A CO_2GeoNet Initiative

ENOS D 2.2

Report on Uncertainty Quantification of capacity estimates: coherent framework & applications

Date	June 2019
Author(s)	Jean Charles Manceau* BRGM, Marie Rousseau BRGM, Jeremy Rohmer BRGM
Corresponding author (*)	jc.manceau@brgm.fr

Number of appendices	1
Customer	EC
Project name	ENOS
Project number	GA No 653718

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 653718

Contents

1 Executive summary

One of the objectives of Work Package 2 of the ENOS project ("Ensuring storage capacities and cost-effective site characterisation") is to **quantify the reliability of storage capacities estimates particularly for onshore deep saline aquifers**. An accurate CO_2 storage capacity estimation is an absolute precondition for the success of storage operations. As such, the **development of approaches and methods for establishing capacity estimations** participates to increase the visibility of administrations who need to assess the feasibility of storage operations in their countries, and the confidence of operators, emitters and investors, who need elements to analyse the economic potential of a given CO_2 storage project.

Any efficient management of storage site (e.g., site selection, injection operations, etc.) should then rely on a "good picture" of what is unknown when assessing the storage capacity with a given dynamic model: this is the purpose of uncertainty assessment. Within this context, the objective of the present deliverable is twofold.

1. **Objective 1:** importance ranking (sensitivity analysis, SA). The first objective is to develop a framework for getting a better insight in the role played by these different forms of uncertainties (ENOS Task 2.1.2, step 1);

2. **Objective 2:** reliability quantification (uncertainty quantification, UQ). The second objective is to develop methods for propagating the afore-mentioned uncertainties by estimating the quantiles P5, P50, P95 of the capacity estimates (ENOS Task 2.1.2, step 2). These measures are standards for reliability of reserves in the domain of O&G typically.

Performing robust importance ranking and uncertainty quantification is however a challenging task and classically requires most of the time a **large amount of simulations** (at least > 500), which might not be feasible with dynamic models designed to estimate reliable capacity estimations. To deal with this issue, the proposed approaches for

- SA by benchmarking a **large number of procedures** for importance ranking adapted to relatively low number of simulations and inputs represented by modelling scenarios.
- UQ by **developing a strategy** using a proxy (low fidelity model) established from a detailed model and by testing it to predict over time the quantiles of the capacity indicators.

Though project delays did not let us apply these procedures on an ENOS test case, it should be underlined that the developments are generic and can be applied to any dynamic models dedicated to capacity estimates. Besides, we believe that the considered test case (of onshore injection CO_2 in the lower Triassic sandstone formation based on a potential project in the Paris basin (France) as described by Manceau and Rohmer (2016)) is sufficiently realistic (in terms of quality, quantity and types of data in early stages of CO_2 storage test site) to consider the described recommendations potentially meaningful for future (or on-going) onshore CO_2 storage projects.

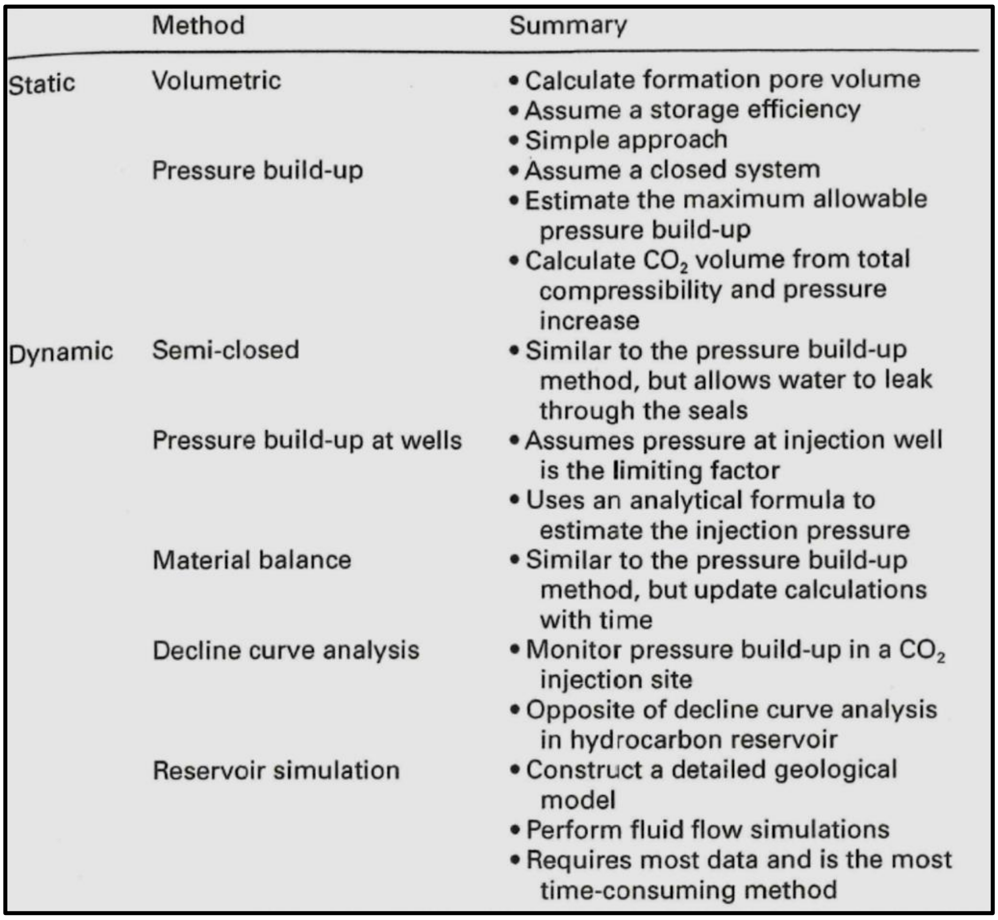
2 Introduction: context & objectives

One of the objectives of Work Package 2 of the ENOS project ("Ensuring storage capacities and cost-effective site characterisation") is to **quantify the reliability of storage capacities estimates particularly for onshore deep saline aquifers**. An accurate CO_2 storage capacity estimation is an absolute precondition for the success of storage operations. As such, the **development of approaches and methods for establishing capacity estimations** participates to increase the visibility of administrations who need to assess the feasibility of storage operations in their countries, and the confidence of operators, emitters and investors, who need elements to analyse the economic potential of a given CO_2 storage project.

CO_2 storage capacity estimation has therefore been the subject of numerous studies. Among the most cited works, one can list Bachu et al. (2007), who described the challenges to be faced when assessing capacity and presented theoretical estimation methodology adapted to coal beds, oil and gas reservoir, and deep saline aquifers; Zhou et al. (2008), who developed a simple analytical storage capacity estimator, dedicated to the quick assessment of CO_2 storage capacity in closed and semi-closed saline aquifer systems; and Kopp et al. (2009a, 2009b), who presented a method for quantifying storage capacity applied to several reservoir setups.

Pickup (2013) provides an exhaustive overview of the different methods for assessing CO_2 storage capacity (see Table 1). Two main types of approaches can be found:

- **static approaches**, i.e. independent of time, estimate storage capacity either by calculating the pore volume available for storage, or by calculating the pressure build(up (compressibility methods). These methods are relatively simple to apply, and are therefore well adapted to preliminary stages of a project for site screening for instance, or for a global estimation at la large to very large scale.
- **dynamic approaches**, that accounts for transient process in the storage capacity estimation. The capacity estimates therefore varies with time (Pickup, 2013). These approaches can be set-up through the use of analytical or numerical dynamic modelling (reservoir modelling). Generally analytical methods consider simple geometry, homogeneity, whereas numerical ones may be based on a realistic geological model and account for some heterogeneities; in addition, analytical solutions often neglect more physical processes than numerical calculations. However, analytical solutions are much faster than numerical ones and allows a quicker assessment of capacity estimates.



	Method	Summary
Static	Volumetric	• Calculate formation pore volume • Assume a storage efficiency • Simple approach
	Pressure build-up	• Assume a closed system • Estimate the maximum allowable pressure build-up • Calculate CO_2 volume from total compressibility and pressure increase
Dynamic	Semi-closed	• Similar to the pressure build-up method, but allows water to leak through the seals
	Pressure build-up at wells	• Assumes pressure at injection well is the limiting factor • Uses an analytical formula to estimate the injection pressure
	Material balance	• Similar to the pressure build-up method, but update calculations with time
	Decline curve analysis	• Monitor pressure build-up in a CO_2 injection site • Opposite of decline curve analysis in hydrocarbon reservoir
	Reservoir simulation	• Construct a detailed geological model • Perform fluid flow simulations • Requires most data and is the most time-consuming method

Table 1. Description of existing methods for assessing CO_2 storage capacity (after Pickup, 2013)

The present study being dedicated **to site-specific reliable** storage capacities estimates, the presented worked is focused **numerical simulations–based methodologies**. According to Pickup (2013), reservoir simulation modelling is indeed “essential for making a more informed estimate of CO_2 storage capacity at a chosen storage site”. However, the reliability of storage capacity estimation depends on the dynamic model built and on the geological model on which it is based.

The use of dynamic simulations for computing a realistic capacity estimation has to face several major difficulties, which are further discussed below:

- **The procedure and choice of the indicator(s) for assessing dynamic capacities**;
- **The management of the uncertainties** stemming from the dynamic modelling process;
- **The computational time required to run** refined dynamic models to ensure the most reliable capacity estimation.

The procedure and choice of the indicator(s) for assessing dynamic capacities.

The capacity is theoretically defined **as a storable quantity** but to our knowledge, **neither a real definition of dynamic capacities which is by definition dependant on time (cf. Pickup, 2013), nor a modelling procedure to assess such capacities have been established**.

For instance, Jin et al. (2012) used as capacity estimates, the so-called **storage efficiency factor** (defined as the ratio of pore volume containing CO_2 to the total pore volume). Numerical simulations are used to estimate such efficiency value at different sites in UK using the following constraint : injection of 15 Mt/a, with fifteen wells with **a target injection**

rate of 1 Mt/a potentially **reduced in case a maximum pressure build-up is reached**, continuation of the injection either during 15 years (alternative 1) or **until the total field injection rate is reduced to one quarter of its initial value** (alternative 2).

In a recent paper published by Zulqarnain et al. (2017), capacity estimates are computed on a specific site in Louisiana (USA); dynamic simulations are used to compute capacity estimates in Mt, which are also translated in storage efficiency factor. Dynamic simulations are run at **constant rate**, and they are **stopped when either a critical bottom-hole pressure is reached or when the CO_2 plume reaches a given boundary**; the **cumulated injected mass after the injection has stopped** is considered as the storage capacity.

The discussion of the most appropriate capacity estimator is not in the scope of the study presented in this report; nevertheless **we propose new indicators with the purpose of launching the discussion on this topic**. In our study (see section 4.4), we propose to run dynamic simulations **at a constant bottom-hole pressure** (with an overpressure equal to 10 % of the initial pore pressure) **during a given duration** (10 years). The chosen indicators for estimating storage capacities are **times series of 1) the cumulated injected mass (in Mt), and 2) the footprint (spread) of the gaseous CO_2 plume (in km^2)**.

Management of the uncertainties stemming from the dynamic modelling process

The uncertainties coming from the dynamic modelling procedure **may take several forms** (see for instance comprehensive reviews by Beven, 2016; Caers & Scheidt 2011).

The first category stems from **the difficulties in estimating the input parameters (in a broad sense) of models/analysis due to the limited number, poor representativeness (caused by time, space and financial limitations), and imprecision of observations/data**. Some examples are the intrinsic permeability of a given rock layer, the initial pressure/temperature conditions at depth, etc.

The second category is **closely related to the process of model setting up**. Pappenberger and Beven (2006) give the following definition of models "[…] an abstract construct to represent a system for the purposes of reproducing, simplifying, analysing, or understanding it. Any model is based on a perceptual model (summary of our (personal) perceptions on how a system responds), which gets translated into a conceptual model (mathematical description and implementation as a procedural model (computer code))". This means that models are **necessarily simplified representations of the phenomena and necessarily based on assumptions**. Uncertainty can then appear in the **structure/form of the model**, which depends on the choice of input variables, dependencies, processes and so forth regarded as relevant and prominent for their purpose in the model. Yet, in some cases, a set of different models (e.g. differing in their structure and input variables) are either considered equally adequate (e.g., they equally fit the observations), or they are associated with different confidence levels. In the present study, **we consider situations where the modellers have difficulties in unambiguously choosing among a set of plausible modelling assumptions (like a physical law or a boundary condition)**. In practice, this can be treated by assigning to the set of plausible modelling choices/assumptions **an indicator variable**, which takes up discrete values; for instance Manceau and Rohmer (2016) defined an indicator variable taking up the value {1, 2,…,10} to account for a set of 10 equally plausible modelling relative permeability law). We acknowledge that more complicated situations may exist where the number and nature of the input parameters may also change form one modelling scenario to another. This situation is out of the scope of the present study and the interested reader can refer to the recent study by Dai et al. (2017) for new methodological developments on that matter.

Any efficient management of storage site (e.g., site selection, injection operations, etc.) should then rely on a "good picture" of what is unknown when assessing the storage capacity with a given dynamic model: this is the purpose of uncertainty assessment. Within this context, the objective of the present deliverable is twofold.

1. Objective 1: importance ranking (sensitivity analysis, SA). The first objective is to develop a framework for getting a better insight in the **role played by these different forms of uncertainties** (ENOS Task 2.1.2, step 1);

2. Objective 2: reliability quantification (uncertainty quantification, UQ). The second objective is to develop methods for **propagating the afore-mentioned uncertainties by estimating the quantiles P5, P50, P95 of the capacity estimates** (ENOS Task 2.1.2, step 2). These measures are standards for reliability of reserves in the domain of O&G typically[1]. ().

Dynamic model computational time

Performing robust importance ranking and uncertainty quantification is a challenging task and classically **requires most of the time a large amount of simulations** (at least > 500), which **might not be feasible with dynamic models designed to estimate reliable capacity estimations**. To deal with this issue, the proposed approaches for both SA and UQ either are **adapted to relatively low number of simulations** (see objective 1: importance ranking), or **make use of a proxy (low fidelity model)established from a detailed model** (see objective 2: reliability quantification).

Nota bene

All the developments on importance ranking and uncertainty quantification in the context of long-running dynamic modelling **are supported with the real case of onshore injection CO_2 in the lower Triassic sandstone formation based on a potential project in the Paris basin (France) as described by Manceau and Rohmer (2016)**.

