HAL
open science

Infilling missing data of binary geophysical fields using scale invariant properties through an application to imperviousness in urban areas

Auguste Gires, Ioulia Tchiguirinskaia, Daniel Schertzer

To cite this version:

Auguste Gires, Ioulia Tchiguirinskaia, Daniel Schertzer. Infilling missing data of binary geophysical fields using scale invariant properties through an application to imperviousness in urban areas. Hydrological Sciences Journal, 2021, 66 (7), pp.1197-1210. 10.1080/02626667.2021.1925121 . hal-03424307

HAL Id: hal-03424307

https://enpc.hal.science/hal-03424307v1

Submitted on 10 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Infilling missing data of binary geophysical fields using scale invariant properties through an application to imperviousness in urban areas

Auguste Gires, Ioulia Tchiguirinskaia, Daniel Schertzer

Hydrologie Météorologie et Complexité, Ecole des Ponts ParisTech, Champs-sur-Marne, France

* Corresponding author: auguste.gires@enpc.fr

Infilling missing data of binary geophysical fields using scale invariant properties through an application to imperviousness in urban areas

High resolution modelling is needed to improve the understanding and management of storm water in cities. It requires data, which is not always available. Hence the growing importance of handling missing data. Here, we use impervious areas in cities as case study. They are responsible for rapid run-off that can generate surface flooding. A methodology to handle such binary missing data relying on scale invariance properties is presented.

It uses a previous study, which showed on ten peri-urban areas that imperviousness exhibits scale invariant features from meter to kilometre, to generate realistic scenarios for the missing impervious data. More precisely, fractal fields are commonly simulated thanks to a simple binary multiplicative cascade process (beta-model). Here we condition it to the available data.

Numerical simulations are used to confirm theoretical expectations. It is then implemented to infill missing impervious data on a 3 km^2 catchment and the corresponding uncertainty is quantified. Type or paste your abstract here as prescribed by the journal's instructions for authors. Type or paste your abstract here as prescribed by the journal's instructions for authors. Type or paste your abstract here as prescribed by the journal's instructions for authors. Type or paste your abstract here.

Keywords: fractal, multiplicative cascade, missing data, imperviousness

Introduction

Missing data, which can arise due to measurement devices malfunctioning, errors in measurements, natural hazard as well as budget reductions, are ubiquitous in hydrology. The topic is likely to remain crucial in the coming years given the overall tendency, on the one hand to rely more and more on data driven approaches, and on the other hand to physically model catchments at higher and higher resolution. Both approaches require

large amounts of high quality data. It is, by the way, a topic that goes beyond the field of hydrology (Garciarena and Santana, 2017). This explains the motivation of the general, mathematical approach developed by Salvadori et al. (2000) to define a multifractal objective analysis, based on conditioned multifractal conditioning and interpolation.

Given the increasing need for data without gaps, significant efforts to develop infilling techniques have been carried out. It should be noted that the action of filling missing data is also called imputation, completion, reconstruction or patching depending on the authors and/or fields of application. The interested reader is referred to a recent review on the topic of missing data in hydrology by Aissia et al. (2017). Numerous approaches have been suggested, mainly for time series, with various levels of complexity (Aissia et al., 2017; Dumedah and Coulibaly, 2011): mean substitution, time series analysis (using a previously calibrated model to fill the gaps), interpolation (linear regression, kriging, inverse distance weighing...), fuzzy ruled-based methods, pattern recognition, k-nearest neighbor, artificial neural network, expectation minimization or copula based approaches. Some techniques can be applied on single series while others require the use of multiple variables and data from neighbouring stations.

These techniques have primarily been applied to low resolution (typically daily) series of rainfall and stream flow (Bárdossy and Pegram, 2014; Giustarini et al., 2016; Kim and Pachepsky, 2010; Miró et al., 2017; Sivapragasam et al., 2015; Mwale et al., 2012), but also to temperature (Williams et al., 2018; Coulibaly and Evora, 2007; Oyler et al., 2015), soil moisture (Dumedah and Coulibaly, 2011; Dumedah et al., 2014), ground water level (Gill et al., 2007), hydraulic conductivity (Tchiguirinskaia et al., 2004), or evapotranspiration (Abudu et al., 2010).

In this general context, the main purpose of this paper is to introduce and validate a new model to infill missing data of binary geophysical fields. Its underlying concept is introduced through a first application dealing with the overall issue of the handling of a specific missing input data required for high resolution (typically few meters of tens of meters) hydro-dynamic modelling in urban areas. Indeed there is a growing demand for such kind of modelling by researchers and stakeholders to improve the understanding and management of storm water in urban areas, and notably for accurate prediction of local pluvial flooding. A current limitation is the fact that when dealing with high resolution of the order of few meters, there is often a lot of missing data with regard to land cover, topography, or soil properties (Ichiba et al., 2018). Imperviousness is of paramount importance in cities because impervious areas are responsible for rapid run-off that can generate surface flooding. The objective of the paper is to develop, and test on an existing case study as a proof of concept, an innovative technique to infill the missing data concerning the distribution of impervious areas. Being treated as binary question, i.e. whether a given pixel is impervious or not, this enables to deploy an approach tailored to this particular feature.

The goal of this paper is to develop an approach that preserves the key property of scale invariance and is stochastic. Indeed, a previous study carried out on 10 European urban and peri-urban areas has shown that imperviousness exhibits scale invariant features on scales ranging from few meters to few kilometers for all the case studies (Gires et al., 2017). It was possible to assess for each area a fractal dimension, which enables to parsimoniously characterize in a scale-invariant way the space filled by a geometrical set in its embedding space. It ranged from 1.6 to 2 according to the area and quantified the level of urbanization. A stochastic technique to infill the missing data is furthermore much needed in order to obtain not only a single deterministic

outcome, but an ensemble of realisations actually yielding to an empirical probability distribution of the outcome for all the pixels where data on imperviousness is missing. Such approach enables to address the issue of uncertainty of the generated values for the missing data, which is seldom done while much needed (Aissia et al., 2017; Bárdossy and Pegram, 2014).

The so called β-model (Frisch et al., 1978) appears as a relevant starting point for the wanted approach, which can be seen a special pedagogic case of Salvadori et al. (2000). Indeed, it is used to simulate fractal fields through a simple stochastic binary discrete multiplicative cascade. Using such model enables to not only preserve the scale invariance but actually intrinsically rely on it. A methodology to condition the β-model, i.e. ensure that the generated field of imperviousness has the correct values at the available locations, is developed and constitutes the core of the paper.

In section 2, the conditional β-model is described after a reminder on the computation of fractal dimensions and the β-model. In section 3, its implementation and possibilities of use are discussed through numerical simulations and an illustration with actual rainfall occurrence patterns. Finally, results on its implementation with imperviousness data over a 3 km^2 semi-urbanized catchment of the Paris area along with its hydrological impacts are presented and discussed in section 4. It should be stressed that the model and case study were already available, which suits well the purpose here which is to illustrate the possible uses of the developed conditional β-model on a practical case. It does not enable to explore all the possibilities of the conditional β-model which are nevertheless discussed. Main conclusions are in section 5.

Description of the model

The notion of fractal dimension

Before presenting the β-model and its conditional version, it should be briefly reminded what is a fractal dimension and how it is computed. Let us consider a bounded geometrical set A of outer scale l_0. It can be either in 1D or 2D. Its fractal dimension D_F quantifies its sparseness, i.e. how much space it fills across scales. The fractal co-dimension c_F of a field is simply equal to :

$$c_F = d - D_F \qquad (1)$$

where d is the dimension of the embedding space (d = 1 in 1D for time series and d = 2 in 2D for maps).

In order to compute D_F, let us represent A at different scales l and introduce the notion of resolution λ defined as the ratio between the outer scale and the observation scale ($\lambda = \frac{l_0}{l}$). $N_{\lambda,A}$ is number of non-overlapping elements of size l (time steps in 1D or pixels in 2D) needed to completely cover A. If A is a fractal object, then $N_{\lambda,A}$ is power law related to the resolution in the limit $\lambda \to +\infty$, and the corresponding characteristic exponent is called fractal dimension. Mathematically, it means that we have:

$$N_{\lambda,A} \approx \lambda^{D_F} \qquad (2)$$

This approach to roughly estimate fractal dimensions is called box counting (Hentschel and Procaccia, 1983; Lovejoy et al., 1987). In practice, the analysis is initiated at the maximum resolution and the field is then iteratively upscaled by increasing the observation scale l by a factor of two at each step.

Brief description of the β-model

The so called β-model was initially introduced to model turbulence (Frisch et al., 1978). It enables to simulate fractal fields through a simple binary discrete random multiplicative cascade process. These cascades were later generalized to handle not only geometrical sets (i.e. binary fields) but fields exhibiting various values (Schertzer and Lovejoy, 1987). See Schertzer and Tchiguirinskaia (2020) for more recent developments including vector fields. Besides atmospheric turbulence, it has also notably been used to model rainfall occurrence (see Over and Gupta, 1996, and Molnar and Burlando, 2005, for examples among others).

Discrete random multiplicative cascades use an iterative process to distribute in space and time an average intensity initially homogeneously distributed on a large scale structure. A cascade step consists in: (i) Dividing a structure into λ_1^d sub-structures with *d* being the dimension of the embedding space. The commonest choice is to set λ_1 =2. It enables to maximize the number of cascade steps for a given final size of simulation. (ii) Affecting an intensity to the sub-structure equal to the intensity of the parent structure multiplied by a random ``multiplicative increment" denoted $\mu\varepsilon$.

As a result, after *n* steps, i.e. at resolution $\lambda = \lambda_n = \lambda_1^n$, a value of the generated field $\mu\varepsilon_n$ is simply equal to the product of all the corresponding multiplicative increments:

$$\varepsilon_n = \varepsilon_0 \prod_{i=0}^{n} \mu\varepsilon_i \qquad (3)$$

In the specific case of the so-called β-model used here, the multiplicative random increment is the simplest possible and has only two states. It is either dead (equal to 0) or alive (>0) with the following probabilities:

$$\Pr(\mu\varepsilon = 0) = 1 - \lambda_1^{-c} \qquad (4)$$

$$\Pr(\mu\varepsilon = \lambda_1^c) = \lambda_1^{-c} \qquad (5)$$

As a consequence of Eq. 3, ε_n has only two possible values, either 0 or λ_n^c. c is the characteristic parameter of the model and is equal to the fractal co-dimension of the geometrical field made of the 'alive' portion of the generated field. Fig. 1shows some simulations in 1D and 2D with various c to give the reader an intuitive feeling of the physical meaning of c. One can note the visible square structures, notably in 2D. This feature is indeed not realistic and a limitation of this model. It is due to the discrete nature of its underlying construction scheme (see Schmitt, 2014for a continuous version of the β-model). Conservation of the average activity is ensured since $\langle \mu\varepsilon \rangle = 1$. It should be noted that this conservation is true only on average, i.e. on numerous realisations, and not for individual ones.

