficient training of the constraining functions, it was enough to use at most 150 to 200 forecasts, spread out over several months of the year and

covering diverse enough conditions in order to avoid overfitting particular seasonal phenomena. Using more forecasts resulted only in marginal interpolation improvements, as can be seen in Table 2 for the PSNR metric, expressed in dB.

Depending on the specific pollutant considered, the Root Mean Squared Error (RMSE) metric between the original forecast numerical data and the reconstructed/interpolated data, was on average reduced by a factor of up to one half compared to the RMSE between the original forecast data and the untreated heatmap data. Similarly, the Peak Signal-to-Noise Ratio (PSNR) metric showed improvements ranging from about 2.0 to 6.0 dB, depending on the pollutant and forecast considered.

| N | CO | $SO_2$ | $NO_2$ | $O_3$ | $PM_{10}$ | $PM_{25}$ |
|---|---|---|---|---|---|---|
| 10 | 5.968 | 1.987 | 2.628 | 4.731 | 2.933 | 2.740 |
| 25 | 6.124 | 2.051 | 2.594 | 4.586 | 2.972 | 2.781 |
| 50 | 6.162 | 2.029 | 2.548 | 4.605 | 2.928 | 2.768 |
| 75 | 6.185 | 2.116 | 2.580 | 4.614 | 2.947 | 2.777 |
| 100 | 6.240 | 2.108 | 2.592 | 4.625 | 2.949 | 2.794 |
| 150 | 6.224 | 2.089 | 2.575 | 4.623 | 2.962 | 2.797 |
| 200 | 6.231 | 2.083 | 2.577 | 4.632 | 2.968 | 2.784 |
| 400 | 6.233 | 2.102 | 2.579 | 4.628 | 2.974 | 2.799 |

*Table 2: Improvement of the PSNR metric (in dB) for interpolating different pollutant types depending on the number N of forecast datasets used for training the constraining functions*

In Table 3, the average improvement yielded by the application of the interpolation, on a control set of 100 forecasts of each kind are shown. On average, the PSNR metric is improved (larger values are better); the ΔRMSE metric (difference between RMSE values) also improves (negative values are better). ΔRMSE (%) indicates the percentage of the RMSE reduction between the RMSE of heatmaps (RMSE-f) and of their reconstruction (RMSE-r), versus the unmodified model data, used as a reference. RMSE values of different pollutants are not directly comparable.

| | CO | $SO_2$ | $NO_2$ | $O_3$ | $PM_{10}$ | $PM_{25}$ |
|---|---|---|---|---|---|---|
| **PSNR [dB]** | 6.171 | 2.071 | 2.584 | 4.630 | 2.071 | 2.780 |
| **RMSE-f** | 19.415 | 0.582 | 0.664 | 11.407 | 1.629 | 1.256 |
| **RMSE-r** | 9.680 | 0.468 | 0.494 | 6.704 | 1.162 | 0.915 |
| **ΔRMSE** | -9.736 | -0.114 | -0.171 | 4.630 | -0.467 | -0.341 |
| **ΔRMSE (%)** | -50.144 | -19.650 | -25.673 | -41.229 | -28.668 | -27.167 |

*Table 3: Average improvement yielded by the proposed algorithm by reconstruction a control set of 100 forecast heatmaps, for each pollutant type. RMSE metrics are expressed in $\mu g/m^3$.*

In general, training was most successful for heatmaps created from linearly spaced quantization scales and with an even distribution of values across their quantization scales (like the ones of the $O_3$ and CO pollutants), which helped avoid phenomena like having too little data points available for higher-class bins. The particular monthly and seasonal periods of the datasets used for training and for control didn't result in appreciable variations in the final results.

In Figures 6c and 6d, the absolute RMSE maps of the model data versus their lossy heatmap and of the model data versus their heatmap-based reconstruction (for a reference forecast) are shown, in terms of CO concentration units ($\mu g/m^3$). The original model data are shown in Figure 6a, their lossy heatmap in Figure 6b and their reconstruction in Figure 6e. The reconstructed version appears smoother than the untreated heatmap, and preserves features like hills and valleys present in the model data, while resulting in a noticeably less noisy RMSE map.