According to the description of work of the ENOS project, the application of the proposed approaches should have been performed on *a priori* selected sites: Hontomin, Spain (fractured carbonate aquifer), and GeoEnergy Test Bed, UK (faulted Permo-Triassic sandstone aquifer).

Because of delay in the dynamic model provision, it has been decided in 2017 to apply the SA procedure on a dynamic model developed in a previous project (FP7 ULTimateCO_2); **this dynamic model was used for assessing the long term fate of injected CO_2 in deep saline aquifers** (Manceau & Rohmer, 2016); the SA was therefore carried out to this end.

In 2018, no additional models was available, and therefore, it has been decided to also apply the UQ procedure on the FP7 ULTimateCO_2 dynamic model but to modify the initial objective of the model: **the dynamic model was thus run with a capacity estimation aim, in order to better respond to the ENOS WP2 objective.**

Though the delays did not let us apply the procedures on an ENOS test case, it should be underlined that the developments described for importance ranking and uncertainty quantification are **generic** and can be applied to **any dynamic models dedicated to capacity estimates**. Besides, we believe that **the considered test case** (of onshore injection CO_2 in the lower Triassic sandstone formation based on a potential project in the Paris basin (France) as described by Manceau and Rohmer, 2016) is **sufficiently realistic** (in terms of quality, quantity and types of data in early stages of CO_2 storage test site) **to consider the described recommendations potentially meaningful for future** (or on-going) **onshore CO_2 storage projects**.

The report is organized as follows. In a first section (Sect. 3), we describe the methods used to fulfil objective 1 & 2; the case study is presented in Sect.4. In Sect. 5 and Sect. 6, the methods for SA and UQ are applied on the case study.

[1] Px is a statistical confidence level for an estimate, when probabilistic Monte Carlo type evaluations are adopted: Px is defined as x% of estimates exceed the Px estimate. P95 and P05 are low and high estimates respectively.

3 Methods

Dealing with uncertainties require responding to two different objectives as explained in the introduction section: 1) assessing the importance of the different uncertainties existing throughout the flow modelling process, and 2) assessing how those uncertainties impact a given output of interest (e.g. capacity estimates). This section describes the approaches followed for dealing with these SA and UQ issues.

3.1 Objective 1: importance ranking

To get a better insight in the role played by the different forms of uncertainties encountered within the dynamic modelling workflow, we can apply global sensitivity analysis (denoted GSA), which is a powerful setting to provide valuable information by addressing the following questions:

- what sources of uncertainty contribute the most to the uncertainties in the flow simulation results?
- How to rank these sources of uncertainties?
- How to set priorities for future investigations?

Methods for SA have extensively been used for unravelling the role played by uncertain parameters. A recent study by Tong and Mukerji (2017) tackles this issue for parameter uncertainty in basin and petroleum system modelling.

However, to analyse sensitivity to both parameter and model uncertainties, there is to the authors' best knowledge often no consensus on which methods are best applicable regarding the specificities of the situation. The aim objective of the present study regarding importance ranking is to test the feasibility (and potentially extend the functionality) of the available methods/approaches for dealing with global sensitivity analysis with respect to parameter and model uncertainties.

We restrict the analysis to global techniques (contrary to local method using parametric studies like One-At-A-Time approach, see a discussion by Saltelli and Annoni, 2010), which introduce the less hypotheses regarding the input-output relationship (i.e. model free). Though implementing GSA has been eased over the last years by the increasing computer power, applying it to the large-scale models characterized by a large computation time cost (computation time CPU cost > several hours), still faces difficulties. Therefore, we further restrict the comparison analysis to methods, which do not require too many numerical simulations (of the order of a few hundred). Finally, we select methods which have been applied on real cases in the literature. Four methods are tested (see also the summary in Table 2):

- M-VBSA: combination of variance-based GSA (Saltelli et al. 2008) and metamodeling techniques adapted to situations using continuous and categorical variables (Storlie et al. 2013) Application to CO_2 sequestration modeling can be found in Manceau and Rohmer (2016);
- RF: a machine learning approach based on the random forest technique (Wei et al. 2015: Sect. 6) with application in environmental modelling;
- DGSA: distance-based generalized sensitivity analysis developed Fenwick et al. 2014 and further extended by Park et al. 2016 based on the Regionalized sensitivity analysis RSA method (Spear and Hornberger 1980); Application to reservoir modelling can be found in Tong and Mukerji (2017);
- PAWN: a density-based GSA (aka moment-independent) developed by Pianosi & Wagener (2015) with application in environmental modelling.

Method	Description	Description of variable importance measure	Treatment of discrete input variables
M-VBSA	Variance-based Sensitivity Analysis	The main effect is the first order Sobol' index and is interpreted as the contribution of the input variable to the output variance; The total effect	Metamodel adapted to both handle continuous and discrete input

	combined with Metamodel	is interpreted as the global contribution including the main effect and the interactions with the other parameters	variables should be used.
RF	Permutation-based Variable Importance Measure PVIM derived from the Random Forest method	Decrease of predictability (measured by the mean squared error), which is related to the the non-standardized Sobol' total effect index. Significance is estimated using Altmann et al.'s approach.	The method relies on binary partitioning (tree-based approach. It is by construction adapted to handle discrete input variables
RSA	Regional sensitivity analysis based on the CDFs on input parameters	The sensitivity index quantifies the differences between input empirical CDFs of two different groups of simulations defined according to the outputs values.	No adaptation needed for handling categorical inputs.
DGSA	Sensitivity analysis based on the CDFs on input parameters after classification of the output parameter into multiple sets.	The sensitivity index quantifies the difference between input empirical CDFs of multiple groups of simulations defined according to the outputs values. Conditional indices quantifies multiple way interaction among parameters.	No adaptation needed for handling categorical inputs.
PAWN	Moment-independent GSA	The sensitivity index associated to an input parameter quantifies the difference among the CDFs of the output for different fixed values of the input parameter.	Since it is based on the analysis of conditional CDF i.e. CDF provided that the considered input parameter is fixed at a given value, it is well adapted to both continuous and discrete variables.

Table 2. Characteristics of the SA methods

These methods are detailed in the appendices section.

3.2 Objective 2: reliability quantification

The aim of this study regarding UQ is to assess the reliability of the capacity estimates calculated with a dynamic model by propagating the uncertainties encountered within the flow modelling workflow and by assessing different quantiles of the capacity estimates. Classically, quantiles (more generally probability distributions) are also used for evaluating the uncertainties of the potentially recoverable volumes of oil (see for instance the Petroleum Resources Management System[2]); a low, best, and high estimates are defined in the following way:

[2] https://www.spe.org/en/industry/petroleum-resources-management-system-2018/

- Low estimate or P90: “There should be at least a 90% probability (P90) that the quantities actually recovered will equal or exceed the low estimate”.
- Best estimate or P50: “There should be at least a 50% probability (P50) that the quantities actually recovered will equal or exceed the best estimate”.
- High estimate or P10: “There should be at least a 10% probability (P10) that the quantities actually recovered will equal or exceed the high estimate”.

In our study, we propose to evaluate the P5, P50 and P95 on the capacity estimates. Though assessing such quantiles with flow simulations has been eased over the last years by the increasing computer power, performing such estimates using large-scale models (i.e., characterized by a large number of mesh cells – typically over several 100,000s and associated to large computation time > several hours), still faces difficulties.

To alleviate this computational burden, different approaches have been proposed in the literature. A first option relies on the statistical analysis of existing databases of pre-calculated reservoir simulation results. This approach based on response surface techniques (also known as surrogates or metamodels) like Gaussian Processes (e.g., Hamdi et al., 2017), Spline-based techniques (Manceau and Rohmer, 2016), polynomial chaos expansions (e.g., Köppel et al., 2018), etc. have been successfully applied in different contexts. Yet, their performance remains, by construction, related to the size of the training database.

An alternative is to rely on proxy flow models, i.e. low-fidelity models, which are less accurate, but computationally cheaper. Different proxies exist: (1) coarsening the model grid mesh potentially combined with upscaling techniques, which computes equivalent petrophysical properties at a coarser scale than the initial detailed model (Durlofsky,2005); (2) using simplified physics such as streamline simulation (3) approximating dynamic reservoir response based on connectivity (Renard and Allard, 2013) or using Fast Marching methods (Hovadik and Larue, 2011). The bottleneck of the proxy-based approach is related to the simplifications made for building the proxy, which might result in a biased uncertainty quantification. To avoid this problem, the proxies are typically employed only to identify a representative subset of realizations for which the exact model is solved as originally proposed by within the distance-based ranking procedure by Scheidt and Caers (2009), and recently extended to all class of proxy models by Bardy et al. (2019), to compute the quantile of interest (i.e. P05, P50 and P95).

To circumvent the problem of proxy-based biased responses, a possible option is to learn the error model as proposed e.g. by Josset and Lunati (2013), i.e. to build statistical error models that describe the discrepancy between exact and approximate responses. Once the error model is constructed (and validated), it can be used to correct the approximate responses and predict the responses expected from the exact model for all realizations. Recent developments have been proposed for handling functional quantities, i.e. time-dependent variable like time-dependent breakthrough curves (Josset et al., 2015), but, to our best knowledge, very few feasibility assessment are available in the literature and more particularly for the most common situations where the proxies correspond to coarse-scaled flow simulations.

In the present work, we build on the class of error-based methods dedicated to coarse scale reservoir models. Our main contribution is to perform an extensive benchmark, on a realistic case study, different approaches for building dynamic error model, i.e. for time-dependent capacity estimators (see Sect. 4.4).

4 Case study

In this section, we first describe the dynamic model used to simulate the injection of CO_2 (Sect. 4.1). A full detailed description can be found in Manceau and Rohmer (2016). Second, we present the input parameters, the different modelling assumptions, as well as the uncertainties associated to those data and modelling options (Sect. 4.2). Last, the application of the methods described in Sect. 3 on that case study is explained (Sect. 4.3).

4.1 Model set-up

The case study is the one described by Manceau and Rohmer (2016). We consider the hypothetical storage of CO_2 in the lower Triassic sandstone formation based on a potential project in the Paris basin (France). The site has been classically characterized and we summarize in this section the data available for establishing a dynamic model of the CO_2 storage evolution over time. Note that the level of information is representative of a project in preliminary stage of development.

The studied zone is located in the easternmost part of the Paris basin with a chosen injection point located at a depth of approximately 1,000 m and at more than 50 km from the outcrops. A 3D geological model of the storage formation was built (see Figure 1, [Top]) and the corresponding Petrel grid was converted into a grid to be run with TOUGH2 flow simulator (Pruess et al., 1999). 2 different grids have been constructed : one "coarse mesh" composed of 4,030 cells, and one "refined mesh" composed of . 19,671 cells. The "refined mesh" is refined in both vertical and horizontal directions compared to the "coarse mesh" (see Figure 1, [Left]).

The flow simulator includes the equation of state package ECO_2N (Pruess and Spycher 2007), accounting for the properties of brine-CO_2 mixture at classical pressure and temperature of CO_2 storage operations. Dissolution of CO_2 in the aqueous phase is treated by means of local equilibrium solubility, i.e. by considering an instantaneous phase partitioning of water and carbon dioxide between the liquid and gas-like phases. A hysteresis module is used (Doughty 2009) to model residual trapping and to take into account the associated hysteresis effects onto relative permeability and capillary pressure. It follows Land's residual trapping model (land 1968) as well as hysteretic characteristic functions derived from van Genuchten's capillary pressure function (van Genuchten 1980) and based on Lenhard and Parker's relative permeability to brine and CO_2 (Lenhard and Parker 1987).

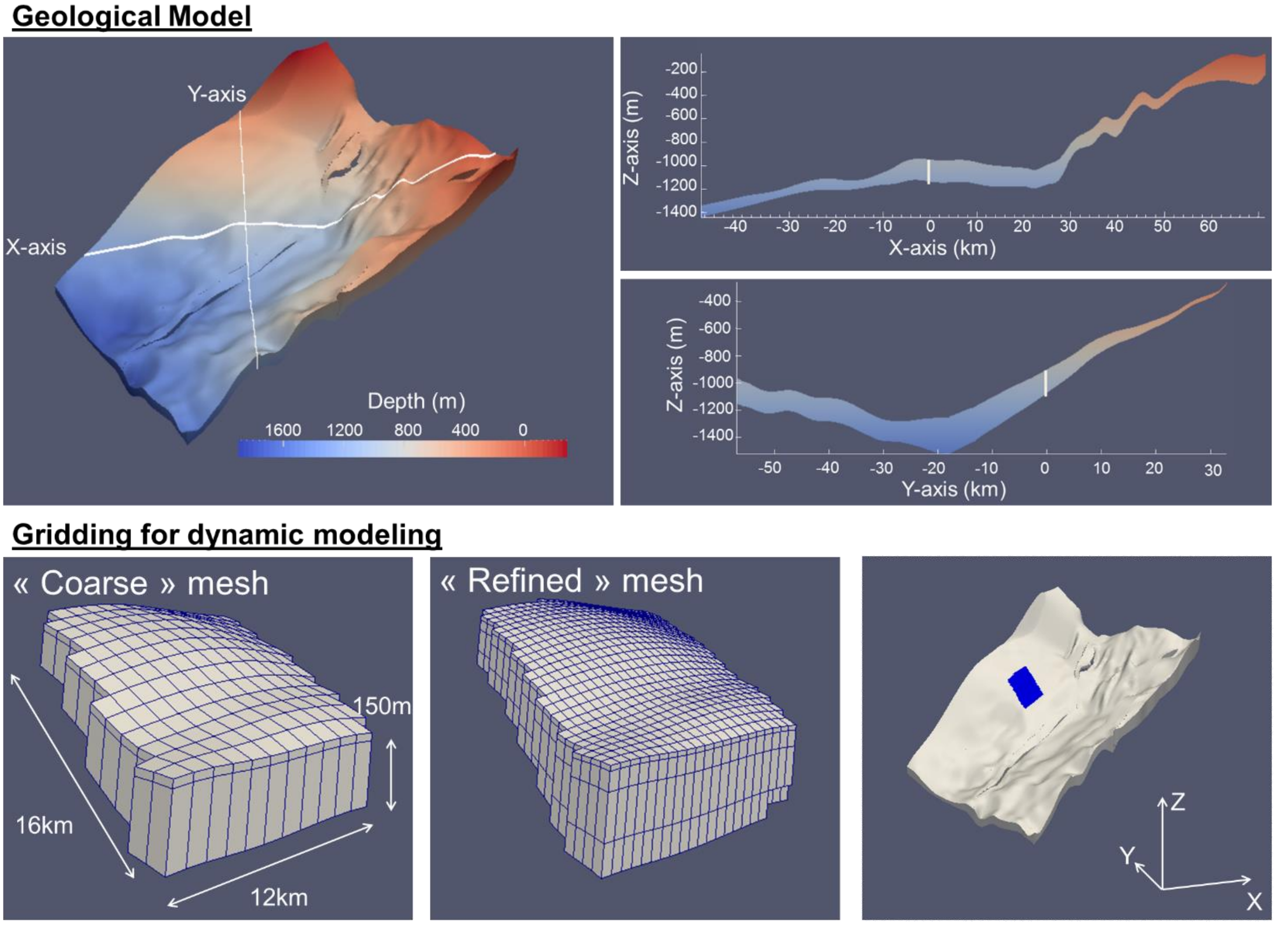


Figure 1. Model set-up (modified after Manceau and Rohmer, 2016) : [Top] - Three views of the geological model of the Triassic formation considered. Left: 3D view. The intersection between the X-axis and Y-axis corresponds to the injection zone; Right: X-axis (West-East) cross section (top) and Y-axis (North-South) cross section (bottom). [Bottom] - The two different meshes used in this study (only a small part of each entire mesh is shown on the figure and corresponds to the blue area on the top-right picture of the total model)

4.2 Input Parameter and Modelling assumptions

The main input parameters are the absolute flow properties of the reservoir formation, which are considered homogeneous at the scale of the reservoir. The associated parameter uncertainty is modelled by probability distributions derived from the available permeability and porosity measurements derived from well data (see Figure 2 [Top]).