The conditional β-model

In this sub-section, the conditional β-model which is at the core of this paper is introduced. The developed process is illustrated in Fig. 2, which will be used for pedagogical purposes, while the text description is generic. More precisely, we start with a field at a given resolution made of zeros (in white), ones (in black), and missing/unknown data (in yellow). The goal is to simulate realistic values for the missing portion of the field, while ensuring that the final field has the desired statistical properties; i.e. that it exhibits a fractal behaviour with the correct fractal dimension.

In order to achieve this, we suggest to use a conditioned version of the previously introduced β-model. The assumption is that the field is generated through a multiplicative cascade process as described in section 2.1. This means that the final field is actually fully defined by all the multiplicative increments of its underlying cascade process (Fig. 2.a). For a series of length 2^n, the increments to be filled can be denoted $\mu\varepsilon_{k,i}$, where k is the cascade step ranging from 0 to n and i ranges from 1 to 2^k. They are

displayed in yellow on Fig. 2.a, since at the beginning of the process they are unknown. Hence, the whole concept of the conditional β-model simply consists in affecting a value to each of the cascade's multiplicative increments enabling to both reproduce the available values and the statistical features. This is what is called conditioning a β-model in this paper.

The general idea of the process presented in this paper is actually rather similar to the one discussed in Tchiguirinskaia et al. (2004). It basically consists in first setting 'manually' the increments needed to ensure that the available values are correctly retrieved (i.e. first level of conditioning of the β-model). In a second steps, the remaining increments are simply stochastically drawn using the laws of Eq. 5, with a parameter c yielding to the wanted fractal dimension (i.e. second level of conditioning of the β-model). In Tchiguirinskaia et al. (2004), it is done with continuous Universal Multifractals (Schertzer and Lovejoy, 1987), meaning that it is not the increments but the generator that need to be set. It might be seen like adding complexity, while enabling to handle non binary fields. However, the process does not ensure that the exact values at the available locations are retrieved, but only approximations. The process is stochastic, meaning that the output is not a deterministic field, but an ensemble of realistic realisations.

More precisely, the three successive steps of the process are detailed with the illustrations in Fig. 2:

- First, the required alive increments are filled, i.e. for each strictly positive value of the field, the chain of increments (one per cascade step) leading to this value is all set to one (i.e. 'alive'). In the example of Fig. 2, there is only one non-null available value and the corresponding increments which then must be set to 'alive' are circled in red in Fig. 2.b.

- Second, the increments that are needed to correctly reproduce the zeros of the initial field are set to zero (i.e. 'dead'). Given that it is a multiplicative process and that a single zero in the increments chain leading to a final value is sufficient to have it equal to zero, the process is slightly more elaborated than for the 'alive' values. In practice, each zero of the initial field is treated one after the other. The order is randomly chosen (in practice a random permutation between their positioning indexes is run and they are treated in the obtained order). Then for a given zero, the chain of increments leading to it is considered. If there are no zeros, then a single increment among the available ones (there might already be some 'alive' increments in the chain, and they should not be changed) is set to zero ('dead') and the others are left unknown. Let us remind that given Eq. 3, a single zero is sufficient to ensure that the final value (i.e. the product of the multiplicative increments) will be equal to zero. Hence there is no need to set more than one to zero. The increments set to zero is randomly chosen among the available ones (in practice, a random permutation between their defining position indexes is run and the first element is selected). If there is already a zero in the chain of increments, then nothing is done. This is illustrated in Fig. 2.c. In this example, the first zero treated is the one circled in red, with its corresponding increment at cascade step 3 set to zero (the ones of cascade step 0, 1 and 2 were already set to one and could not be altered). Then it was the green one (increments at cascade step 1 is set to zero and the ones at cascade step 2 and 3 are left unknown) followed by the dark blue one (increments at cascade step 2 is set to zero while the one at step

3 is left unknown). The last one was the light blue one. For this time step, since a zero was already in its increments' chain, no additional zero was added. This process limits the number of increments set to zero while ensuring that available values are correctly retrieved.

- Third, all the increments which are not yet filled are randomly drawn by using the probabilities distribution of Eq. 5. Once a value has been affected to all the increments, it is straightforward to generate a realisation of the field. See Fig. 2.d for an illustration. See section 3.3 if the parameter *c* is unknown.

Results and discussion with numerical simulations

Individual fields

In this section, the results obtained with individual fields are discussed. It is done through an example shown in Fig. 3. The initial field was generated with the help of a β-model with 7 cascade steps (n=7, leading to series of length 128) and c=0.2. The field exhibits an excellent fractal behaviour (Fig. 4) with $D_F = 0.79$. This value is close to the expected one of 0.8 (=1-c), and the slight difference is normal given that it is a single realisation and ensemble computations with numerous samples would yield much closer results.

Then, a portion p of the field is set to missing data. In this example we used $p = 0.5$ meaning that half of the field is removed. This is done randomly, and here, it turns out that 64 time steps were 'removed' and are now considered as missing. The partial field, which is the one used as input in the conditional β-model is displayed in Fig. 3. Ones are in black, zeros in white and missing data in yellow.

The next step is to implement the conditional β-model to simulate possible realisations for the missing data. Since it is a stochastic process, various realisation ($N_{realisation}$) can be generated. Here we used $N_{realisation} = 100$. For each realisation, a percentage of hits (a hit rate), i.e. a time step correctly guessed, can be computed by comparing the generated values with the initial ones for all the missing data. In this example, we find that the average % of hits ($\%_{avg}$) over all the realisations is 81% (with a standard deviation of approximately 5%). It should be noted that all the realisations exhibited the same excellent scaling behaviour with similar fractal dimensions ($D_F = 0.79 \pm 0.01$).

The average value over all the realisations is displayed in Fig. 3 on the 4th row with shades of grey. White and black correspond to 0 and 1 respectively. In addition to improve clarity, the same quantity is plotted as a time series in yellow ranging from 0 to 1. This quantity is actually an empirically computed probability for the time step to be equal to one, which corresponds to a probability distribution in this binary framework. Depending on the application, the whole information contained in this probability distribution could be used. In the context of this paper, and in order to remain in a binary framework, we suggest to use it to extract a 'most probable' (*mp*) field by setting a time step to one if this average is greater than 0.5 and zero otherwise. This rather natural threshold of 0.5 is actually not arbitrarily chosen. Indeed, a sensitivity analysis showed (not displayed here) that it corresponds to the values leading on average over numerous samples to the maximum percentage of hits for this 'most probable' field. This is the field showed in the last row of Fig. 3. The percentage of hits computed on this field ($\%_{mp}$) is equal to 87 in this example. Similar results are found with other initial fields.

Similar results are found in 2D and illustrated in Fig. 5 for an initial field obtained with 5 cascade steps and c=0.2. The same p=0.5 as before was used and yielded in this case 49% of missing data. D_F on the initial field is equal to 1.81 here with an excellent scaling behaviour. On this example we find $\%_{avg} = 79$ and $\%_{mp} = 85$.

Influence of the various parameters

In this section, the sensitivity of the results to the parameters c of the β-model and p is investigated. In order to achieve this, the same process as the one described in the previous section is implemented on 200 samples of initial fields for each set of parameters (c, p). Results are presented for simulations in 1D, but similar ones are obtained in 2D.

Fig. 6 shows the % of hits on missing data (called 'hit rate' after) as a function of p for c=0.1. Black lines are for the average over realisations while red ones are for the 'most probable' fields (see previous section). The uncertainties retrieved over the 200 initial samples is represented as follow: the solid line is the 50% quantile, and the dash ones are the 10 and 90% quantiles. As expected, the % of hits decreases with increasing values of p, and it is now quantified. The performance in terms of scores appears to be 5 to 10% better with the 'most probable' approach, confirming its relevancy. Such results are found for all the other values of c. Finally, it should be noted, that the width of the uncertainty interval is rather great (typically 15-20%) meaning that there is a significant variability of the results from one sample of initial field to the other.

Fig. 7 displays the % of hits on missing data as function of p for various values of c. To keep a readable graph, only the 50% quantile over the 200 samples of initial fields with the 'most probable' approach is plotted. It corresponds to the red solid line in

Fig. 6. The decrease of scores with increasing values of p is found for all c. One can note that even in the worst case, i.e. removing 90% of the initial field with c=0.2, the % of hits remains greater than 70%. Such achievement is possible because the developed model relies on the underlying robust structure of a multiplicative cascade. For a given value of p, there is a tendency of the scores to be better for greater values of c. It suggests that the conditional β-model is better at reproducing the zeros than the ones. It is likely to be due to the fact that generating an alive time step requires having all the multiplicative increments alive which turns out to be trickier to get. It should be noted that this tendency actually arises only after a sufficient number of cascade steps is reached (typically n>6-7).

How to proceed when c is unknown?

A tricky point is how to proceed when the parameter c of the β-model is unknown. Indeed, in the previous section, c was always assumed to be known. In such situations, it is suggested to implement the following algorithm, where c_{algo} is the parameter c used for the simulations:

- Step 1 : set c_{algo} to a value. As we will see later, it can be done rather arbitrarily (actually the output of the algorithm is not sensitive to this initial value), but a natural and simple choice is to set it to $d - D_{F,ini}$ where $D_{F,ini}$ is the fractal dimension of the field assuming the missing data is equal to zero.
- Step 2: compute the 'most probable' field using this c_{algo}.
- Step 3: estimate the fractal dimension of the retrieved field and update c_{algo} to d minus this fractal dimension.
- Step 4: repeat step 2 to 4.

Here are some results obtained in 1D with an initial field generated with $c=0.2$ and p=0.7. The fractal co-dimension of the initial field is equal to 0.19. Fig. 8 displays the evolution of the parameter c_{algo} as well as $\%_{mp}$ as a function of the number of iterations in the algorithm. For illustrative purposes, the algorithm was run three times with three different initial values for c_{algo} (0, 0.3 and 1). Running the algorithms several times with the same initial c_{algo} yields the same results. The first striking feature is that the algorithm rapidly converges toward an asymptotic value of the fractal dimension after only few iterations (Fig. 8.a). Furthermore, this values does not depend on the initial value of c_{algo}. The combination of these two features suggests that the algorithm is robust and can simply be stopped as soon as two successive estimates of c (i.e. the fractal co-dimension of the alive portion of the field) differ by less than 0.05. The final value of c retrieved here is of 0.20, which is quite close from the input value of 0.19. Similar results are found for other values of c and p in both 1D and 2D.

Implementation on rainfall data

The purpose of this section is to carry out a first implementation of the developed conditional β-model on actual geophysical data. A 5 min rainfall times series of length equal to 2048 time steps is used. It corresponds to a duration of roughly 7.1 days. The series starts on 2019-06-02 00:00:00 (UTC). Studied time series is displayed Fig. 9. The data was collected on the roof of the Carnot-Cassini building of the Ecole des Ponts ParisTech campus near Paris by an OTT Parsivel^2 disdrometer (Battaglia et al., 2010; OTT, 2014). It is part of the TARANIS observatory of the Fresnel Platform of École des Ponts ParisTech (https://hmco.enpc.fr/portfolio-archive/fresnel-platform/). Interested reader is referred to Gires et al. (2018), which presents in details the available data base with some data samples for a similar measurement campaign. The total cumulative

depth of the series is of 29 mm. It contains 323 rainy time steps, i.e. ≈ 16 %. Its occurrence pattern, i.e. the geometrical set corresponding to the rainy time steps, exhibits an excellent scaling behaviour with a fractal dimension of 0.71 (see Fig. 10).