In Figure 6f, the 2D FFT spectrum of the reference CO forecast of Figure 6a is shown, revealing that model data is not, in general, a band limited signal. While most energy is concentrated on lower bands and there is a noticeable DC bias (shown as a horizontal line), the forecast is obviously not sufficiently band limited for inverse spectrum reconstruction conditions to apply [2,5,8]. In such conditions, interpolation methods that operate in the spatial (rather than in the frequency) domain, may be preferable, as there are less problems with finding sufficient quantities of spectrally significant threshold-crossing points, and generally less corner cases to handle.

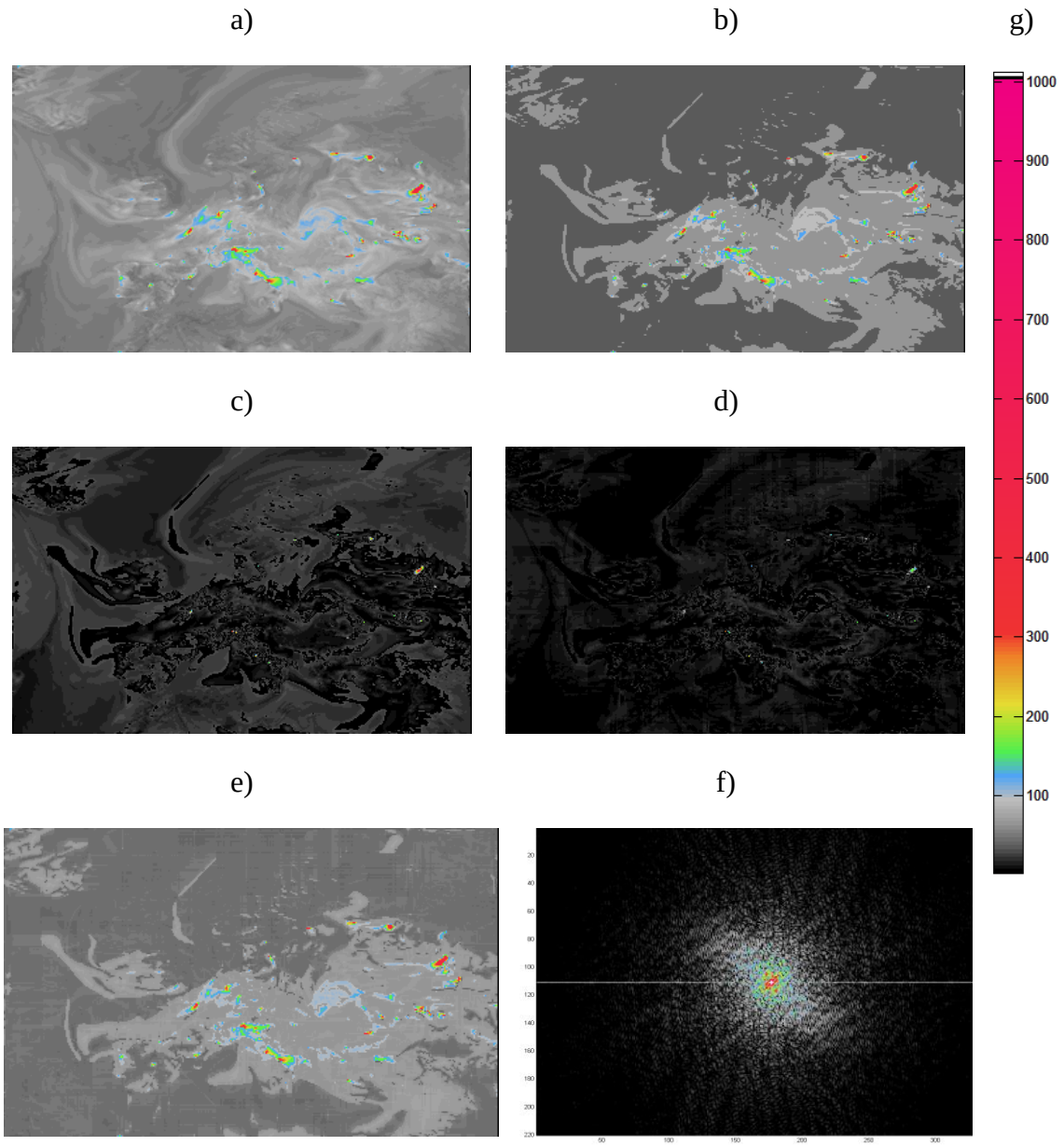


*Figure 6: a) the reference CO forecast b) its quantized version c) the difference map of the model data versus their heatmap version d) the difference map of the model data versus their heatmap-*

*based reconstruction e) reconstructed version with the proposed algorithm f) The 2D FFT spectrum of the reference forecast g) the common colour scale, expressing CO concentration in $\mu g/m^3$. The same color scale is used in all images.*

## 4. Discussion and conclusion

In this paper, a method for the domain-specific interpolation of air quality heatmaps is presented.

The method proved particularly effective at interpolating chemical weather data, providing consistent improvements in the Peak SNR (PSNR) and reductions in the Root MSE (RMSE) metrics for several types of air pollutant heatmaps. In this way, the method can be used to recover a closer representation of the original models' data, by using significantly less data and bandwidth and in a way which is transparent to the model publishers themselves, making it ideal for streaming detailed Chemical Weather coverage information on the internet.

The method was fine-tuned to a specific application domain and using datasets from a specific provider, but it is expected to be adaptable to other domains where heatmaps representing 2D signals with similar spectral and spatial characteristics are employed, as long as a way for constructing or training constraining functions suited to the specific domain is provided.