Contrary to porosity and permeability, no available information has been found on the multiphase flow parameters of this sandstone formation. Multiphase flow parameters (relative permeability law kr) are not known for our case study and ten existing databases (with preprocessing as described by Manceau and Rohmer, 2016), were used (see Figure 2 [Bottom]). The uncertainty related to the choice of the kr law is considered to be of model type.

Data regarding capillary pressure for CO_2/brine system are even more limited than relative permeability ones. To assess the importance of this parameter in our study given this lack of data, two extreme modelling scenarios are considered: 1. a case of significantly low capillary pressure (equivalent to no-capillary pressure) and 2. a case of strong capillary pressure (named high capillary pressure in this paper). No change (hysteresis) between drainage and imbibition was considered.

The permeability vertical anisotropy is unknown in the reservoir. Therefore, we considered three different anisotropy scenarios: an isotropic case (ani=kv/kh=1), a laminated case (ani=0.1) and one scenario in-between (ani=0.5).

The hydrogeology of the region has been studied on a large scale. However, in the specific area of injection the local pressure gradient is not precisely known. Two extreme scenarios are therefore considered, one with initial hydrostatic conditions and another one with a strong hydraulic gradient in the injection area. As an extremely strong gradient, we use 0.01 m/m (equivalent to ca. 100 kPa/km), which is in the high range of the review of hydraulic gradient in geologic basins done by Larkin (2010).

The range of salinity being relatively narrow, we consider that the uncertainties on this parameter are low and use a mean value corresponding to 20 $g.l^{-1}$. In a similar way, the CO_2 migration after injection is limited and therefore the temperature variations are expected to be low in the area of interest: the temperature is therefore considered constant with a value of 45°C (estimation of the temperature at 1,000 m).

Table 1 summarizes the hypotheses and choices made regarding the uncertainties accounted and presented in this section.

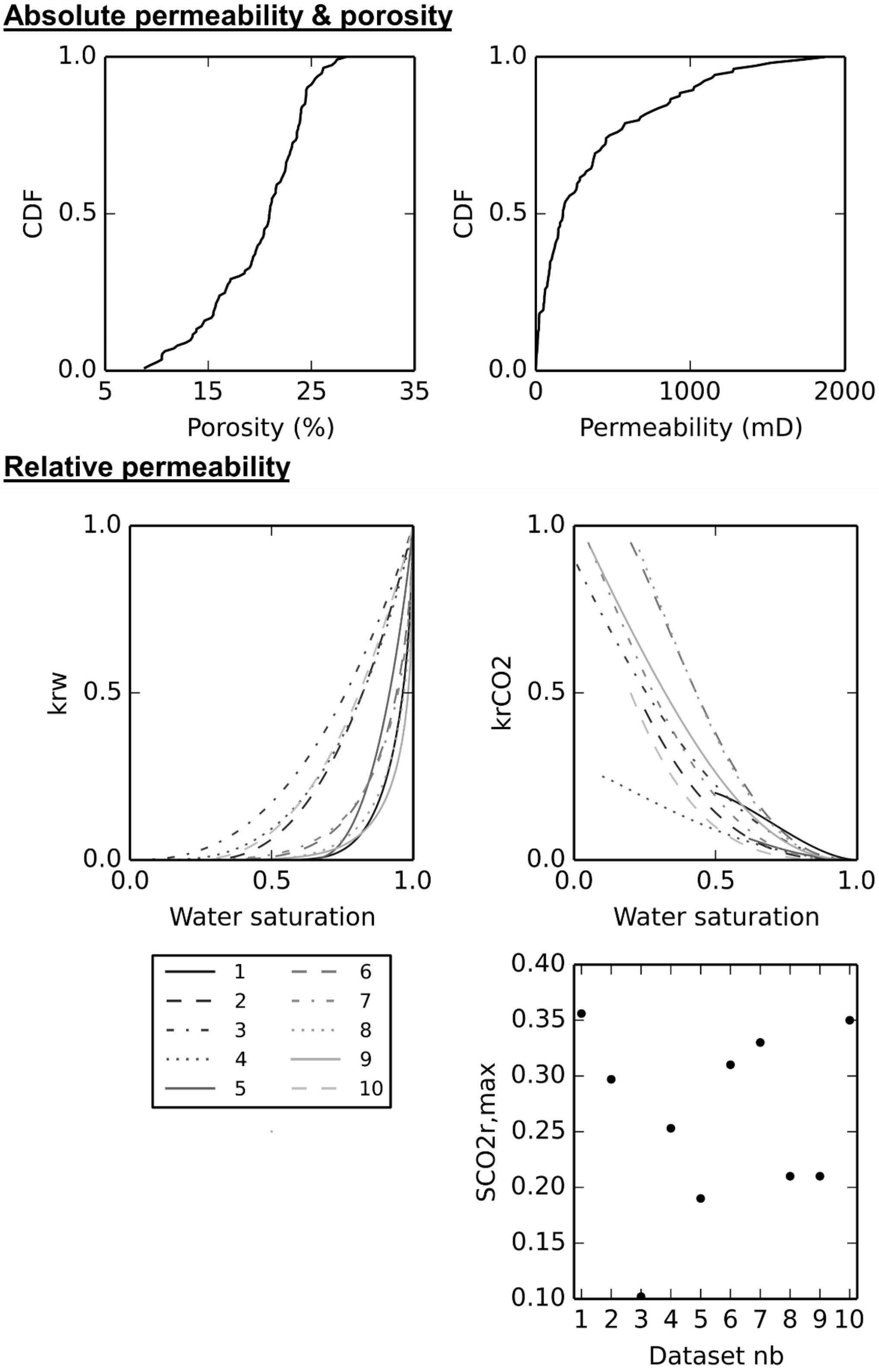


Figure 2. [Top] Cumulative probability distribution function considered in this study for porosity (left) and absolute permeability (right). [Bottom] (top) Water -krw- and CO_2 -krCO_2- relative permeability dataset considered for the drainage process as a function of the water saturation, and maximum residual CO_2 saturation for each dataset

Parameter	Symbol	Uncertainty type	Representation
Porosity	phi	Parametric	Empirical probability distribution
Permeability	kabs	Parametric	Probability density
Permeability anisotropy	ani	Model	3 scenarios (k_v/k_h = 0.1; 0.5 and 1)
Regional hydraulic gradient	vreg	Model	2 scenarios (hydrostatic and 0.01 m/m)
Relative permeability	kr	Model	10 scenarios (10 different relative permeability datasets)
Capillary pressure	pc	Model	2 scenarios (no-capillary pressure and “strong” capillary pressure)

Table 3. Uncertainty classification of the case study

4.3 Application of importance ranking (SA) techniques to the case study

The different sensitivity analysis methods described in Sect. 3.1 have been applied to the dynamic model detailed above.

The importance of model and parameter uncertainties (see Sect 4.2) have been assessed on post injection trapping of mobile CO_2. More precisely, the outcome of interest (for importance ranking) is the amount of mobile CO_2 after 150 years the injection stops, considering a constant injection rate of 1 Mt/a during 30 years (output of interest called OP1 in Manceau and Rohmer, 2016). Only the coarse mesh was used for the importance ranking analysis.

Up to 1,000 samples of the input parameters were randomly generated. For each of these input configurations, a 250 years (including the 30 year-CO_2 injection) 3D flow simulation was run to compute the output of interest.

We first analyse the results of SA considering in turn each method (“within-method” analysis). We pay a special attention to investigating the impact of the number of simulation results N (ranging from 50 to 1,000) as well as the robustness to the parametrisation of each method (i.e. the sensitivity of the SA results to the parameters’ values of each method). Second, we compare the results between the different methods (“between-method” analysis) and formulate recommendations for practical implementation.

4.4 Application of reliability quantification (UQ) techniques to the case study

The UQ method described in Sect (3.2) was implemented to the case study, and we recall that the final objective was to evaluate different quantiles of capacity estimators to quantify their reliability.

In practice, the dynamic model was used in the following way: CO_2 injection is performed at a constant bottom-hole pressure, during an arbitrary duration of 10 years. The overpressure considered to allow CO_2 to be injected in the reservoir has been taken at a value of 10% compared to the initial pore pressure. This overpressure ensures an acceptable pressure build-up (see for instance Jin et al., 2010).

Storage capacity can be constrained notably by the pressure build-up in the formation (ensured by the constant bottom-hole pressure), and by a too large extension of the CO_2 plume implying potentially larger risks (presence of leaking well, conflicts of use, etc.). Therefore two capacity estimators has been computed from the simulations:

1) the cumulated injected mass (in Mt),

2) the footprint (spread) of the gaseous CO_2 plume (in km^2), i.e. the surface of the mobile free-phase CO_2 plume directly in contact with the caprock formation (more precisely the projection of this surface on the horizontal plane is computed[3]).

It is important to note that the evolution of these two capacity estimators is evaluated: therefore, the UQ is made on these estimators time-series (10 years long time series with a time step of 1 year).

The UQ methodology followed in this study is based on a "proxy" model, whose responses are compared to an "exact" model. For the application to the case study, the "exact" model is the dynamic model constructed with the "refined" mesh, while the proxy model is the dynamic model constructed with the "coarse" mesh. 540 samples of the input parameters were randomly generated to run the proxy model. For a comparison between the "exact" quantiles and the quantiles obtained with the UQ methodology, the exact model was also run with the same input data sample despite the large computational time required to run those simulations.

Two examples are given to illustrate the evaluated estimators. They are coming from two simulations run with the refined mesh, leading to similar results in terms of cumulated injected mass (after10 years) but different results in terms of footprint. The input parameters for these two simulations are given in Table 4. The storage capacity indicators are provided in Figure 3, and a visualisation of the CO_2 plume after 10 years of injection is shown at Figure 4.

Parameter	Simulation A	Simulation B
Porosity	*0.09*	*0.22*
Permeability	*$1.14e^{-12}m^2$*	*$6.7e^{-13}$*
Permeability anisotropy	*0.1*	*0.1*
Regional hydraulic gradient	*hydrostatic*	*strong*
Relative permeability	*scenario 2*	*scenario 10*
Capillary pressure	*no Pc*	*strong*

Table 4. Input parametrers for simulations A et B shown as examples

[3] The surface output parameters are calculated by considering the mobile gas when the mobile free-phase CO_2 saturation is superior to 0.001.

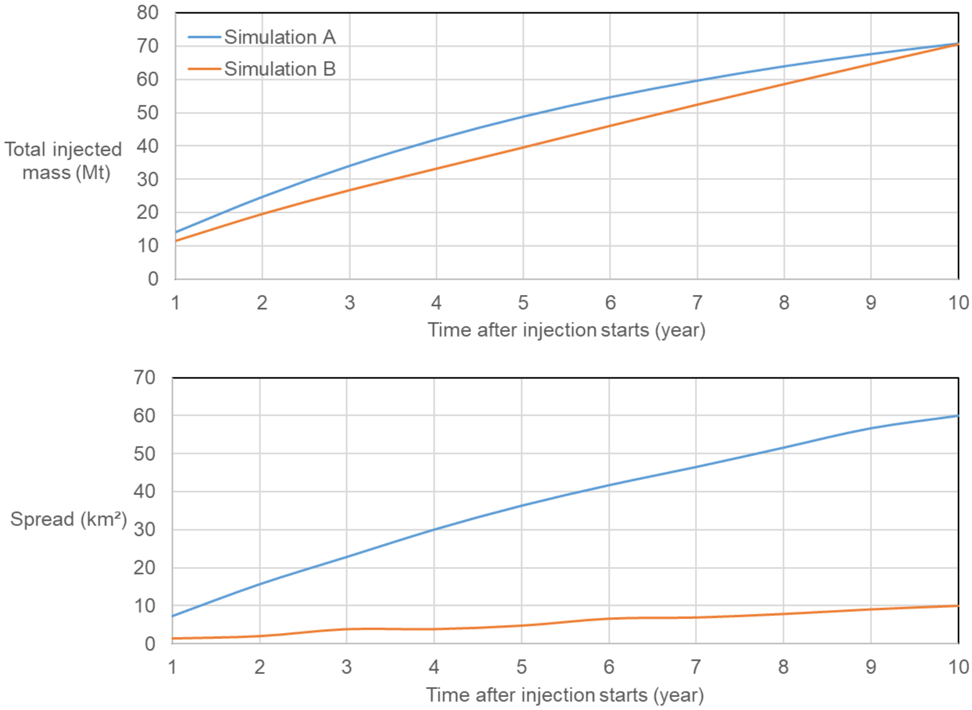


Figure 3. Storage capacity indicators (i.e. time series of the total injected mass, in Mt, and of the CO_2 plume spread, in km²) for simulation A and simulation B.