The implementation of the conditional β-model is illustrated on Fig. 11. The 2048 time steps long occurrence pattern is displayed on top. Below, a partial initial field with approximately half of the time steps set to unknown is shown (obtained with p=0.5). It basically means that 50 % of the recorded data is set aside, and it is tried to reconstruct it with the help of the developed algorithm. Such degraded data was used as input into the algorithm to check its validity. The algorithm was initiated with c_{algo} = 0.1, and two iterations were enough to ensure convergence. More precisely, the first iteration yielded a c equal to 0.28, while the second one yielded a value of 0.28 as well, which is also the expected value ($1 - D_{F,rainfall}$). Fig. 11 displays the outcome after this second iteration of the algorithm: one realisation, average, and 'most probable' field. $\%_{mp} = 94$ is found, which is very good. Since there are numerous zeros in the field, the hit rate (the % of correct hits on missing data) by assuming that all the missing time steps are equal to zero was also computed and found to be equal to 85%. This highlights the benefice of the conditional β-model. In summary, this first implementation of the conditional β-model on actual data enables to confirm its relevancy and efficiency.

It was also tested to threshold the series before implementing the conditional β-model, i.e. to artificially set to zero all the values below a given threshold T. It turns out that as long as $T <\approx 1$, the resulting field still exhibits good fractal behaviour ($r^2 > 0.98$ in the computation of D_F). The conditional β-model also performs well with $\%_{mp}$ scores for p=0.5 equal to 97 and 98% for T equal to respectively .2 and 1 mm.h^{-1} (the hit rate assuming all missing data is equal to zero is equal to 91 and 95 %). For larger thresholds the quality of the scaling becomes dubious meaning that the developed

approach is no longer relevant. These findings: (i) are consistent with the intuitive notion of multifractality, i.e. that a multifractal field consists in a hierarchy of embedded fractal fields. (ii) confirm that a simple threshold is not appropriate to address this issue and that the notion of singularity, i.e. a kind of scale invariant threshold would be required. This would result in a qualitative shift which is outside the scope of this paper. Some elements in this direction can be found in Tchiguirinskaia et al. (2004).

Infilling missing data of imperviousness

Presentation of the model and studied catchment

The model and case study are actually exactly the ones that were used in Gires et al. (2018), so only the main elements are summarized here.

The hydrodynamic model used is Multi-Hydro, which is developed at Ecole des Ponts ParisTech (El Tabach et al., 2009; Giangola-Murzyn et al., 2012; Ichiba et al., 2018). It is a fully distributed, physically based, scalable model that basically consists in a interacting core between widely validated existing models, each of them representing a portion of the water cycle in urban environment.

In this paper, only the surface and drainage model are used. The surface model is based on TREX (Two dimensional Runoff, Erosion and eXport model, Velleux et al., 2011), which models surface flow through a diffusive wave approximation of 2D Saint-Venant equations, and infiltration through a simplification of Green and Ampt equation. The drainage model is based on SWMM, developed by the US Environmental Agency (Storm Water Management Model, Rossman, 2010), which models flow in the sewer network through a numerical solution of 1D Saint-Venant equations. See Ichiba et al. (2018) and references therein for examples of implementations.

Multi-hydro requires to rasterize available GIS (Geographical Information System) data into a regular grid with a fixed pixel size. For the determination of the unique land use class for a given pixel, a priority order based on the hydrological importance is used. A pixel's land use class will be the one of the class with the highest priority level it contains, regardless of its surfacic significance. The order used here is: gully (because the two way interactions between surface and drainage flow is handled through these pixels), road, house, forest, grass. This is not the only possible approach to rasterize and comparison with one based on surface significance can be found in Ichiba et al. (2018).

The catchment that is studied in this paper, is a 3.017 km^2 semi-urban area located in Jouy-en-Josas (Yvelines County, South-west of Paris). It is mainly on the left bank of the Bièvre River, which basically flows from West to East in the South of the catchment. There is a rather steep slope in the middle of the catchment, with an altitude difference of roughly 100 m between the North of the catchment and its outlet (South East). More details on the catchment and its flooding history can be found in Gires et al. (2018).

Fig. 12.a and 13.a display a representation of the studied catchment with pixels of size 2 and 10 m respectively. The pixels in yellow correspond to data not available, which were simply filled with 'grass' in Gires et al. (2018). They represent 16.2 % of the pixels with 2 m pixels and 5.6 % with 10 m pixels. As it can be seen in Fig. 12.a and 13.a, they are primarily located within the urbanized portion of the catchment around the buildings, meaning that they could typically correspond to small gardens attending the individual house or private drive ways / parking lots. Such areas were not identified on the available GIS data (BD ORTHO, professionnels.ign.fr). This suggests that the automatic treatment performed to obtain the data from the available areal photographs

does not enable to distinguish such type of area at the required resolution. Improving the treatment of such pictures to refine the data could be a relevant approach, but it is another field of expertise. Here, the implementation of the conditional β-model will enable to distinguish in a simplified binary way whether these pixels are pervious or impervious, and hence behave in a very different manner hydrologically speaking.

Results and discussion on filling the gaps with the conditional β-model

In this section, we implement the conditional β-model on the land use representation of the Jouy-en-Josas catchment. More precisely, two classes of pixels are distinguished in order to retrieve the binary framework of the previous sections: impervious pixels (i.e. here the pixels corresponding to gully, road and building), and pervious ones (i.e. here forest and grass).

The geometrical set consisting of the impervious area of this catchment was found to exhibit an excellent scaling behaviour, with a fractal dimension equal to ≈ 1.7 (see Fig. 14, for a computation on a 1024 x 1024 pixels area) (Gires et al. 2017, 2018). Such fractal behaviour suggests that it is possible to use the developed conditional β-model to fill the missing data.

For technical reasons, the β-model requires to work with square fields whose size is equal to a power of two, since in the discrete cascade model, each structure is divided into two sub-structures which enables to maximise the number of cascade steps for a given final size. In practice, here, the field is embedded in a larger one matching the requirement, which slightly increases computation costs. The added pixels are considered as missing. After the conditional β-model has been run, they are simply removed and all the outputs analysis are carried out on the initial field. In order to limit possible bias associated with the addition of large areas with missing data, the fractal

dimensions in the conditional β-model are computed only on the portion corresponding to the initial area.

Fig. 12.b and 13.b display a representation of the studied catchment with pixels of size 2 and 10 m respectively, with missing data filled using the 'most probable' approach. The missing data found to be impervious ('alive' in the β-model) is in blue, while the one found to be pervious ('dead' in the β-model) is in green (considered to be grass). As it could be expected, given the lower initial amount of impervious pixels at 2 m and the priority rule set, a greater portion of pervious areas was generated at 2 m. The quality of the scaling behaviour is slightly improved with this most probable field (r^2 goes from 0.998 to 0.999), and the fractal dimension estimate is slightly increased as well from 1.72 to 1.76 (Fig. 14).

The current version of the hydro-dynamic model only enables to have a limited number of land use classes. Hence, it was chosen to use the binary output of the conditional β-model in the form of the 'most probable' field. However, in terms of perspectives, with models having a different approach to represent land use distribution, the actual probability distribution of the output for a given pixel could be used. Here, for example, it could mean using the average output over numerous realisations to define a given level of imperviousness for each pixel; rather than the binary approach presented here. The purpose of this is simply to illustrate the other possibilities of the developed conditional β-model. Their actual implementation should be carried out in further investigations.

Results and discussion on the hydrological consequences

Hydrodynamic simulations were then carried out with the help of the Multi-hydro model with pixels of size 10 m. The initial land use field with missing data taken

as grass, along with the 100 realisations of realistic ways of filling the 5.6% of missing data, and the most probable field were used as input. All the other inputs are kept the same. A moderate rainfall event that occurred over roughly 10 hours on 9 February 2009 and resulting in an average cumulative depth of 9.4 mm over the catchment is used. Distributed C-band radar rainfall provided by Meteo-France is used. This event was already studied in Gires et al. (2018) and presented in more details there.

As in Gires et al. (2018), a pseudo coefficient of variation CV'_{95} is used to quantify the variability within the 100 realisations of the generated ensemble (here of land use distribution). For a given quantity, it consists in taking half of the difference between the 5 and 95 % quantile divided by the 50 % quantile (median) over the 100 realisations:

$$CV'_{95} = \frac{1}{2}\frac{Q_{95\%} - Q_{5\%}}{Q_{50\%}} \qquad (6)$$

where Q is the studied quantity, here peak flow or maximum water depth for a given pixel.

The outcomes of the simulations are shown in Fig. 15. In panel (a) the flow at link #507 is visible. It drains the water from the North of the catchment. A zoom during the peak flow is in Fig. 15.b. The relative difference at the time of peak flow between the 'initial' field and the 'most probable' one is of 2%. The uncertainty range is limited and the CV'_{95} between the various realisations is of only 0.5%. These variations are limited, but given that there was only 5.6% of missing data, it was expected. Implementing it with the 2 m representation for which there are 16.2% of pixels which data is missing would yield stronger variability.

In term of maximum depth for each pixel, the difference between the 'most probable' case and the 'initial' one is displayed in Fig. 15.c. It reaches 2 cm and is

obviously mainly located on the pixel corresponding to missing data. The CV'_{95} for this maximum depth is displayed in Fig. 15.d. It can reach more than 5% for some pixels, in the area that concentrated the more missing data.

Conclusions and perspectives

In this paper, a conditional version of the β-model is presented. It is designed to fill the gaps of missing data. The β-model is a binary discrete multiplicative cascade process meaning that a final field is entirely defined by the multiplicative increments of the underlying cascade. Hence, in the conditional β-model, the multiplicative increments necessary to obtain the desired values at the available locations are set, and then the remaining ones are simply stochastically drawn. As such, an ensemble of possible fields can be obtained, from which a 'most probable' field is be derived. It corresponds to setting to 1 the portion of the field for which the probability of having a 1 exceeds 0.5. This threshold maximizes the performance of the model. The main advantages of this conditional model with regards to other approaches are: (i) it is physically based in the sense that it preserves and actually relies on underlying scale invariant properties of the studied geophysical fields; (ii) it can intrinsically work at any resolution; (iii) requiring only one parameter (the fractal dimension of the studied field), it is parsimonious; (iv) it requires limited computational power.