Ideally, it is expected that when the constraining functions are constructed based on dataset training or by using analytic formulas and models which are close enough to an acceptable ground truth for a particular domain, they will exhibit wider applicability beyond their originating datasets.

For example, training functions trained on the SILAM dataset could be used in different AQ datasets, after accounting for the different spatial and quantization domains. This can be the object of future work, as access to more reference datasets in the same (AQ) or other environmental science domains is gained.

## 5. Acknowledgements

The authors would like to thank the Finnish Meteorological Institute for granting access to their SILAM integrated AQ modelling system's datasets. This publication was sponsored by the "IKY Fellowships of Excellence for Postgraduate Studies in Greece – Siemens Program" at the time of writing.

[1] T. Balk, J. Kukkonen, K. Karatzas, A. Bassoukos, and V. Epitropou, "A European open access chemical weather forecasting portal," *Atmospheric Environment*, vol. 45, no. 38, pp. 6917-6922, December 2011.

[2] L. Donoho, "Compressed Sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289-1306, April 2006.

[3] V. Epitropou, K. Karatzas, J. Kukkonen, and J. Vira, "Evaluation of the accuracy of an inverse image-based reconstruction method for chemical weather data," *International Journal of Artificial Intelligence*, vol. 9, no. 12, pp. 152-171, September 2012.

[4] J. Kukkonen et al., "A review of operational, regional-scale, chemical weather forecasting models in Europe," *Atmospheric Chemistry and Physics*, vol. 12, pp. 1-87, 2012.

[5] M. Mishali and Y.C. Eldar, "Blind Multiband Signal Reconstruction: Compressed Sensing for Analog Signals," *Signal Processing, IEEE Transactions on*, vol. 57, pp. 993-1009, 2009.

[6] M. Sofiev, P. Siljamo, I. Valkama, M. Ilvonen, and J. Kukkonen, "A dispersion modelling system SILAM and its evaluation against ETEX data," *Atmospheric Environment*, vol. 44, no. 4, pp. 674–685, February 2006.

[7] S. Vrochidis et al., "Extraction of environmental data from on-line environmental information sources," in *IFIP Advances in Information and Communication Technology*, vol. 382, 2012, pp. 361-370.

[8] A. Zakhor, "Reconstruction of two-dimensional signals from level crossings," *Proceedings of the IEEE*, vol. 77, no. 1, pp. 31-55, January 1990.