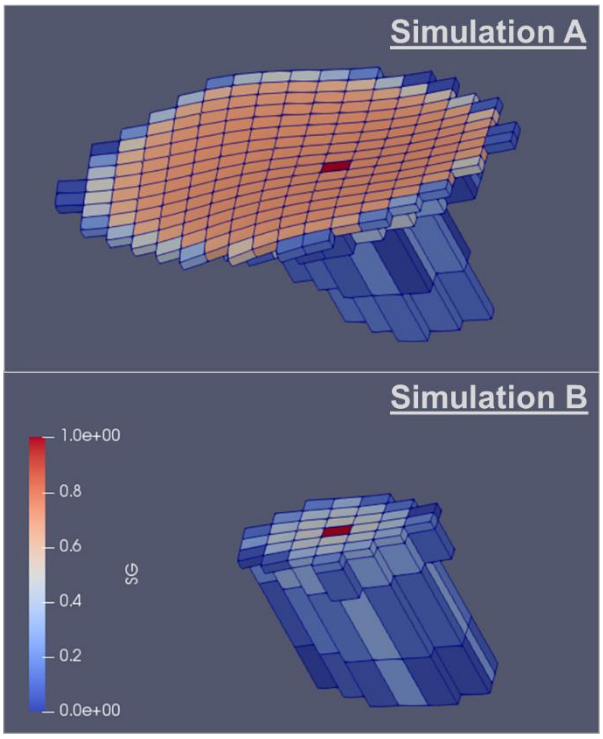


Figure 4. Concentration and spatial extension of the mobile free-phase CO_2 plume for simulation A and B (time step = 10 years)

5 Application for importance ranking (objective 1)

5.1 Within-method analysis

5.1.1 M-VBSA

The analysis of the influence of N is investigated for values ranging from 50 to 250. Above N=250, the computation becomes too expensive and the analysis lacks of meaning because metamodels are specifically dedicated to handle situations of low number of simulations. The tested metamodel is the ACOSSO (Adaptive COmponent Selection and Shrinkage Operator) method adapted to the treatment of both continuous and categorical input variables (Storlie et al. 2013). The main and total effects are computed using the Monte-Carlo-based algorithm described by Homma and Saltelli (1996) using N_{mc}=10,000 random samples. Confidence intervals are estimated using the bootstrap-based approach described using 100 random replicates; their widths both reflecting the error related to the use of a metamodel (instead of the true simulator) and the Monte-Carlo error (for the computation of the Sobol' indices).

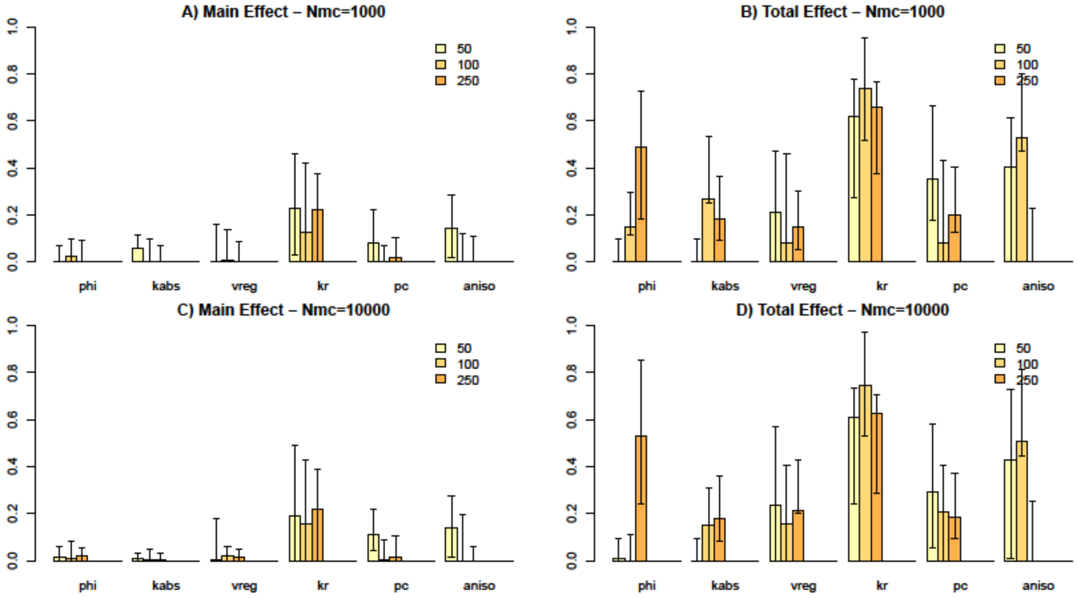


Figure 5. Sobol' indices computed using the ACOSSO metamodel trained with N simulations (N ranging from 50 to 250). A) Main effect (first order Sobol' index) given Nmc=1,000 random samples; B) Total effect given Nmc=1,000 random samples; C) Main effect (first order Sobol' index) given Nmc=10,000 random samples; D) Total effect given Nmc=10,000 random samples.

Figure 5 highlights several aspects:

- The identification of the most important parameter (here kr law) based on the value of the corresponding main effect (of the order of 20%, Figure 5 A&C) is little affected by either N or N_{mc};
- The impact of N is very large for the identification of the second most important parameter, which appears to be very difficult for N=50 (by judging on the value of the main effect as well as on the width of the confidence intervals);
- The impact of N_{mc} is less marked compared to the one of N;
- The identification of the negligible parameters (based on the value of the corresponding total effect, Figure 5 B&D) can largely be impacted by N. For instance the values of the total effect for anisotropy (denoted ani) are large for N≤100 (above 10%), but appears to be almost nil for N=250. The situation appears to be reversed for phi, which appears to be non-negligible for N=250 with a very large total effect (~50%);

- Whatever N or N_{mc}, the analysis of the large differences between the main and total effects clearly highlights a complex input-output relationship.

5.1.2 RF

A RF model is set up considering the commonly-used parametrisation (see e.g., Liaw and Wiener (2002), i.e. mtry=(number of input variables)/3=2 and nodesize=5). Given this set up, Figure 6 A) provides the evolution of the PVIM value and the 95%-confidence lower and upper bounds derived from the random permutation procedure of Altmann et al. (2010) using 100 iterates, given the number of numerical simulations. The PVIM values are normalised by the maximum value given each number of simulation results. Figure 6 B) provides the corresponding significance p-value so that an input variable is considered negligible for p-value >5%.

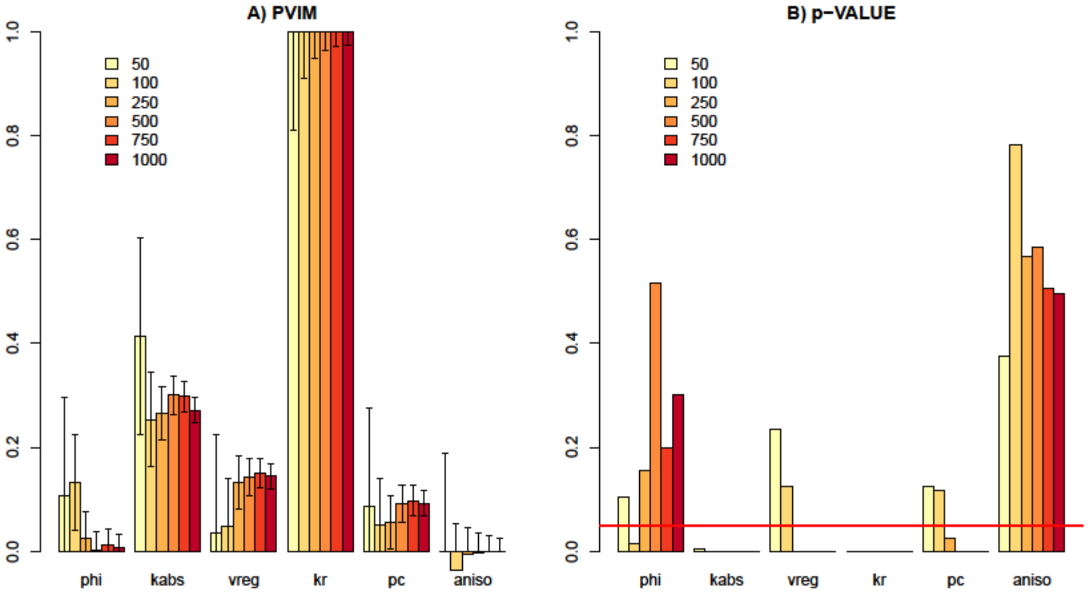


Figure 6. A) Permutation-based Variable Importance measure PVIM of RF model (set up with ntree=1,000; mtry=2 and nodesize=5); the 95%-confidence intervals are computed using the random permutation approach of Altmann et al. (2010); B) Significance p-value derived from the approach of Altmann et al. (2010). The input parameter is classified as negligible for p-value above the threshold of 5% (horizontal red line).

The following observations can be made:

- The ordering as well as the relative PVIM value of the most important parameters is little influenced by N;
- The kr law is systematically identified as the most important parameter whatever N;
- the PVIM values rapidly stabilise for N≥100 and the width of the confidence interval rapidly decreases with N;
- The screening of the negligible variables is impacted by N; For instance, the capillary pressure law pc and the regional hydraulic gradient vreg are identified as negligible (by adopting a threshold at 5%) for N≤100;
- The porosity phi is more complicated to analyse, because it switches from negligible (for N=50) to important (for N=100), and to negligible again for N≥250. This may be explained by the random nature of the procedure for p-value calculation.

5.1.3 RSA

To implement RSA, a criterion to separate two groups of outputs has to be defined. We decide to run de RSA technique for two different thresholds, namely the 10% (two groups with one containing all the lowest values) and the 90% quantile (two groups with one containing all the most extreme values). *N* values have been chosen to range from 50 to

250. We recall that the sensitivity index produced by RSA cannot be used for factor screening. Confidence intervals have been estimated with a bootstrap procedure (100 bootstrap samples). These results are shown in Figure 7.

The following conclusions can be drawn:

- We clearly see the influence of the threshold value chosen to divide the output population into two different groups. What can be seen is the importance of kr law for a threshold separating the lowest output values and that of the kabs parameter for a threshold separating the highest output values. It can be of great interest if the threshold has a real signification (acceptability limit or performance value for instance); however, it provides only little information on the influence of each parameter on the entire output range.
- The impact of N has no real importance for deriving the most influential parameter, which appears clearly at low N. However, a higher N is needed to reach enough stability to find out other important parameters and allows a ranking. Quasi-stability occurs for all ranking for the three last N values 500, 750 and 1000.
- Moreover, a relatively wide confidence interval may be noticed even for high N values (for N=1000, CI=0.19 in average for P10 threshold and CI=0.16 for P90 threshold) This may be explained by the choice of two extreme thresholds regarding the output distribution (P10 and P90), inducing a small number of samples in one group and avoiding reaching stability in the confidence intervals computation.
- The stability of the indices is not completely reached even for high N values especially for the P10-threshold case (not influencing the ranking though). The choice of threshold may also explain this observation as a better stability is reached with the 90% quantile and the median as threshold.
- Finally, a two-sample Kolmogorov-Smirnov (KS) test can be performed to assess the statistical significance of the results. The minimum sensitivity index for which the hypothesis that two distributions are the same can be rejected is, for a 95% confidence level (5% significance level), 0.43 for N=100 and 0.14 for N=1000. This highlights the fact that high N values are needed to detect slightly important parameters.

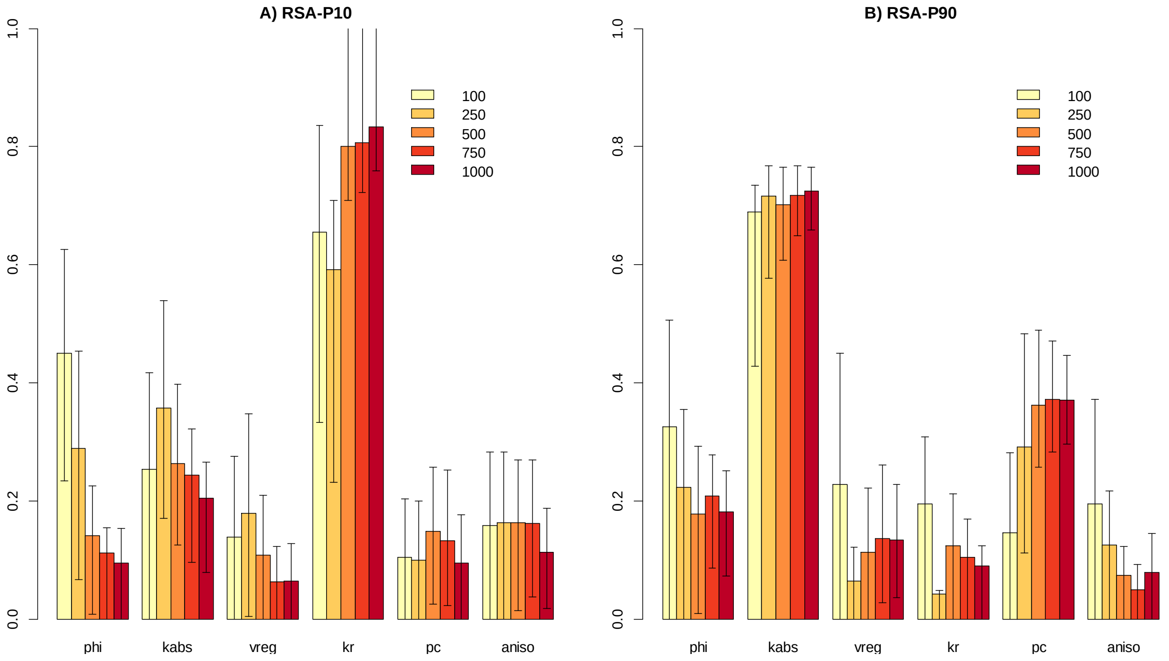


Figure 7. RSA sensitivity index with 95% confidence bounds with a threshold corresponding to the 10% quantile of the output distribution (left) and to the 90% quantile of the output distribution (right)

5.1.4 *DGSA*

The DGSA approach has been implemented for *N* ranging from 50 to 1,000. Given the fact that a class should not contain less than 10 simulations, the set of outputs have been clustered in 2 and in 5 different classes with the k-means clustering technique. The main sensitivity indices for the different *N* is provided in Figure 8 (2 classes) and in Figure 9 (5 classes), under 2 forms: the unstandardized indices gives information on the stability of the computed indices, while the standardized ones furnish insights on the stability of the ranking among parameters.