This approach was first tested on numerical simulations, notably to quantify the sensitivity of the conditional β-model to the various parameters, i.e. mainly the fractal dimension of the studied field and the percentage of missing data. When randomly removing up to 90% of a field, the hit rate with this approach when trying to reconstruct the original field is greater than 70 %. Such figures can be achieved thanks to the underlying robust structure inherited from the cascade process. The algorithms were

then tested on a rainfall occurrence pattern, which enabled a first validation of the developed approach on geophysical data. Finally, the conditional β-model was used to infill the gaps of missing data of imperviousness at high resolution over a 3 km^2 area nearby Paris. The ensemble of possible fields was then used as input in an available hydro-dynamic model in order to quantify the uncertainty associated to these missing data. In this specific case, due to the available data (only 5% of missing data) and resolution used for running the model, the hydrological consequences are limited, with for example less than 2% of differences between the peak flow simulated with the various inputs. Such uncertainty in surface and sewer flow is small and anyway smaller than the one associated with other sources of uncertainty such as the spatio-temporal rainfall distribution, or the roughness coefficient variability for which a single value per land use type was used. Despite the limited impact, this implementation nevertheless enabled a successful proof of concept of the developed methodology, which was the purpose of this paper. Future studies with other case study, and notably cases where the impact of missing data is more striking (for example in areas where measurements are sparser) would enable to confirm the usefulness of the developed approach.

In addition, it would be needed in further investigations to carry out a comparison of this conditional β-model with existing methods such as the ones already discussed in the introduction; and also with other underlying geophysical fields. This would notably enable to identify more clearly its advantages and limitations depending on the context.

In this paper, a simple and rather pedagogical conditional model is developed. Yet, this conditional β-model is applicable and has been successfully implemented with rainfall and imperviousness in urban areas. Further investigations should be devoted to

shifting from binary fields to full fields and also removing the visible square structures, which would require using continuous multifractal fields.

Acknowledgments

The authors greatly acknowledge partial financial support from the Chair of Hydrology for Resilient Cities (endowed by Veolia) of the Ecole des Ponts ParisTech.

References

Abudu, S., Bawazir, A.S. and King, J.F., 2010. Infilling missing daily evapotranspiration data using neural networks. *Journal of Irrigation and Drainage Engineering*, 136 (5), 317-325, 2010. doi: 10.1061/(ASCE)IR.1943-546 4774.0000197.

Battaglia, A., *et al*., 2010. PARSIVEL snow observations: a critical assessment. *Journal of Atmospheric and Oceanic Technology*, 27 (2), 333–344. doi:10.1175/2009JTECHA1332.1.

Bárdossy, A. and Pegram, G., 2014. Infillling missing precipitation records - a comparison of a new copula-based method with other techniques. *Journal of Hydrology*, 519, 1162-1170. doi: https://doi.org/10.1016/j.jhydrol.2014.08.025

Ben Aissia, M.-A., Chebana, F. and Ouarda, T., 2017. Multivariate missing data in hydrology – a review and applications. *Advances in Water Resources*, 110, 299-309. Doi: https://doi.org/10.1016/j.advwatres.2017.10.002

Coulibaly, P. and Evora., N.D., 2007. Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, 341 (1), 27 – 41.

Dumedah, G. and Coulibaly, P., 2011. Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. *Journal of Hydrology*, 400 (1), 95-102. Doi: https://doi.org/10.1016/j.jhydrol.2011.01.028.

Dumedah, G., Walker, J. P. and Chik, L., 2014. Assessing artificial neural networks and statistical methods for infilling missing soil moisture records. *Journal of Hydrology*, 515, 330-344. Doi: https://doi.org/10.1016/j.jhydrol.2014.04.068

El-Tabach, E., *et al.*, 2009. Multi-Hydro: a spatially distributed numerical model to assess and manage runoff processes in peri-urban watersheds. In: E. Pascheet, et

al., eds. *Final Conference of the COST Action C22, Road map towards a flood resilient urban environment*. Paris, France: Hamburger Wasserbau-Schriftien

Frisch, U., Sulem, P.-L. and Nelkin, M., 1978. A simple dynamical model of intermittent fully developed turbulence. *Journal of Fluid Mechanics*, 87 (4), 719-736. doi: 10.1017/S0022112078001846.

Garciarena, U. and Santana, R., 2017. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, 89, 52 – 65. doi: https://doi.org/10.1016/j.eswa.2017.07.026

Giangola-Murzyn, A. *et al.*, 2012. Multi-Component Physically Based Model to Assess Systemic Resilience in Paris Region. *Proceedings Hydro-Informatics Conference*, Hamburg, 14-18 July 2012, Germany.

Gill, M. K., *et al.*, 2007. Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. *Water Resources Research*, 43 (7). doi: 10.1029/2006WR005298

Gires, A., *et al.*, 2017. Fractal analysis of urban catchments and their representation in semi-distributed models: imperviousness and sewer system. *Hydrology and Earth System Sciences*, 21 (5), 2361-2375, doi: 10.5194/hess-21-2361-2017

Gires, A., Tchiguirinskaia, I. and Schertzer, D., 2018. Two months of disdrometer data in the Paris area. *Earth System Science Data*, 10 (2), 941–950. doi:10.5194/essd-10-941-2018.

Gires, A., *et al.*, 2018. Multifractal characterisation of a simulated surface flow: a case study with multi-hydro in Jouy-en-Josas, France. *Journal of Hydrology*, 558, 483–495. doi: 10.1016/j.jhydrol.2018.01.062

Giustarini, L., *et al.*, 2016. A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. *Environmental Modelling and Software*, 82, 308 – 320. doi: https://doi.org/10.1016/j.envsoft.2016.04.013

Hentschel, H. and Procaccia, I., 1983. The infinite number of generalized dimensions of fractals and strange attractors. *Physica D: Nonlinear Phenomena*, 8, 435 – 444. doi: https://doi.org/10.1016/0167-2789(83)90235-X

Ichiba, A., *et al.*, 2018. Scale effect challenges in urban hydrology highlighted with a distributed hydrological model. *Hydrology Earth Systems Sciences*, 22, 331–350. doi:10.5194/hess-22-331-2018

Kim, J.-W. and Pachepsky, Y. A., 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, 394, 305 – 314. doi: https://doi.org/10.1016/j.jhydrol.2010.09.005.

Lovejoy, S., Schertzer, D. and Tsonis, A. A., 1987. Function box-counting and multiple elliptical dimension in rain. *Science*, 235, 1036-1038

Miró, J. J., Caselles, V. and Estrela, M. J., 2017. Multiple imputation of rainfall missing data in the Iberian Mediterranean context. *Atmospheric Research*, 197, 313 – 330. doi: https://doi.org/10.1016/j.atmosres.2017.07.016

Molnar, P. and Burlando, P., 2005. Preservation of rainfall properties in stochastic disaggregation by a simple random cascade model. *Atmospheric Research*, 137-151. doi: 10.1016/j.atmosres.2004.10.024

Mwale, F., Adeloye, A. and Rustum, R., 2012. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi – A self organizing map approach. *Physics and Chemistry of the Earth, Parts A/B/C*, 50-52, 34 – 43. doi: https://doi.org/10.1016/j.pce.2012.09.006

OTT, 2014. Operating instructions, present weather sensor ott parsivel 2. Kempten, Germany: OTT.

Over, T.M. and Gupta, V.K., 1996. A space-time theory of mesoscale rainfall using random cascades. *Journal of Geophysical Research-Atmospheres*, 101 (D21), 26319–26331. doi:10.1029/96JD02033

Oyler, J. W., *et al.*, 2015. Creating a topoclimatic daily air temperature dataset for the conterminous United States using homogenized station data and remotely sensed land skin temperature. *International Journal of Climatology*, 35, 2258-2279. doi: 10.1002/joc.4127

Rossman, L.A., 2010. *Storm water management model, user's manual. Version 5.0.* Cincinnati: U.S. Environmental Protection Agency, EPA/600/R-05/040.

Salvadori, G., Schertzer, D. and Lovejoy, S., 2000. Multifractal objective analysis: conditioning and interpolation. *Stoch. Environ. Res. and Risk Anal.*, 15, 261-283.

Schertzer, D. and Lovejoy, S., 1987. Physical modelling and analysis of rain and clouds by anisotropic scaling and multiplicative processes. *Journal of Geophysical Research*, 92 (D8), 9693–9714. doi:10.1029/JD092iD08p09693

Schertzer, D. and Tchiguirinskaia, I., 2020. A Century of Turbulent Cascades and the Emergence of Multifractal Operators. *Earth and Space Science*, 7, e2019EA000608. doi: 10.1029/2019EA000608

Schmitt, F. G., 2014. Continuous multifractal models with zero values: a continuous b-multifractal model. *Journal of Statistical Mechanics: Theory and Experiment*, 2014. http://stacks.iop.org/1742-5468/2014/i=2/a=P02008

Sivapragasam, C., *et al.*, 2015. Infilling of Rainfall Information Using Genetic Programming. *Aquatic Procedia*, 4, 1016 – 1022. doi: https://doi.org/10.1016/j.aqpro.2015.02.128

Tchiguirinskaia, I., *et al.*, 2004. Multiscaling geophysics and sustainable development. *Scales in Hydrology and Water Management*, IAHS Publ. 287, 2004, 113-136.

Velleux, M.L., England, J.F., and Julien, P.Y., 2011. *TREX watershed modelling framework user's manual: model theory and description.* Fort Collins: Department of Civil Engineering, Colorado State University.

Williams, D. A., *et al.*, 2018. A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data. *Environmental Modelling and Software*, 102, 172 – 184. doi: https://doi.org/10.1016/j.envsoft.2018.01.012

Figures:

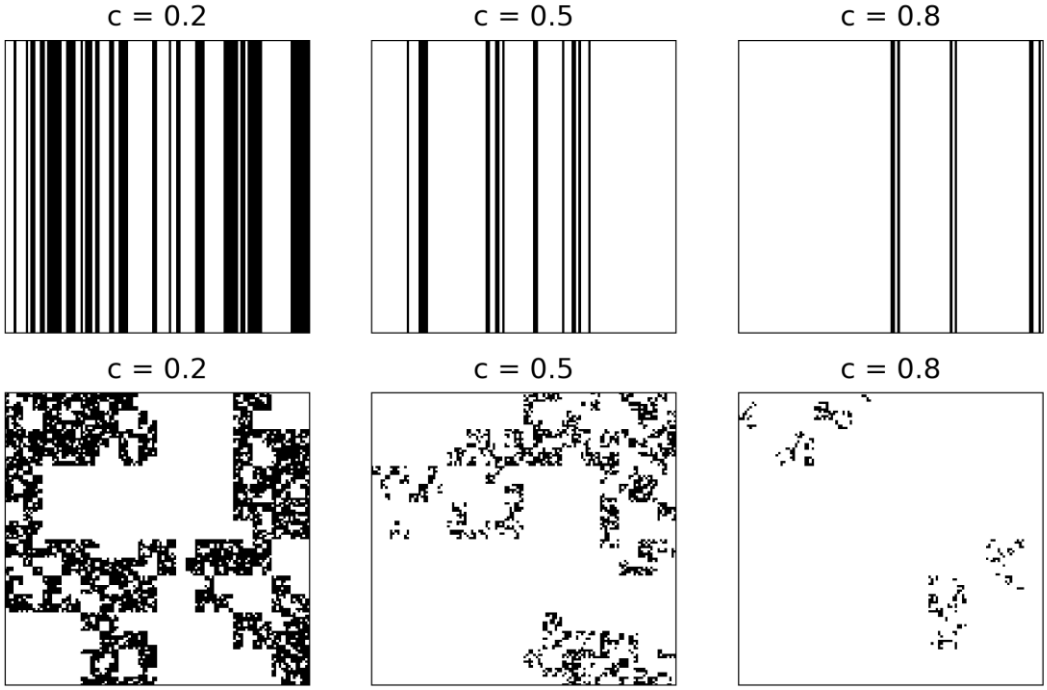

Figure 1: Examples in 1D (up) or 2D (bottom) of fields generated through the implementation of a β-model with various c}