The main insights are the following:

- The non-standardized sensitivity index appears to converge relatively quickly with N and logically the confidence interval is improved as N increases;
- The ranking is less stable but appears to converge for the two cases as soon as N exceeds 500. For the 5-class situation a ranking stability is even found for N>100;
- The most noticeable observation is the influence of the chosen number of classes in the results. Similarly than for the RSA method where the choice of threshold influences the parameter ranking, the choice of the separation of the output set into different groups influences very much the results. On the example shown in Figure 8 (2 clusters), the kr is the most important parameter since the clustering led to one group containing low output values very much influenced by the kr parametrer. On the other example (5 classes), the application of the DGSA technique provide information on the parameter importance on more subranges of the output distribution: in that case, the influence of kr is lower, and the influence of kabs is higher. Note that similar discrepancies could have been observed by still considering 2 clusters but different;
- The statistical significance of the computed sensitivity indices are directly visible on the right side of Figure 8 and Figure 9. According to the test proposed by Fenwick et al. (2014), when the indices are equal or above 1, the null hypothesis (stating that the statistical distributions of a parameter are not different between the classes) can be rejected. We see that in the proposed examples, the null hypothesis is rejected for all inputs but the phi parameter in the 2-class situation and the aniso parameter in the 5-class situation, showing again differences according to the chosen classes;
- Even though the statistical distributions of these two parameters are not different between the classes, they can be important in terms of interactions with other parameters. Following the approach proposed by Fenwick et al. (2014), the null hypothesis stating that there is no statistical significant interaction among parameters is always rejected for N=1000, meaning that no parameter could be fixed. This seems to be due a too low number of simulations for a reliable conditional effect computation. It has to be noted that the interaction analysis depends on bins chosen for each parameter to compute the conditional effects.

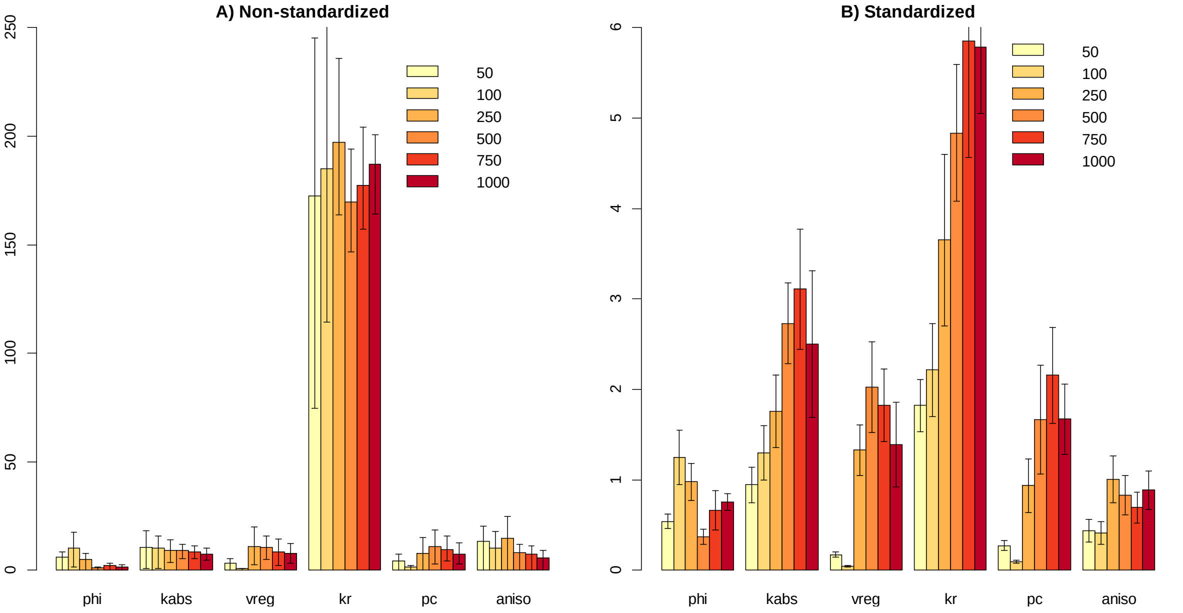


Figure 8. DGSA sensitivity index with 2 classes A) non-standardized (95% confidence bounds computed with the bootstrap technique) and B) standardized (standardization and error bars computed after the technique proposed by Fenwick et al. (2014) and Park et al. (2016))

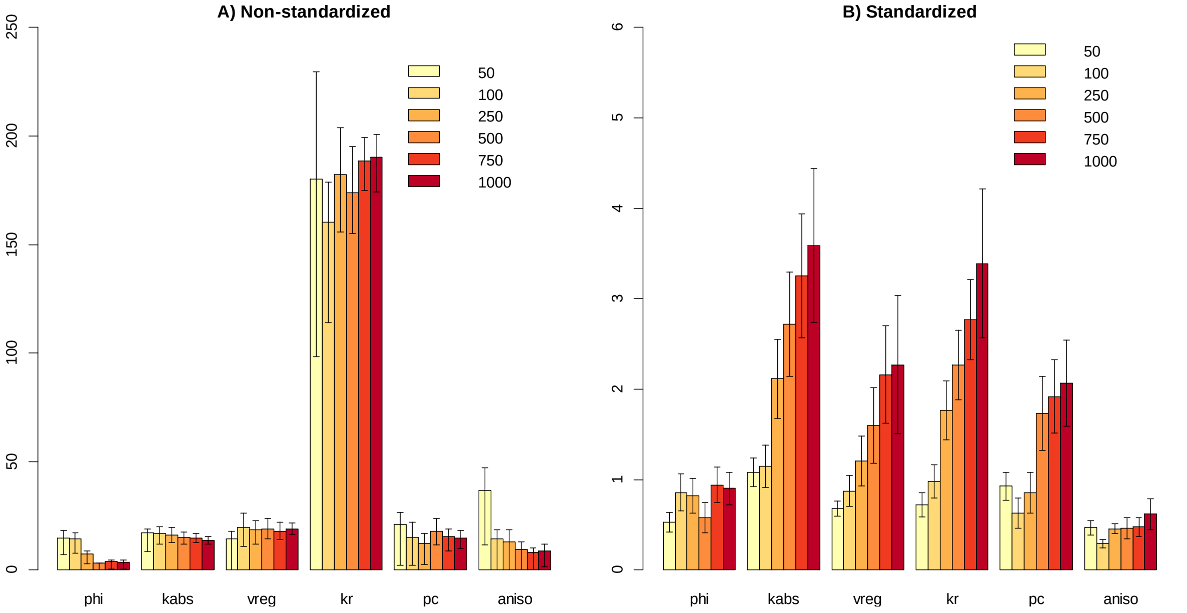


Figure 9. DGSA sensitivity index with 5 classes A) non-standardized (95% confidence bounds computed with the bootstrap technique) and B) standardized (standardization and error bars computed after the technique proposed by Fenwick et al. (2014) and Park et al. (2016))

5.1.5 *PAWN*

The DGSA approach has been implemented for *N* ranging from 250 to 1000, and for two types of statistics, namely the maximum and the median. It is important to note that, a dedicated experimental plan should be built to implement the PAWN approach. In the aim of comparing different methods, we did not follow the numerical procedure as proposed by Pianosi et al. (2015); we instead use the same set of simulations than for the other methods and divided the continuous parameters ranges into subranges (bins) in order to compute the Kolmogorov Smirnov (KS) distance only for discrete CDFs. In order to compute the CDFs for all discrete values and to perform a bootstrap procedure for confidence interval derivation, the analyses started with *N*=250. Therefore, the influence of the bin numbers for the *kabs* and *phi* parameters is also assessed (3 vs. 10 bins).

The results are shown in Figure 10 (for the 3-bin case) and in Figure 11 (for the 10-bin case). The results of the KS testing to detect non influential parameters is shown on Figure 12.

The following conclusions can be drawn:

- The influence of the number of bins for the continuous parameters is very low, as the sensitivity indices, the ranking of parameters and the list of non-influential parameters are very similar.
- The influence of N is not important: the indices are mostly stable between N=250 and N=1000, even though the uncertainty on these values (confidence intervals) decreases when N increases.
- The main changes in the PAWN outcomes come with the choice of the statistical indicator for computing the index. The maximum statistic emphasizes the importance of parameters whose specific values or specific range of input parameters lead to significant changes in terms of output: it is the case of the kr parameter in our example, where one value clearly leads to a different output values distribution (see Figure 10 and Figure 11). The median statistic emphasizes the importance of parameters that impact the output parameters homogeneously along their range of variations, *kabs* for instance in our example.
- The PAWN approach allows deriving the non-influencing parameters. For this purpose, a Kolmogorov-Smirnov test with a 95% confidence level (5% significance level) is performed for each $KS(x_i)$, and the frequency of success for one parameter is derived: a frequency of 1 means that the parameter can be considered as non-influential and can be fixed. The results are relatively stable with N and with the number of bins and would lead to identify *vreg* and *aniso* parameters as non-influential.

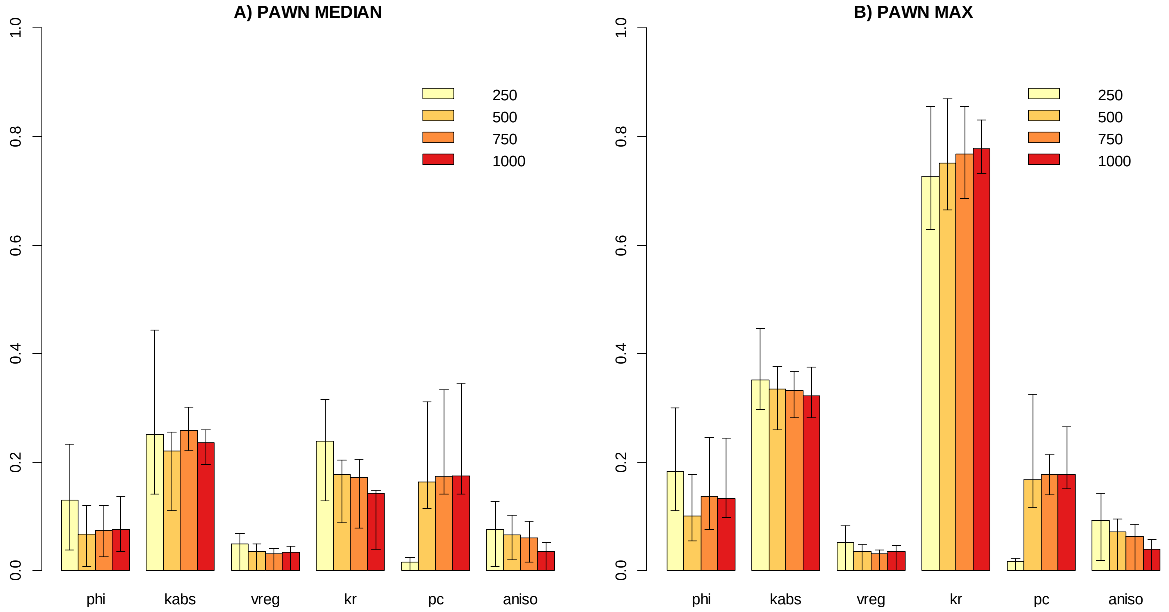


Figure 10. PAWN sensitivity index S_i with 3 bins for the continuous variables: A) PAWN indices computed with the median statistic and B) PAWN indices computed with the max statistic (95% confidence bounds computed with the bootstrap technique).

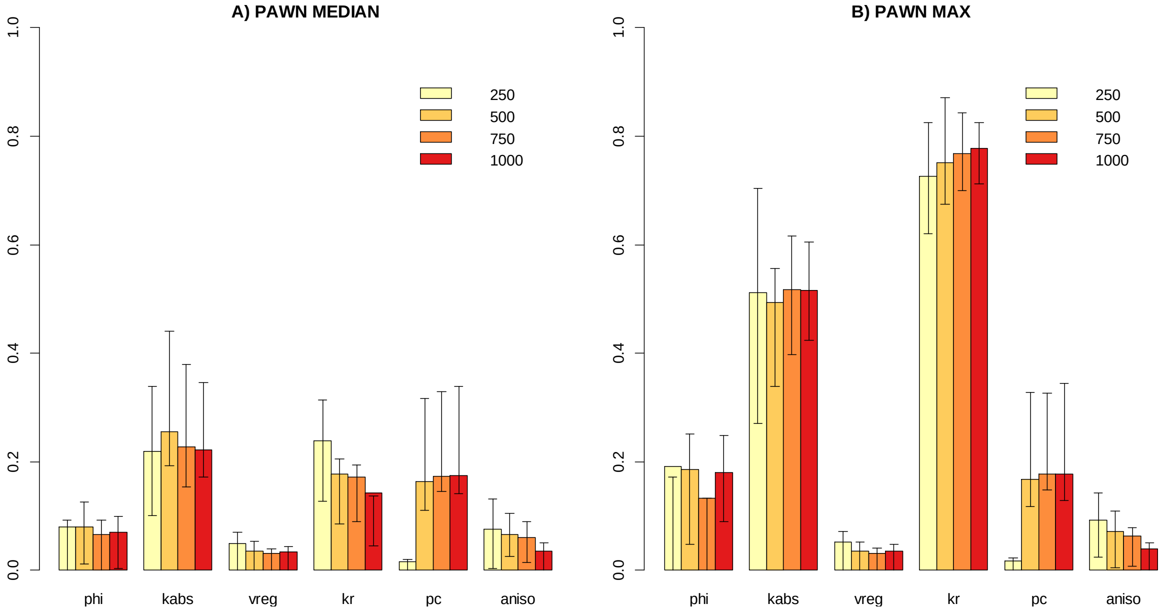


Figure 11. PAWN sensitivity index S_i with 10 bins for the continuous variables: A) PAWN indices computed with the median statistic and B) PAWN indices computed with the max statistic (95% confidence bounds computed with the bootstrap technique).

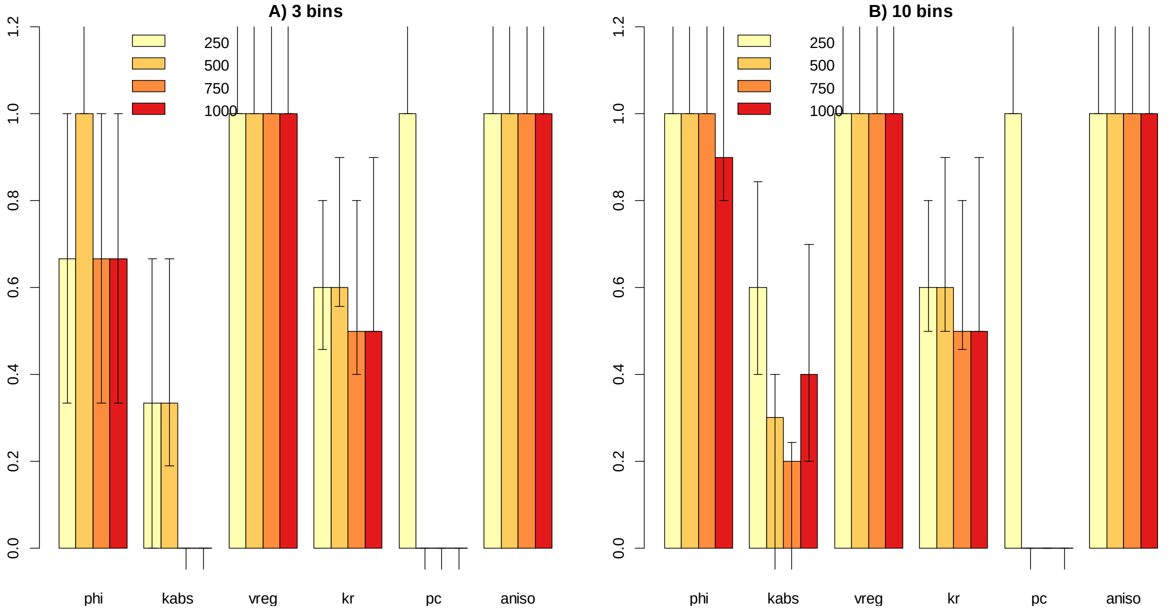


Figure 12. Frequency of success of the Kolmogorov-Smirnov test for each parameter: according to the procedure proposed by Pianosi et al. (2015), a frequency of 1 means that the parameter can be considered as non-influential. A) Considering 3 bins for the continuous variables and B) considering 10 bins for the continuous variables.

5.2 Between-method analysis

We compare the results between the different methods (“between-method” analysis) and formulate recommendations for practical implementation. The sensitivity measure are all normalised with respect to the maximum value reached for each method considering *N*=250 and *N*=100 (Figure 13). Table 5 summarises the main recommendations.

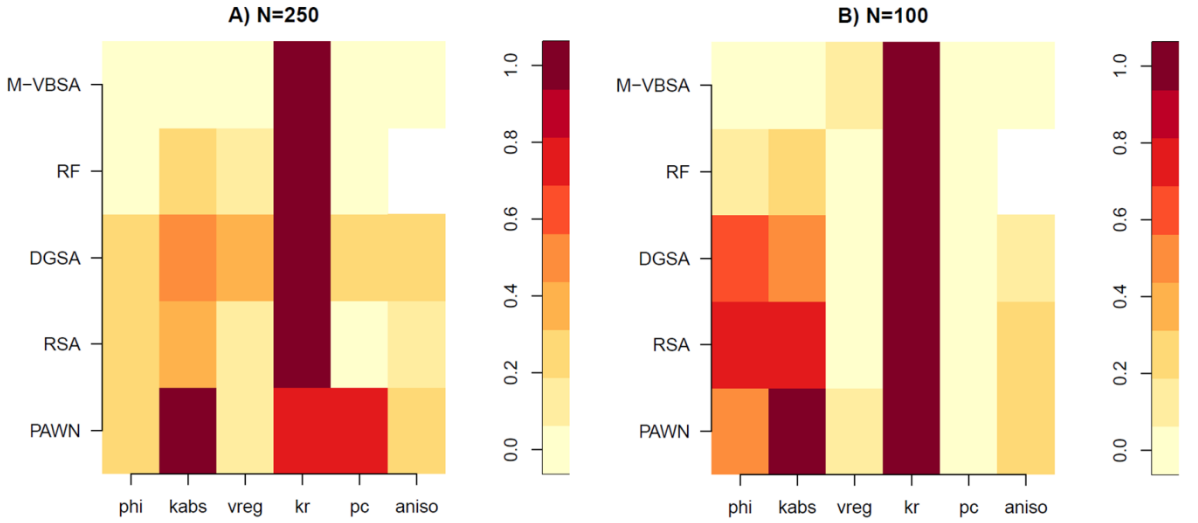


Figure 13. Sensitivity measure normalised with respect to the maximum value reached for each method considering N=250 (A) and N=100 (B).

Several observations can be made:

- The law of *kr* is systematically ranked first whatever the used methods and whatever N except for PAWN, which also identifies *kabs* as a strong contributor to sensitivity;
- The differences of PAWN with the other methods may be related to the interpretation of PAWN sensitivity as a deviation of the whole cumulative probability distribution compared to the use of a specific statistical moment like the variance for M-VBSA and RF (Wei et al. 2015);
- The selection of the second most important parameter is more complicated and differs depending on the type of method and on N: RF, DGSA & RSA all agree on the selection of the second most important parameter, namely *kabs* but only for N sufficiently high;
- Opposite behaviours of variable importance ranking are noticed. DGSA select a large number of parameters with moderate influence (relatively to the sensitivity measure of the most important parameter kr), whereas M-VBSA highlight the low contribution of the other parameters.

Method	**Description**	**Interpretation of VIM**	**Pros**	**Cons**
M-VBSA	Variance-based Sensitivity Analysis combine with Metamodel	The main effect is the first order Sobol' index and is interpreted as the contribution of the input variable to the output variance; The total effect is interpreted as the global contribution including the main effect and the interactions with the other parameters	Intuitive and rigorous interpretation as a proportion of variance; Feedbacks in a large variety of domains	Sensitivity to the number of simulations, which imposes a careful examination of the predictability of the metamodel Convergence analysis when using Monte-Carlo algorithm
RF	Permutation-based Variable Importance Measure PVIM	Interpretation VIM as a decrease of predictability (measured by the mean	Little influence of the RF parameters (*mtry*, *ntree*,	Sensitivity of the p-value algorithm to

	derived from the Random Forest method with significance estimate using Altmann et al.'s approach	squared error), which is related to the the non-standardized Sobol' total effect index;	nodesize and even split rule); Robutsness to the number of simulations	the number of permutations.
RSA	Regional sensitivity analysis based on the CDFs on input parameters	The sensitivity index quantifies the difference between input empirical CDFs of two different groups of simulations defined according to the outputs values.	Easy to compute, and possible even for low number of samples and for categorical inputs. Adapted when the outputs can be naturally divided into two different groups, and useful for factor mapping.	Do not account for interactions among input parameters. No procedure for factor fixing. Cannot handle more than two groups, and very much influenced by the choice of these two groups: may lead to difficulties in interpretation..
DGSA	Sensitivity analysis based on the CDFs on input parameters after classification of the output parameter into multiple sets.	The sensitivity index quantifies the difference between input empirical CDFs of multiple groups of simulations defined according to the outputs values. Conditional indices quantifies multiple way interaction among parameters.	Can handle multiple groups of outputs. Provide a lot of information on interactions among parameters (two-way interactions) Can help fixing parameters to the less influential value/range	The proposed statistical test might lead to a strong sensitivity to the number of simulations. The test for statistically significant interactions require a high number of simulations. Very much influenced by the outputs classification: may lead to difficulties in interpretation.
PAWN	Sensitivity analysis based on the output CDFs changes when fixing inputs	The sensitivity index associated to an input parameter quantifies the difference among the CDFs of the output for different fixed values of the input parameter.	Relatively good convergence for sensitivity analysis, ranking and fixing with the number of simulations.	Dependent on the choice of the statistic: may lead to difficulties in interpretation. A procedure is proposed for factor

			Compared to other density-based GSA, rely on CDF whose approximation is easier than PDF	fixing but interactions are not accounted for.

Table 5. Characteristics of the SA methods for handling parameter and model uncertainties

6 Application for reliability quantification (objective 2)

6.1 Approach / procedure

We recall that the objective of this section, dedicated to uncertainty quantification, is to evaluate the reliability of the capacity estimates calculated with a long running dynamic model by propagating the uncertainties encountered within the flow modelling workflow and by assessing different quantiles of the capacity estimates. The case study is detailed in section 4 and the capacity estimators are two time series, one relative to the total injected mass and another relative to the surface of the CO_2 plume, named respectively mass and surface in the following study.

As explained in section 3.2, the approach followed in this paper relies on a limited set of “exact” simulations and a larger set of low fidelity simulations (proxy model). For the application to the case study:

- We chose as proxy model, the dynamic model built from a coarse grid (in horizontal and vertical direction), compared to the exact model, which is a model built from a refined grid.
- The quantities of interest are twofold: total injected mass, and spread of the CO_2 plume. The application is therefore carried out on these two model responses.
- The final outcomes of the UQ is the estimation of 3 quantiles (P05, P95 and P50) for the two quantities of interest.

An error model is then constructed, based on the limited set of exact simulations, to unbias the proxy model responses; the proxy model is corrected with the error model to predict the exact model responses for the larger set of simulations.

The following section describes, step-by-step, the different stages of the procedure:

1- Construction of the error model
2- Uncertainty propagation

As the quantities of interest are time series, two different approaches are tested:

A. The UQ procedure is applied to each time steps, i.e. a new error model (regression) is constructed for t=1, 2, …, 10 years (application on scalar quantities)
B. The UQ procedure is applied once a Principal Component Analysis is used on the simulation responses time series, with the purpose of diminishing the dimension of the response spaces and therefore the size of the regression problem (application on time series).

6.2 Presentation of the dataset

540 sets of input parameters’ vectors have randomly been generated to cover the domain of variation of the input parameters. A Latin hypercube sampling approach (with the uniform distribution) is followed to tackle the continuous variables, whereas sampling with replacement (with the discrete uniform distribution) is performed for the discrete variables.

Therefore, 540 “low fidelity” simulations have been run: the CPU time is approximately 1 min and 30 sec by simulations.

For methodological purposes, the same simulations have been run with the refined mesh model. The chosen mesh have therefore been chosen to allow a large number of simulations (CPU time of 20 minutes approximately). In a real case, the CPU time of dynamic simulations can typically reach several hours, which makes the direct application of uncertainty quantification hardly feasible.

6.3 Application of the UQ procedure on scalar quantities

Figure 14 summarizes the UQ procedure applied for each time-step of the capacity estimator time series: as explained above in section 3.2, UQ requires a large amount of simulations, but the significant computational time of reservoir simulations makes UQ impossible in practice. The use of a proxy model enables to run a large number of simulations, but such low fidelity models can lead to a biased estimation of the outcomes of interest. To correct this bias, an error model is built from a limited learning set of exact (high fidelity) and low fidelity simulations. Once constructed on a limited learning set of realizations, the error model can be used to predict exact responses from a large set of proxy simulation responses, allowing uncertainty propagation and quantile estimation.

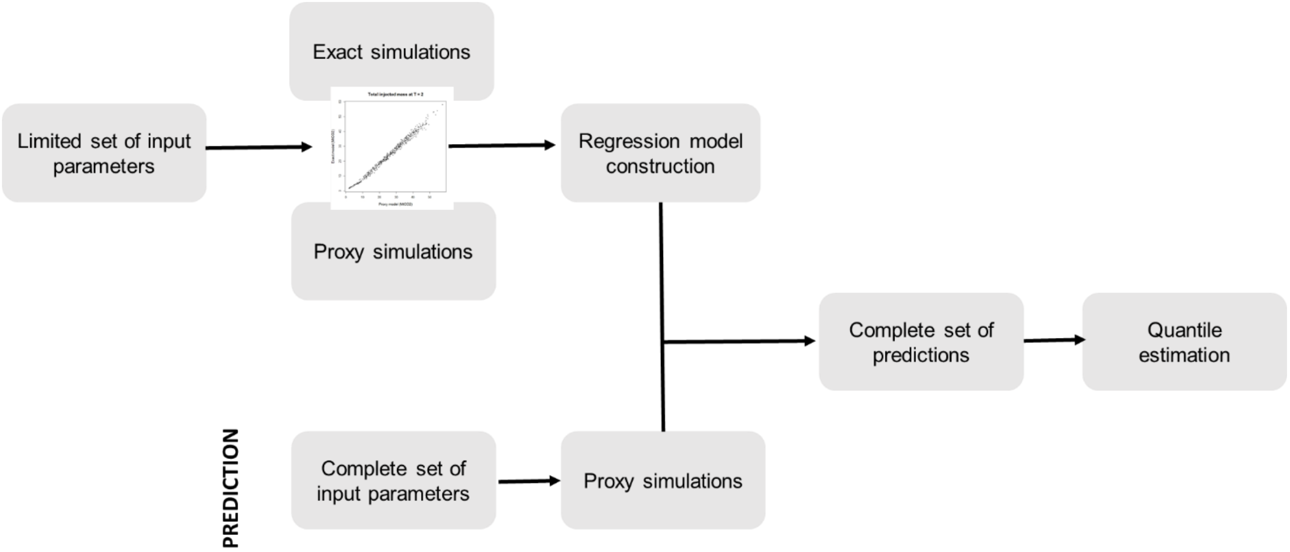

Exact simulations
Limited set of input parameters
Regression model construction
Proxy simulations
Complete set of predictions
Quantile estimation
PREDICTION
Complete set of input parameters
Proxy simulations

Figure 14: Methodology illustration.

6.3.1 Choice of the error model

We recall that the objective of this stage is to construct an error model from a limited set of simulations run both with the proxy and exact model. This error model has to be built for the two capacity estimators and for each time step of the model response time series.

Figure 1515 allows the visualisation of the data using scatter plots at different time step T of the simulation: on the left side the total injected mass (called ‘mass’), on the right the spread of the CO_2 plume (called ‘surface’). The visual inspection shows that a linear relationship between the proxy and the exact model can be expected, for both capacity estimators (mass and surface). The strength of the linear relationship seems stronger for the latest time steps. We can also notice a discontinuous evolution of the surface indicator, especially for that built from the proxy model; this observation is due to the fact that this indicator equals the surface of the plume in contact with the caprock, and therefore is proportional to the size of one grid (and by definition not continuous).

Before going further with a linear error model, the assumptions made when performing linear regression are tested against our data set (see for instance Peña and Slate, 2006). The assumptions to be tested are the following:

1) Independence of the error components
2) Homoscedasticity (constant variance of the error components)
3) Normality of the error components (this hypothesis is not compulsory for establishing the linear model but it is for assessing confidence intervals of the linear model parameters)

Diagnostic/residual plots are displayed on Figure 16, regarding the mass variable at the final time step:

- the plot of residuals against fitted values informs on potential non linear relationships between the proxy model and the exact one that would not be captured by the model (linear model assumption): Figure 16 shows a horizontal red line indicating no non linear patterns.

- the scale location plot is relatively similar to the plot of residuals against fitted values but makes easier the graphical assessment of homo/hetero-scedaticity by using the square root of standardized residuals: Figure 16 shows a non linear red line and data point that does not appear to be equally and randomly spread around that line (higher spread of residuals for larger fitted values), which indicates an heteroscedastic behaviour;
- the normal QQ plot, which indicates if the residuals are normally distributed: Figure 16 show that the residuals data points deviate from a straight line, questioning their normality.