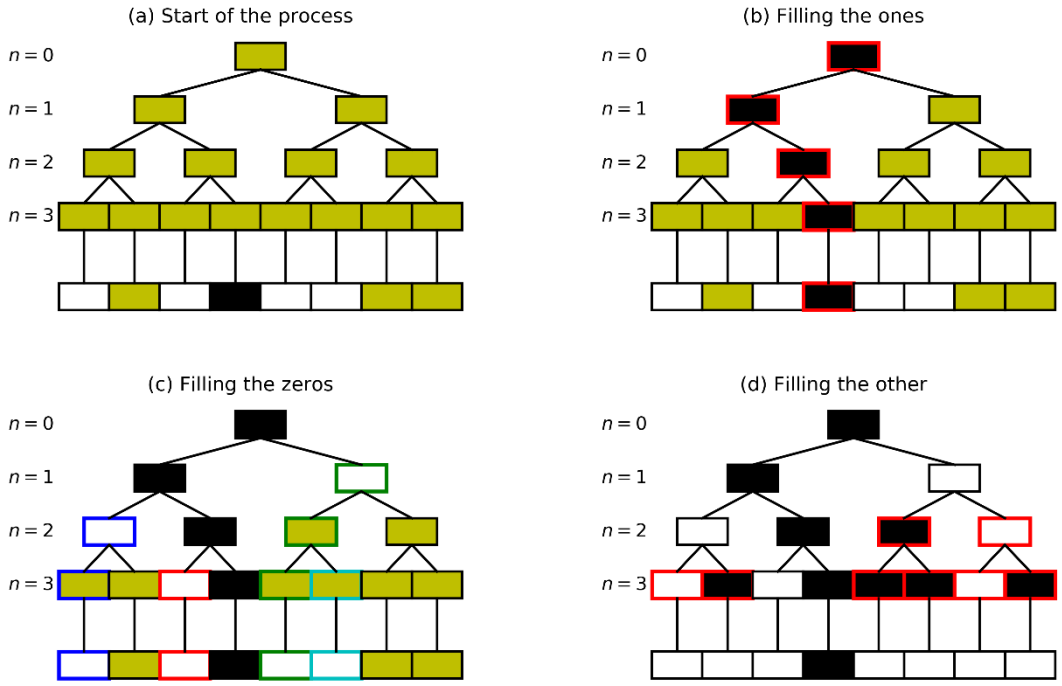

Figure 2: Illustration with a simple case of the successive steps of the conditional β-model. Ones are represented in black, zeros in white, and unknown/missing values in yellow. (a) Start of the process: some data is missing and all the underlying increments of the process are unknown. (b) Filling the ones: all the increments needed to retrieve the ones of the original field are set to one. (c) Filling the zeros: a limited number of increments ensuring the zeros of the original field are retrieved are set to zero. (d) The remaining unknown increment are randomly drawn by using the probability distribution of Eq. 5 More details (including the edge colour highlights can be found in the text.

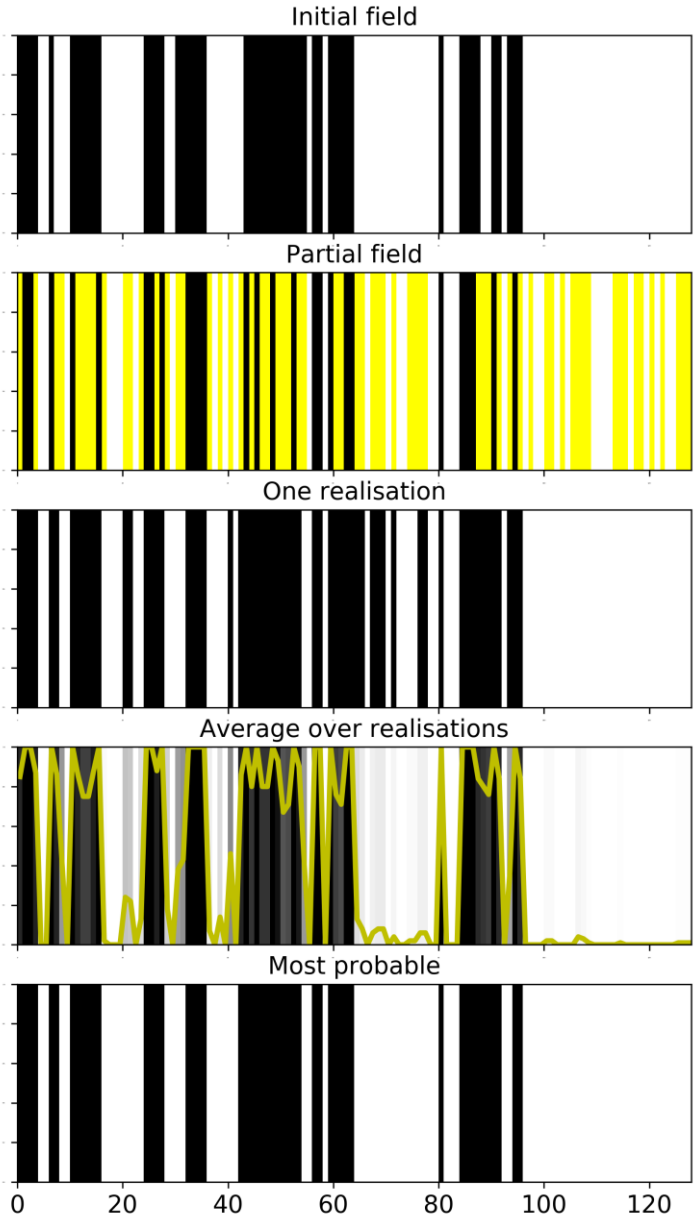

Figure 3: Outputs in 1D of the conditional β-model for an initial field (top) simulated with c=0.2. The content for each sub-figure is described through the title above them and more details can be found in the text. With regard to the 'partial field' sub-figure, the yellow time steps correspond to the missing data (i.e. the time steps whose content has been artificially removed). With regard to the 'Average over realisations' sub-figure, the colour scale goes for 0 (white) to 1 (black) with various levels of grey and the yellow line is simply a representation of the same information as a time series

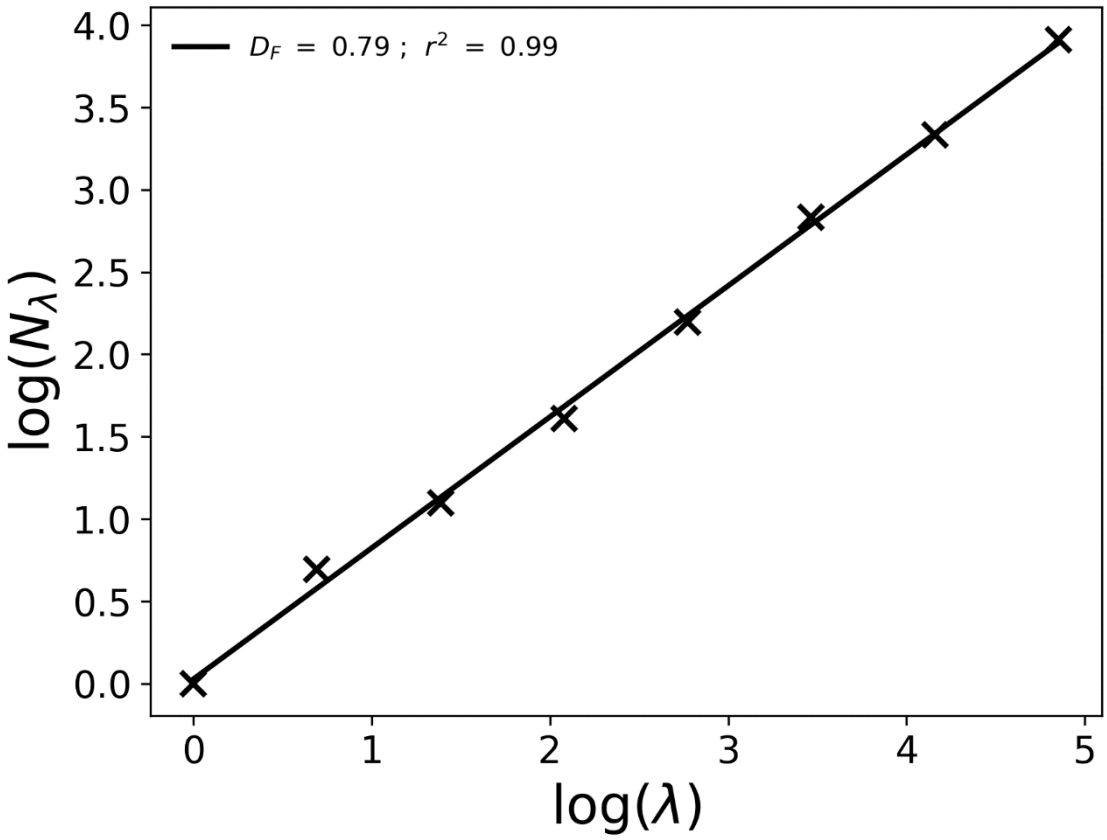

Figure 4: Computation of the fractal dimension (i.e. Eq. 2 in log-log) of the initial field displayed in Fig. 3 (top row)