In addition to that visual inspection, the different assumptions were tested with different statistical tests: Shapiro test to check the normality of the error components, Durbin-Watson test to verify if the residuals are non-correlated and Breusch-Pagan test to evaluate heteroscedasticity. In line with the graphical assessment, the results of the statistical tests show that the assumptions were not well satisfied. In particular, for the final time step, the p-value obtained to perform the Breusch-Pagan test is 2.2e-16, indicating a rejection of the null hypothesis (variance of the residuals is constant) at the 5% significance level.

The same observations (graphically, see Figure 18, and with statistical tests) were done for the spread variable.

To go further, the same procedure has been applied on transformed variables. Transformations (logarithm or square root are commonly applied) have indeed the potential to correct certain violations, potentially enabling the continuation of the regression analysis. In this study, as the variable values are spread over several decades, (in particular the total injected mass), the variables were transformed with the square root function: the simple square root for the surface variable and the double square root for the mass variable.

Significant improvements were noticed regarding the mass variable: Figure 17 show the linear model diagnostic plots with a double square root transformation on the mass variable at the final time step. In particular, less heteroscedaticity is on the scale location plot, and the residuals follow much better a normal distribution. The p-value obtained to perfom the Breusch-Pagan test is with the transformed mass variable is 0.396.

The same observation can be made on the surface variable at the latest time step (Figure 19) transformed with the square root function, even though the improvements are less convincing (notably the normality of residuals). The p-value for the heteroscedasticity without transformation was lower than 2.2e-16 while it is equal to 0.58 when we consider the square root transformed surface.

Note that this verification procedure has been carried out for the 10 time steps T of the simulation.

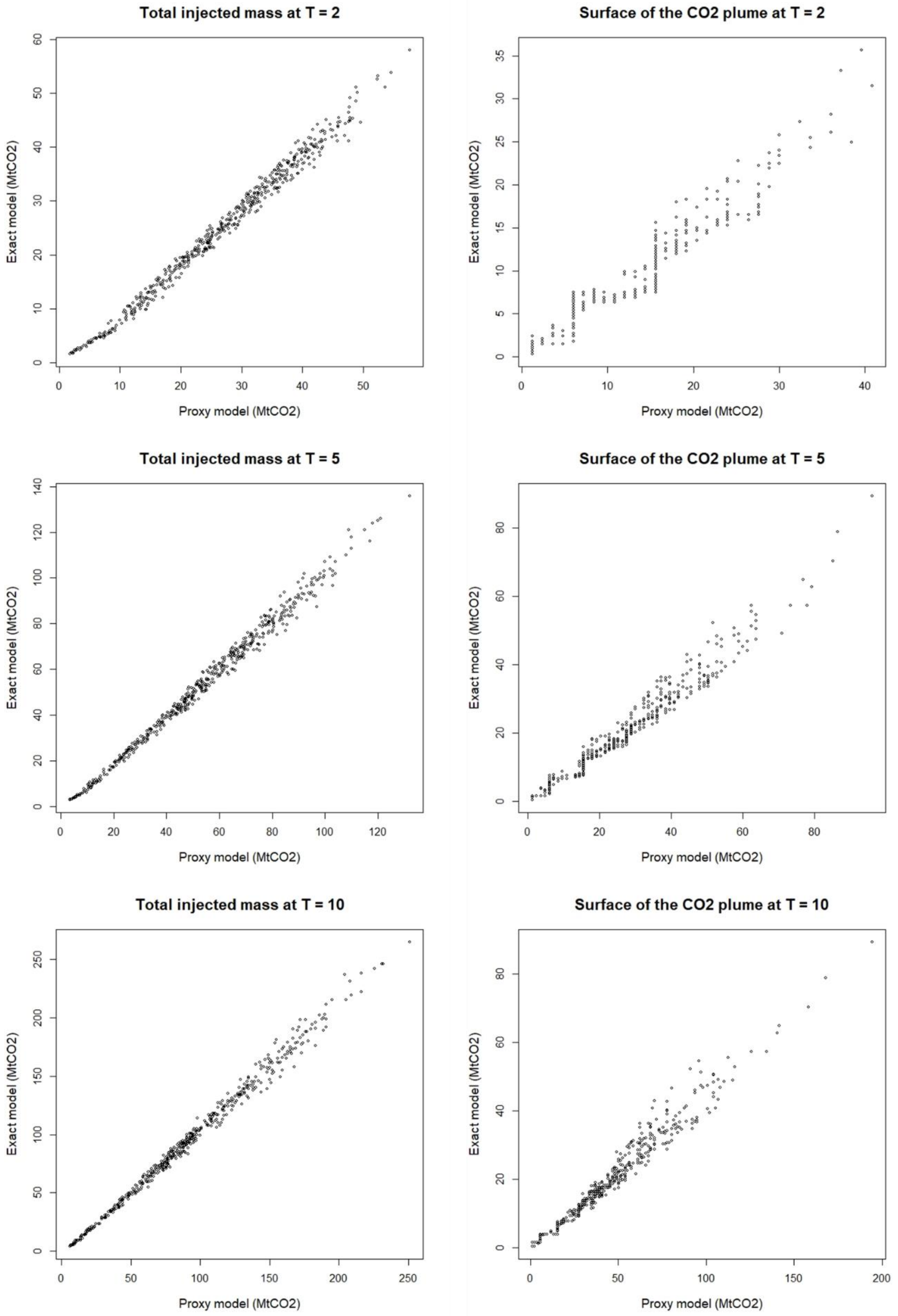


Figure 15: Scatter plots between the proxy model and the exact one at different time steps (T = 2, T = 5, T = 10) for the total injected mass on the left and the surface of the CO_2 plume on the right

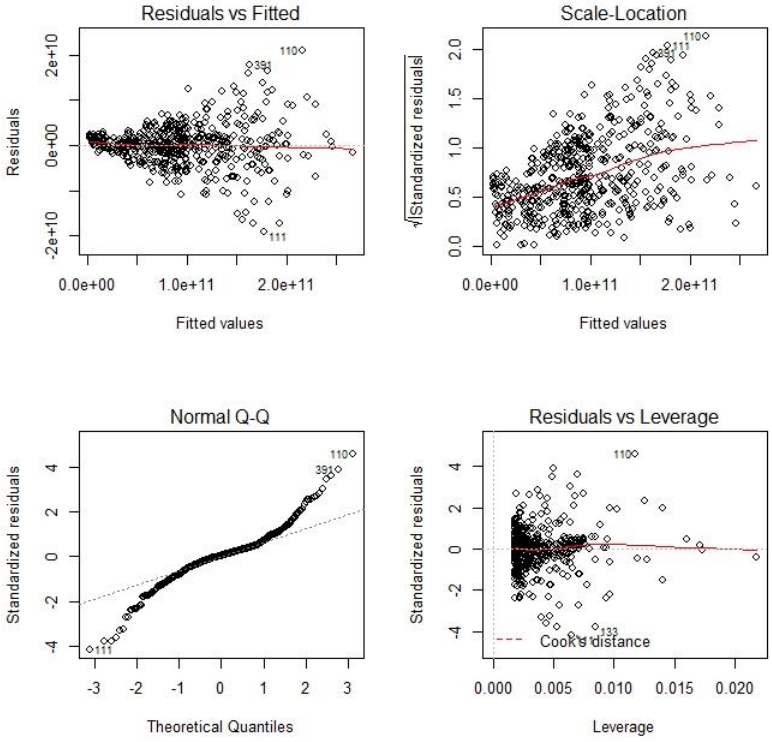


Figure 16: Residual plot of the linear model constructed for the mass variable (final time step), without transformation.

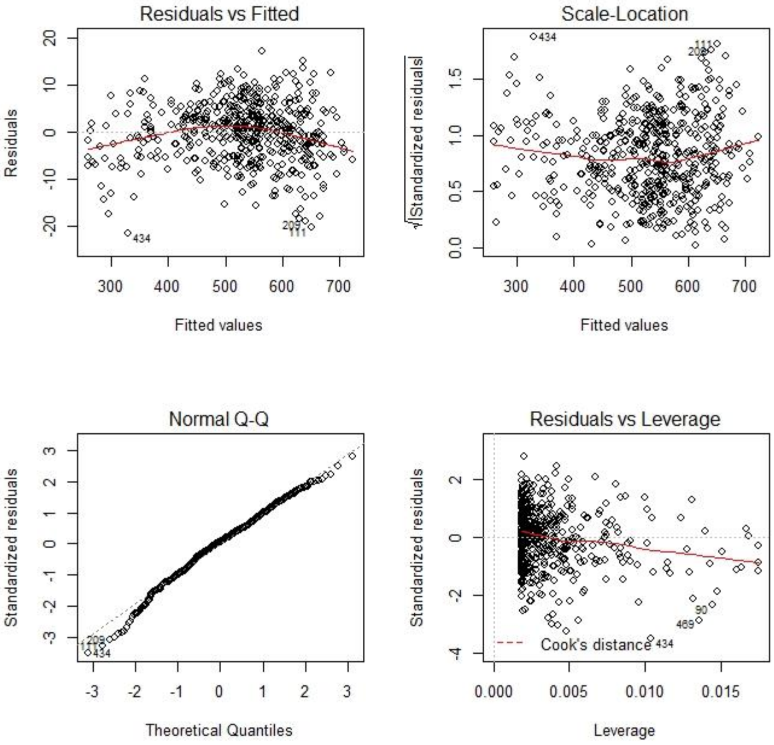


Figure 17: Residual plot of the linear model constructed for the transformed mass variable (final time step), i.e. $\sqrt[4]{mass}$

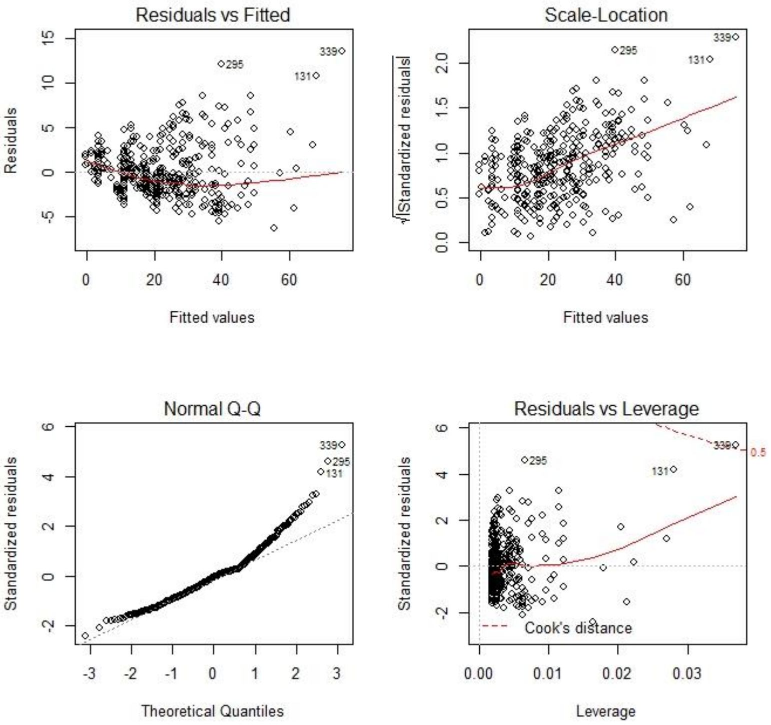


Figure 18: Residual plot of the linear model constructed for the surface variable (final time step), without transformation

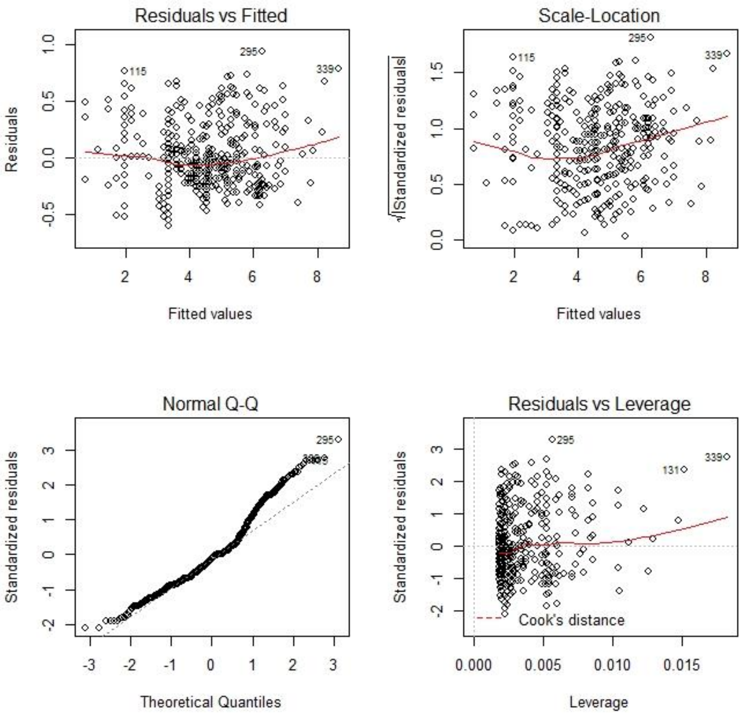


Figure 19: Residual plot of the linear model constructed for the transformed surface variable(final time step), i.e. $\sqrt{surface}$.

6.3.2 *Fitting of the linear model: assessment of the regression quality, confidence intervals and prediction intervals*

The quality of the linear model can be assessed regarding the approximation (goodness of fit) and prediction (goodness of prediction) ability of the regression. The approximation quality is evaluated using the differences between approximated and "true" model outputs (i.e. the errors) and by computing the coefficient of determination R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where the y_i are true model outputs $(\mathrm{i}=1\ldots\mathrm{n})$, $\bar{y}$ the mean out of the model outputs, and $\hat{y}_i$ the approximated values.

A R^2 value close to one indicates that the regression has been successful in matching the observations.

The other key issue is the validation of the approximation i.e. the verification that the regression can estimate with high accuracy not-yet-seen new input data. This can rely on leave-one-out cross-validation (LOOCV) procedures (e.g., Hastie et al., 2009). This technique can be performed as follows:

1) an element is removed from the initial set, and a linear model is constructed using the remaining set;
2) the element removed from the initial set constitutes the validation set;
3) the residual is estimated. This procedure performed for each element of the initial set. Finally, the coefficient of determination R_{CV}^2 (or Q^2) is computed (same formula than above with the predicted values): a coefficient R_{CV}^2 close to 1 indicates that the linear model is successful in matching the observations.