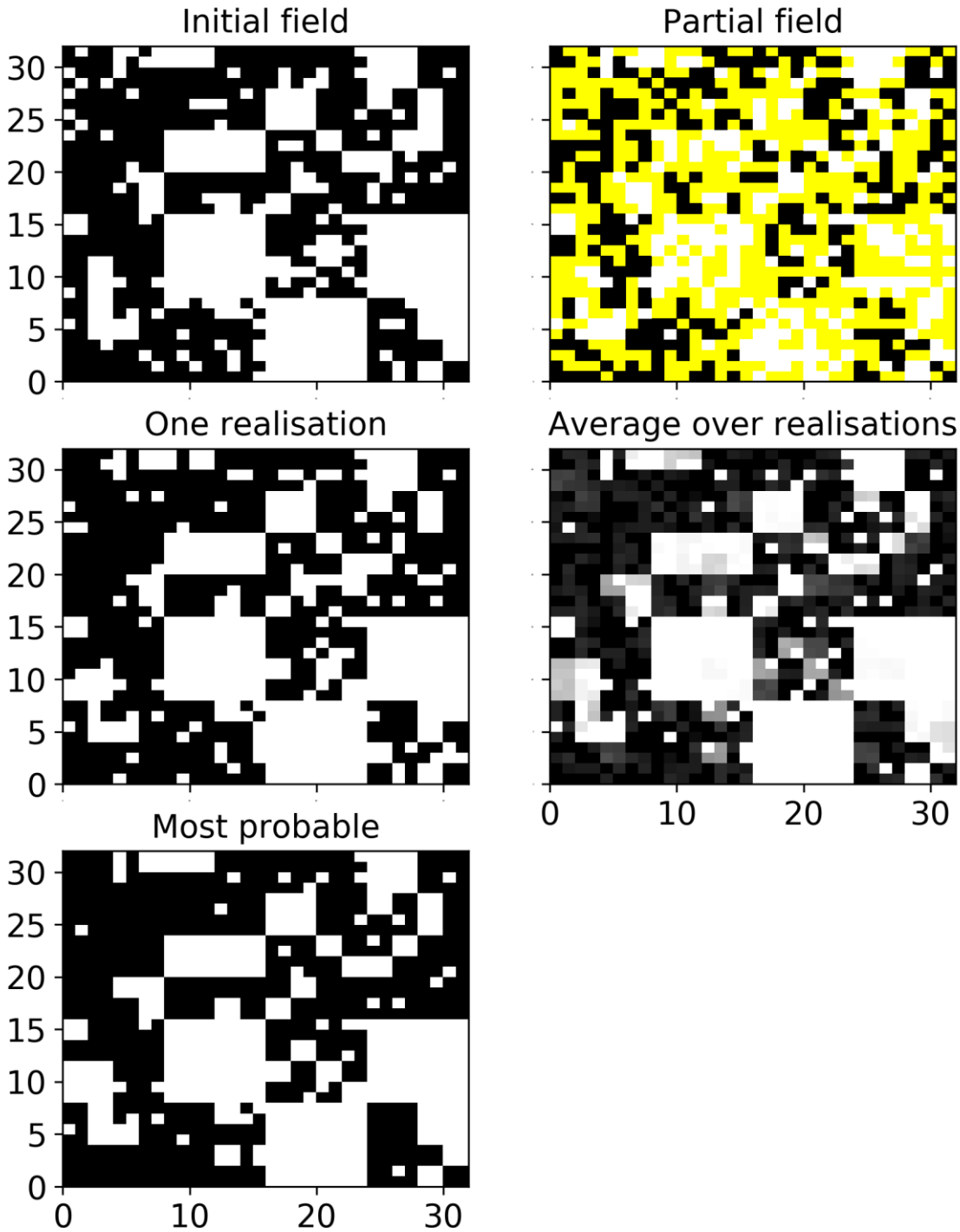

Figure 5: Outputs in 2D of the conditional β-model for an initial field simulated with c=0.2

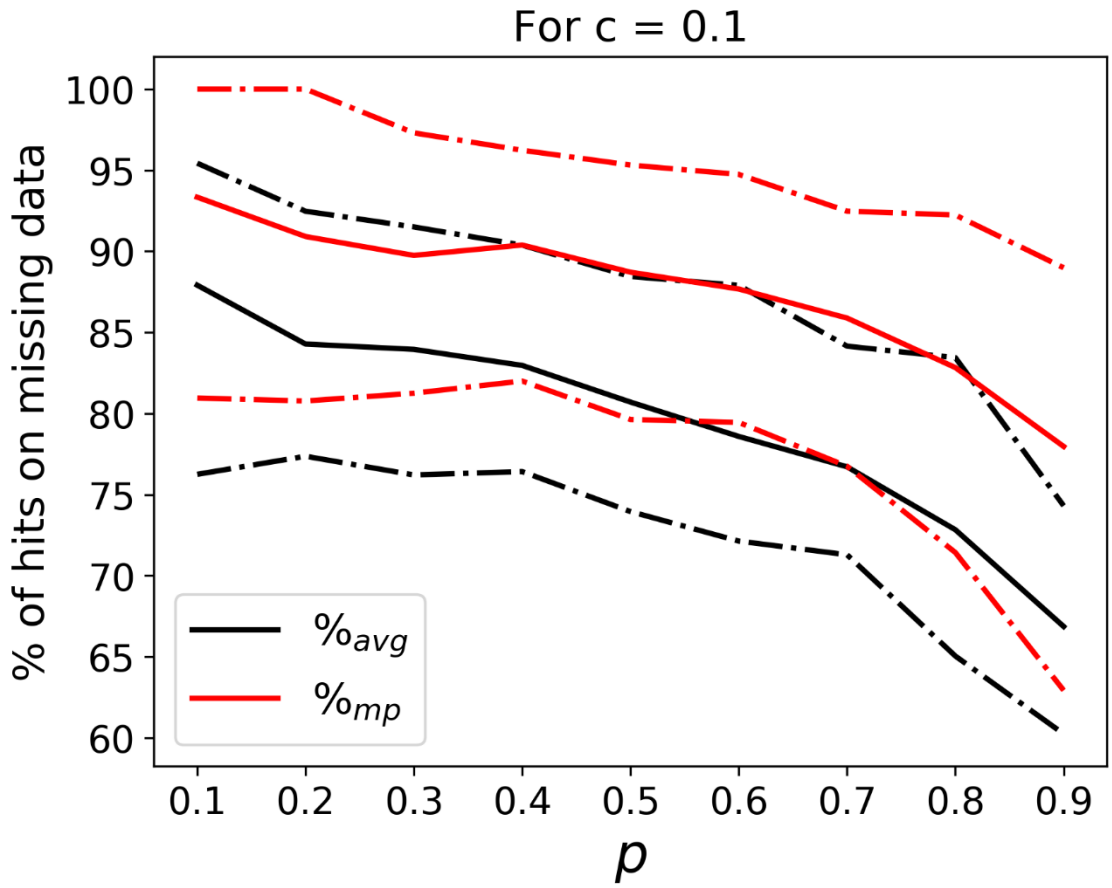

Figure 6: Percentage of hits on the missing data as a function of the proportion *p* of missing time steps inserted in the initial 1D series for *c*=0.1. Black lines are for the average over realisations while red ones are for the most probable fields. Solid lines are for the 50 % quantile over the 200 samples of initial fields while the dash lines correspond to the 10 and 90 % quantiles

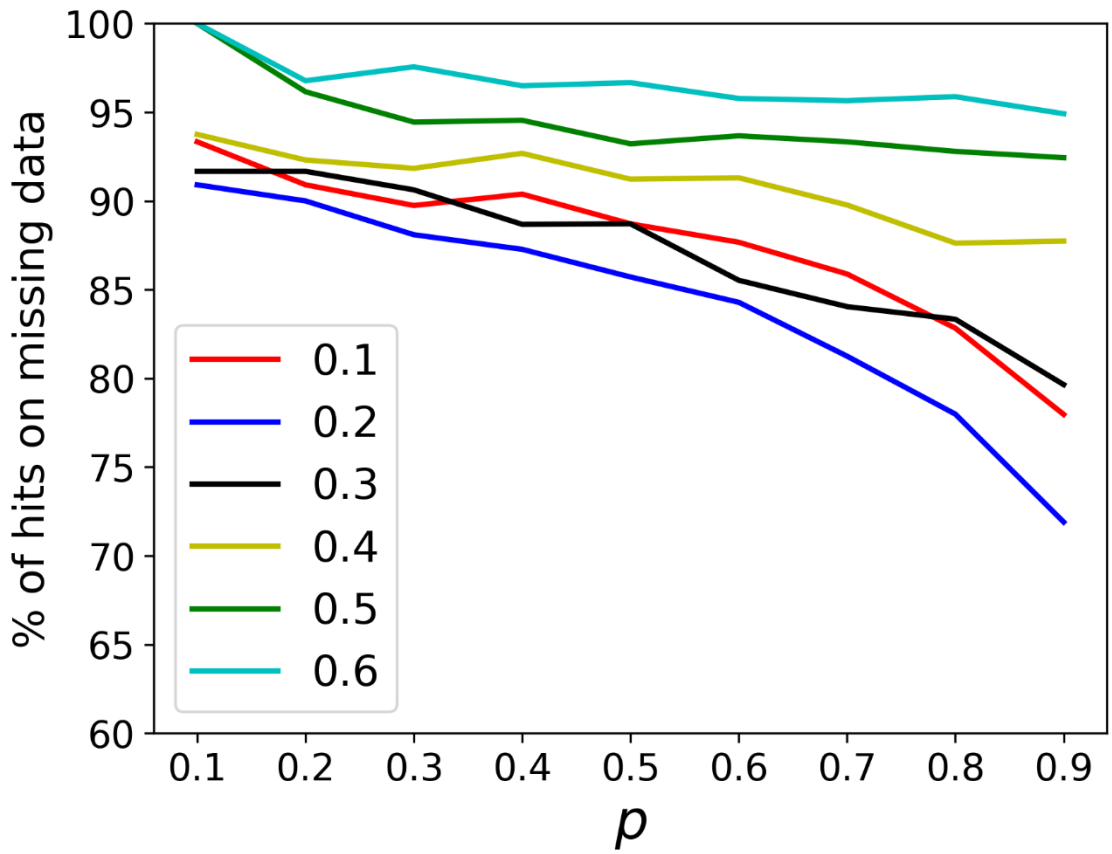

Figure 7: Percentage of hits on the missing data as a function of the proportion *p* of missing time steps inserted in the initial 1D series for various values of *c*. Only the 50% quantile over a set of 200 samples of initial fields through the 'most probable' approach are plotted (solid red line in Fig. 6)

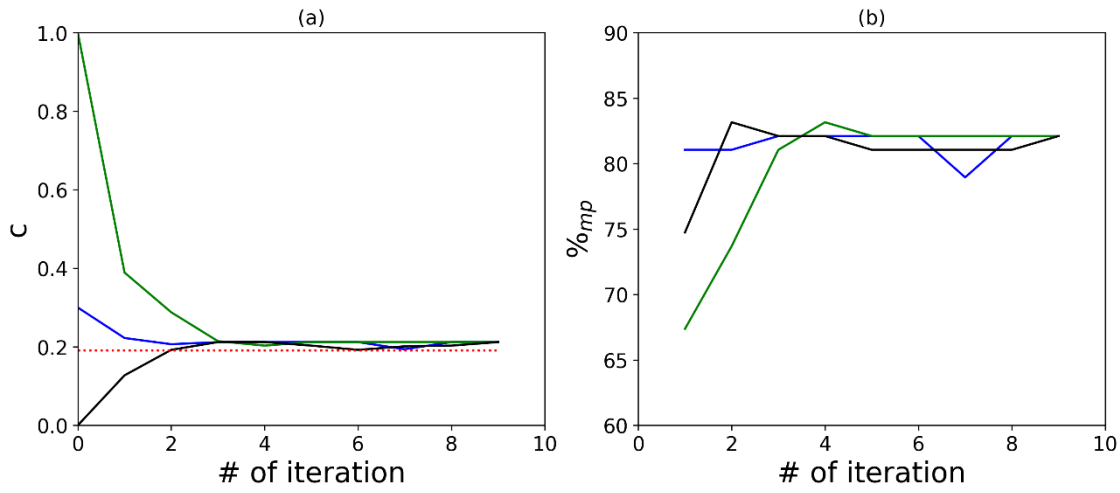

Figure 8: Evolution of c and $\%_{mp}$ as a function of the number of iterations in the algorithm to run the conditional β-model when c is unknown. The initial field was generated with c=0.2 and p=0.7. The algorithm was started with c equal to 0 (black), 0.3 (blue) and 1 (green).

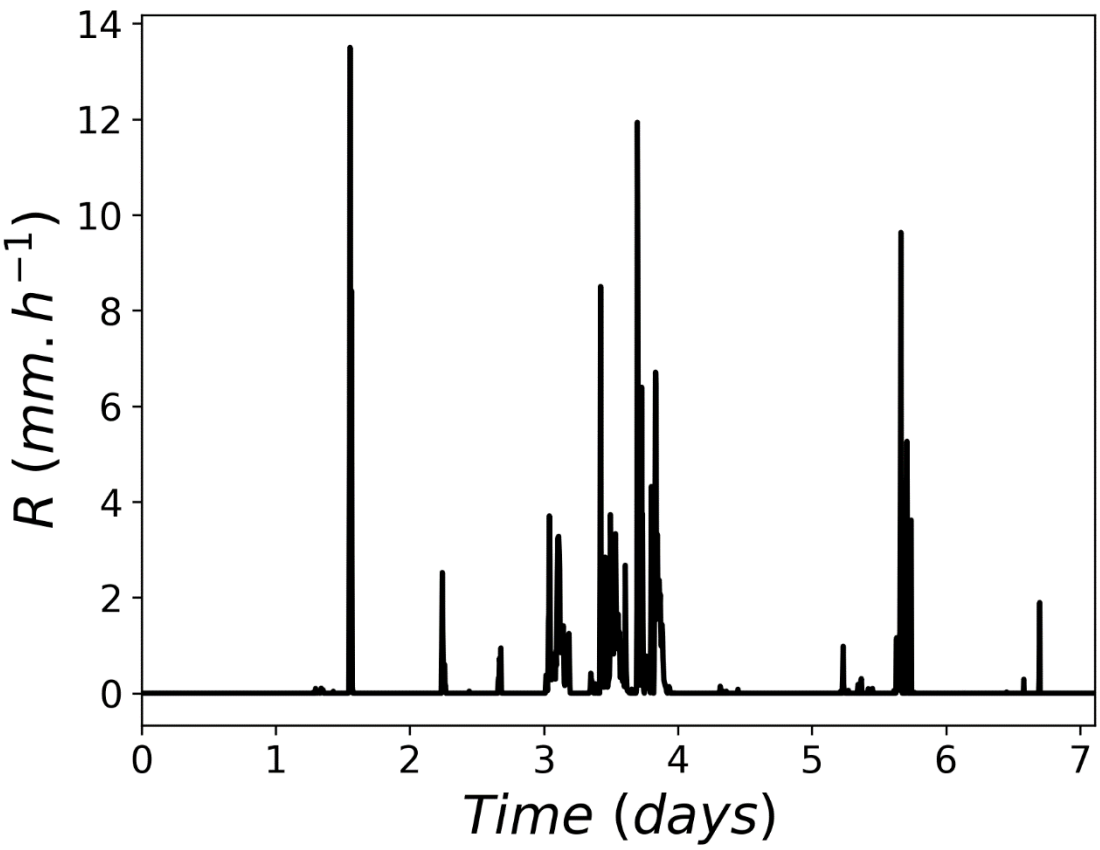

Figure 9: Temporal evolution of the rain rate corresponding to the 5 min time step studied rainfall series. Its length is of 2048 time steps, which corresponds to a duration of roughly 7.1 days. The series starts on 2019-06-02 00:00:00 (UTC)

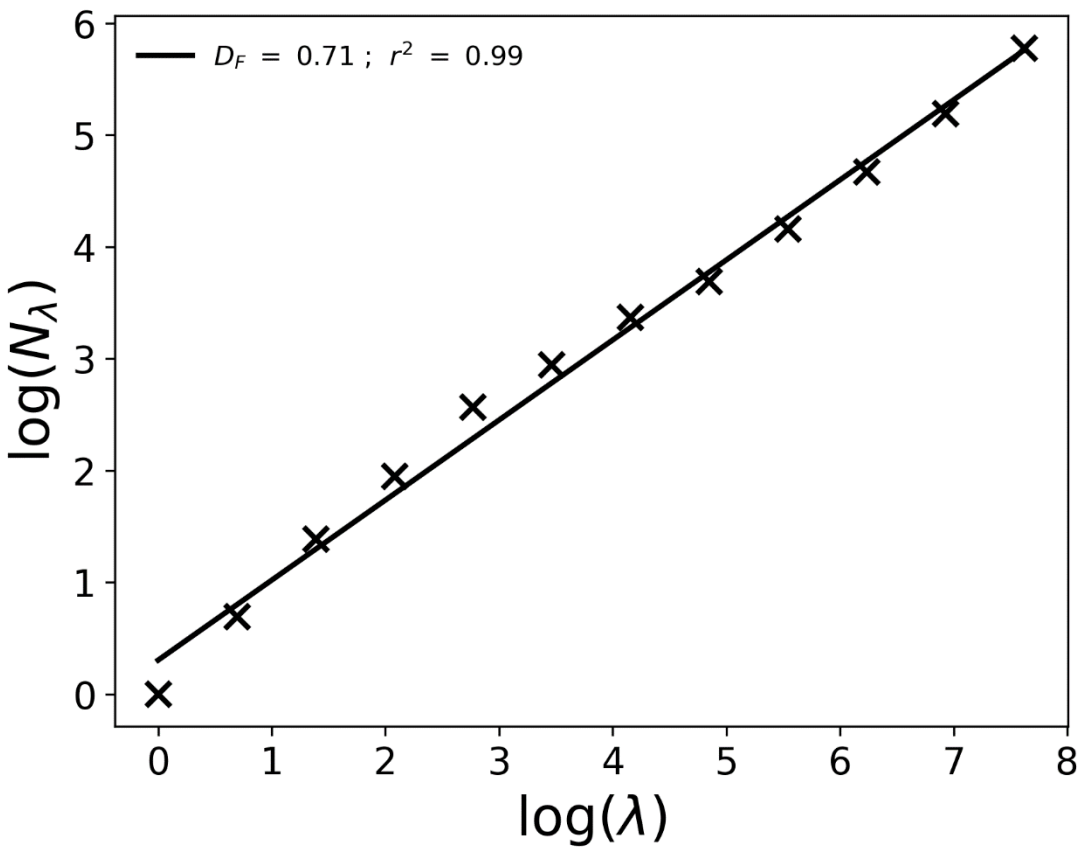

Figure 10: Computation of the fractal dimension (i.e. Eq. 2 in log-log) of the rainfall time series is displayed in Fig. 9.

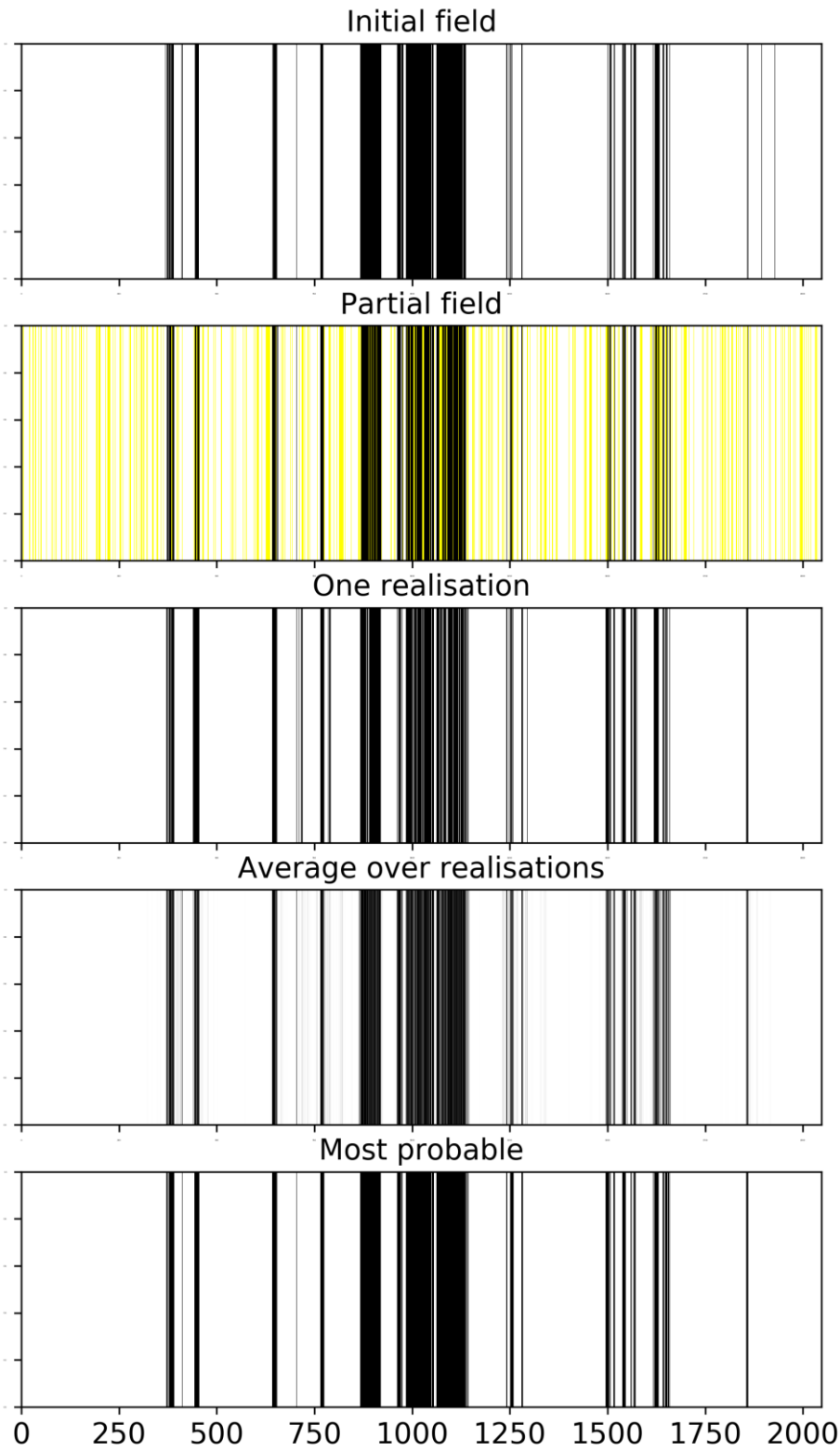

Figure 11: Outputs in 1D of the conditional β-model as in Fig. 3 with an 'initial' field consisting of the rainfall time series displayed in Fig. 9

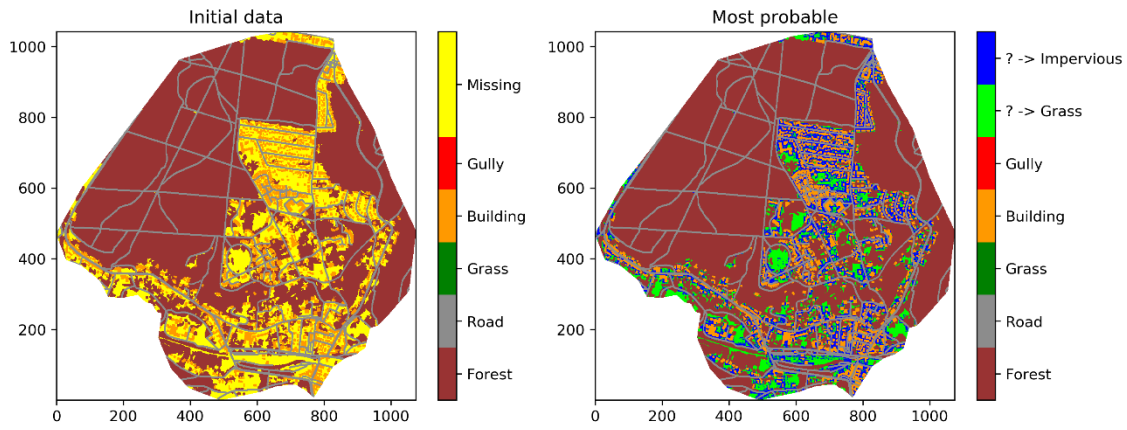

Figure 12: Representation of the Jouy-en-Josas catchment in Multi-Hydro with 2 m pixels. (a) 'Initial' data, hence with missing data. (b) 'Most probable' field where missing data has been replaced by either grass or an impervious area using the developed conditional β-model.

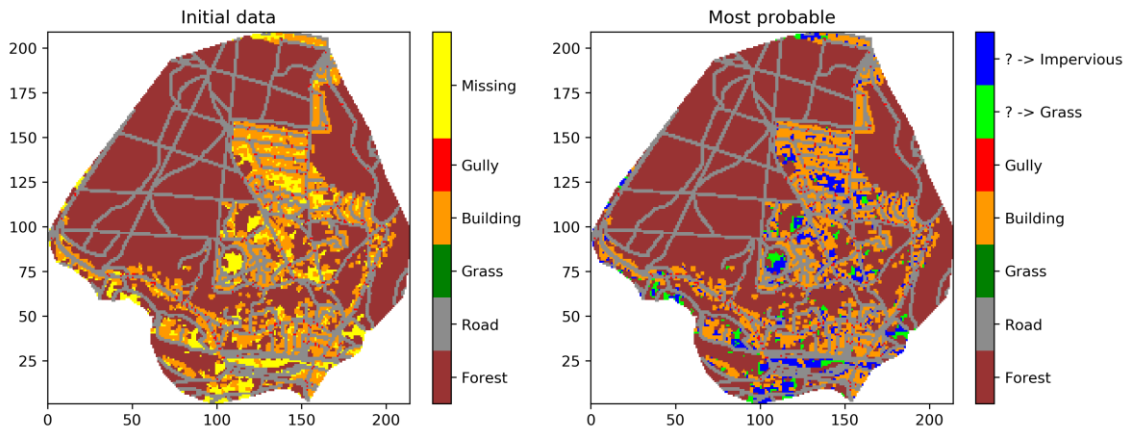

Figure 13: Same as in Fig. 12 size 10 m.

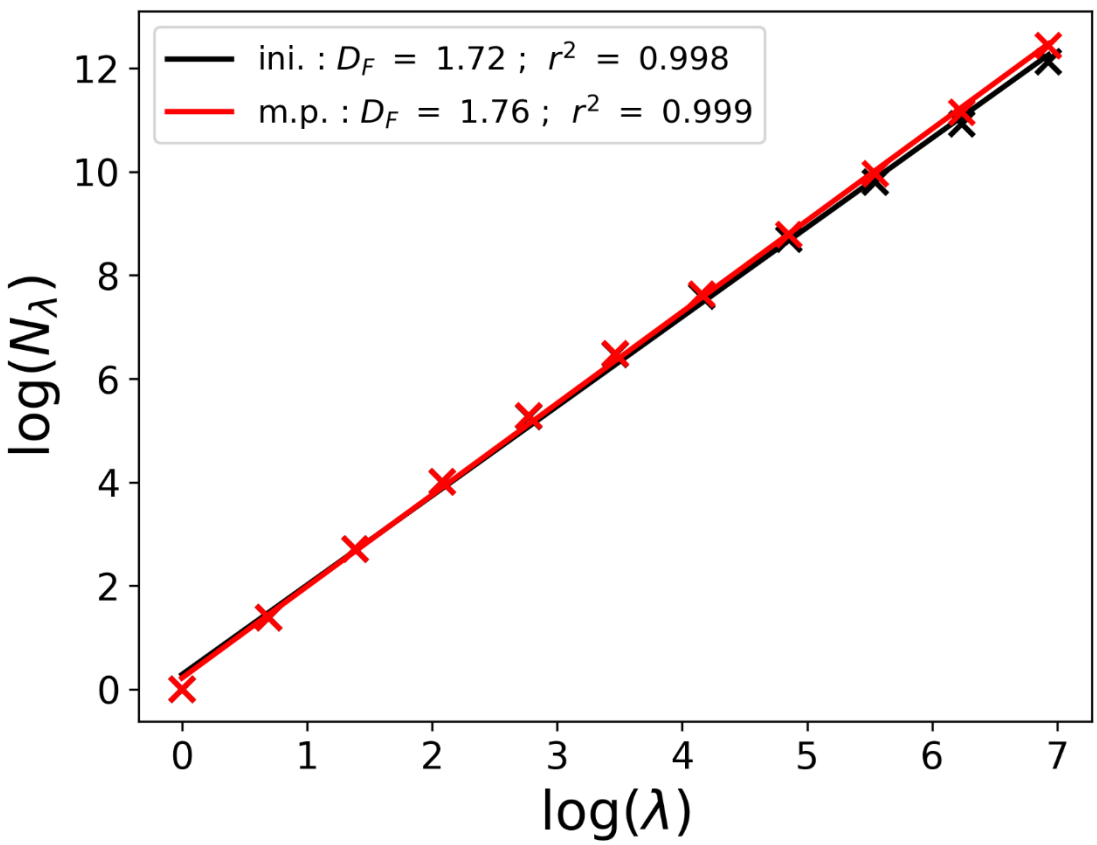

Figure 14: Computation of the fractal dimension (i.e. Eq. 2 in log-log) of the geometrical set made of the impervious areas of the studied catchment at with 2 m pixels (i.e. the field displayed in Fig. 12).

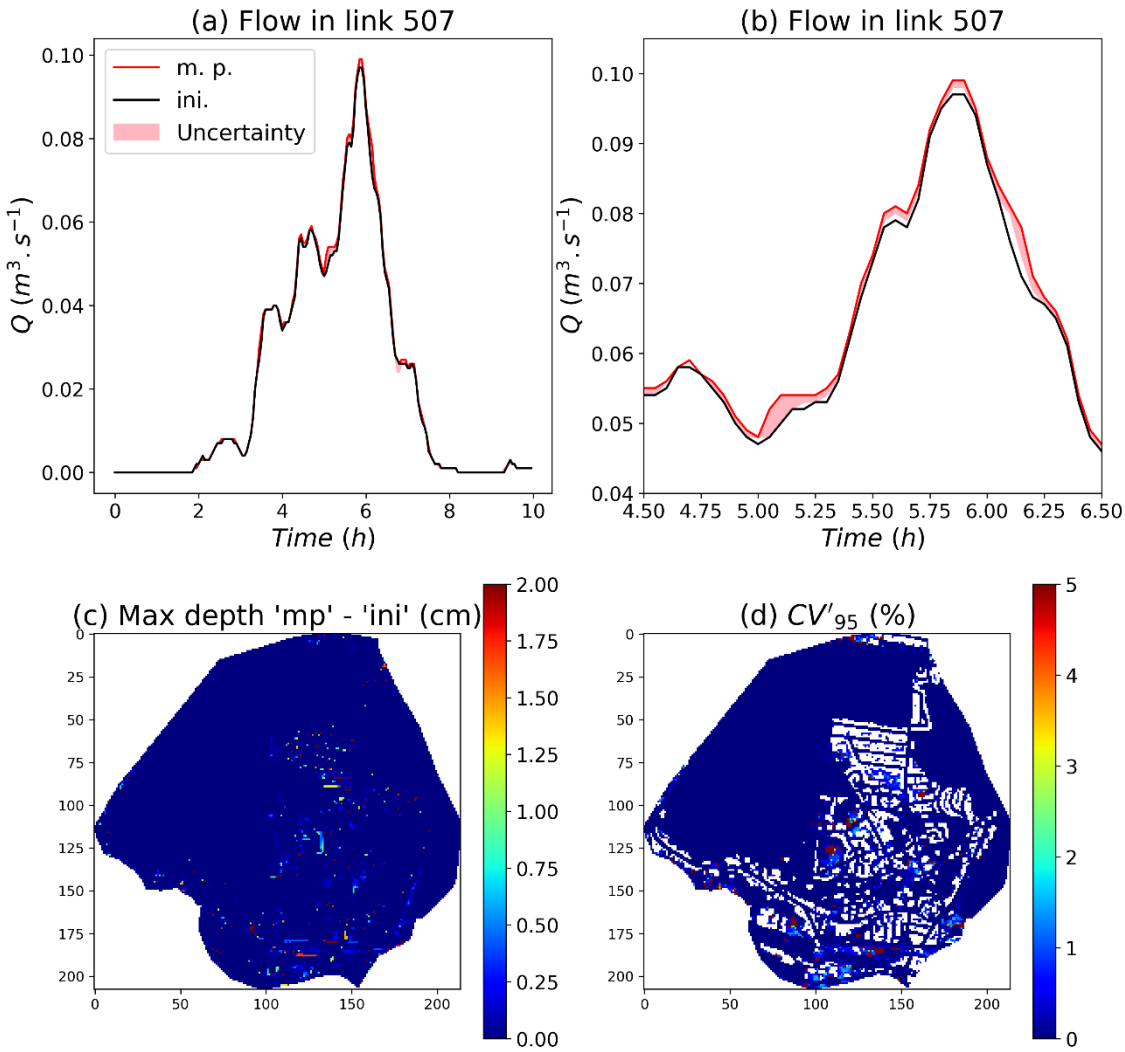

Figure 15: Outcome of the hydrodynamic simulations carried out with the Multi-Hydro model. 'Initial' land use data with the missing data taken as 'grass' is used as input along with the 100 realisations of realistic ways of filling the 5.6% of missing data, and the 'most probable' field. (a) Simulated flows at link #507. Time is indicated since the beginning of the simulation. (b) Zoom of (a) near the peak flow. (c) Map of the difference of simulated maximum water depth between simulations carried out with the 'most probable' field and the 'initial' one. (d) Map of the CV'_{95} coefficient for the maximum water depth.