While the coefficient of determination are relative measure of fit and prediction quality, the RMSE (root mean square error) can also be computed to assess the absolute quality of the model:

$$RMSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

The ability of the linear model to approximate and predict quantities of interest (mass and surface) as a function of the learning set size has been assessed. As mentioned beforehand, we have a sample of 540 simulation couples for both proxy and exact model (nb. for methodological purposes), which allow us to test the optimal learning set to build a satisfactory error model. For each time step T, a sub-sample of size N has been randomly drawn from which a regression model between proxy and exact responses was built. Figure 20 to Figure 22 show the different regression models constructed for the mass variable for different sample size N and at, respectively, time step T=2, T=5, and T=10. These scatter plots of predicted vs observed values are constructed as such: in red, the data points that served to construct the regression model and in black, the data points derived from a Leave-One-Out Cross Validation (LOOCV) analysis performed on each regression model. The data points approximately follow the first bisector (red line), which indicates a good agreement both in terms of approximation and prediction. The criteria R^2 and ${R_{CV}}^2$ are high (minium value of 0.99) and the RMSE's are relatively low (no larger than 7.08), which confirms the satisfactory quality of the regression model. Even if these validation criteria are very good and with few variability, we can observe that R^2 (RMSE's) do not necessarily increase (or decrease) with higher sample size N. This is be explained by the fact that regression model depends on the dataset sampled.

To overcome that limitation, we can perform a bootstrap method, which is a resampling technique used to estimated statistic by sampling a dataset with replacement. The results with the bootstrap method are shown in Table 6: RMSE and R²_CV are between, respectively, 0.988 and 0.996, and 5.72 and 6.62. We can observe that logically ${R_{CV}}^2$ (RMSE's) increase (decrease) depending on the sample size N but also depending on the time step. This corroborates the observation made previously on the strength of the linear relationship that seemed stronger for the latest time steps.

Mass	***T = 2***		***T = 5***		***T = 10***	
N (sample size)	**R²_CV**	**RMSE**	**R²_CV**	**RMSE**	**R²_CV**	**RMSE**
20	0.988	6.45	0.993	6	0.994	6.62
30	0.989	6.18	0.994	5.9	0.995	6.34
50	0.99	6.18	0.994	5.81	0.995	6.32
100	0.99	6.05	0.995	5.77	0.995	6.27
200	0.991	6.05	0.995	5.72	0.996	6.23
300	0.991	6.02	0.995	5.72	0.996	6.23

Table 6: R^2_CV and RMSE derived from the leave-one-out cross validation procedure; with 2000 bootstrap samples of regression models constructed with N learning set size; at time step T=2, T=5 and T=10.

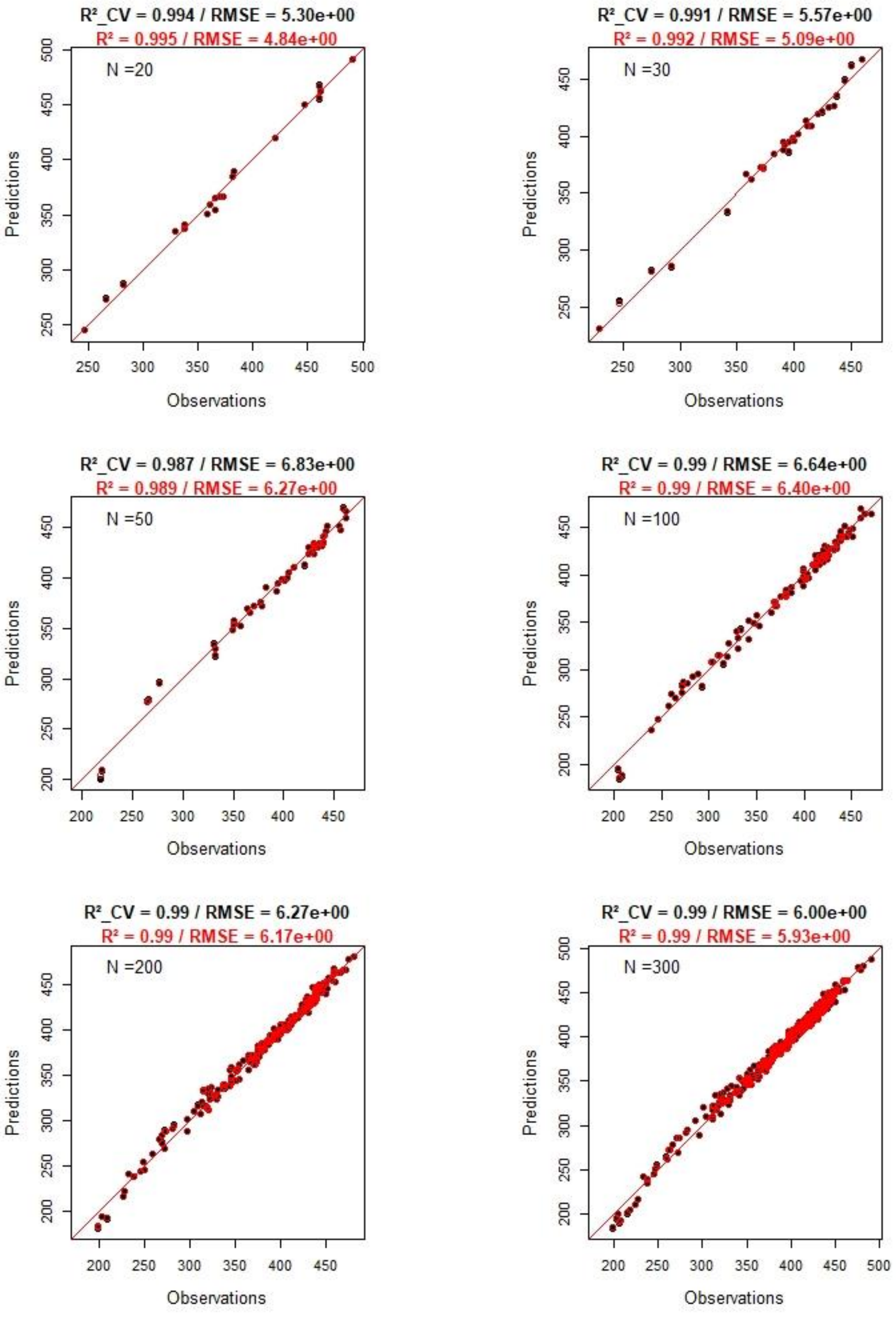


Figure 20: Scatter plots of predicted and observed values for the mass variable, at time step T = 2 and different sample size of the regression model; in red, points that served to construct the regression linear; in black, points used for the leave-one-out cross-validation (LOOCV). For each sample size N, two coefficients are estimated: the coefficient of determination of the regression model (R²) and the root mean squared error (RMSE), for both regression model (in red) and the LOOCV (in black).

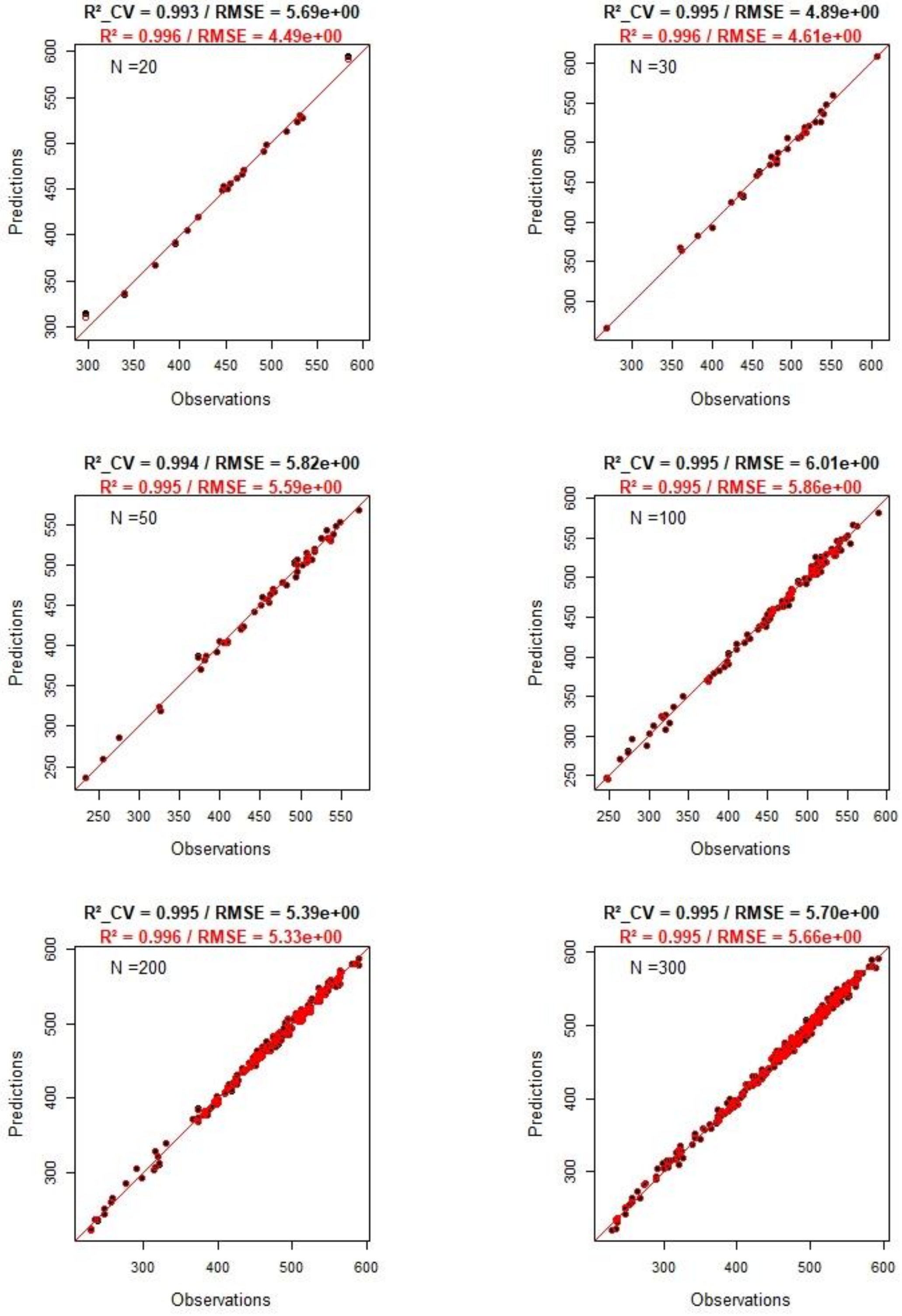


Figure 21: Scatter plots of predicted and observed values for the mass variable, at time step T = 5 and different sample size of the regression model; in red, points that served to construct the regression linear; in black, points used for the LOOCV. For each sample size N, two coefficients are estimated: the coefficient of determination of the regression model (R²) and the root mean squared error (RMSE), for both regression model (in red) and the LOOCV (in black).

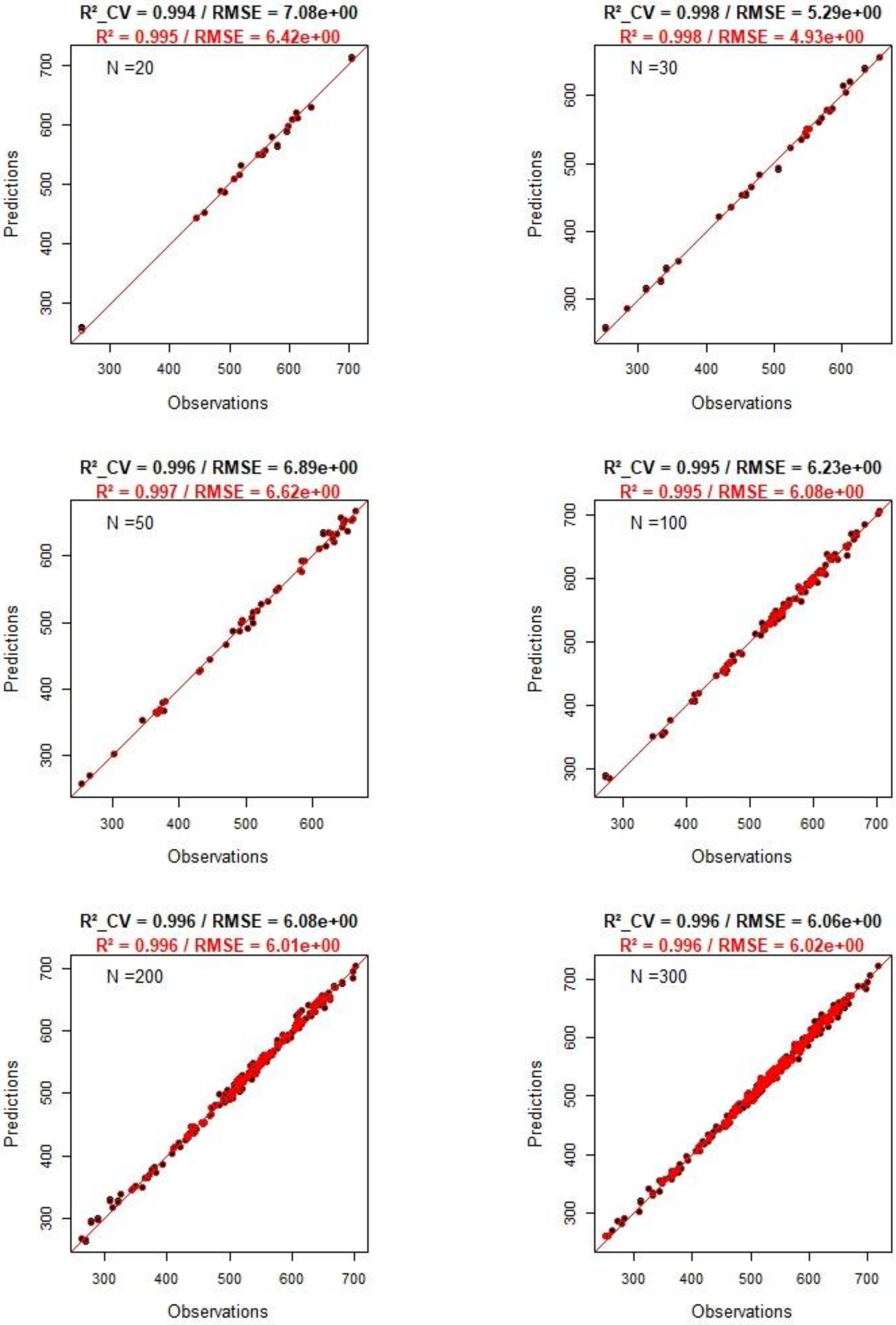


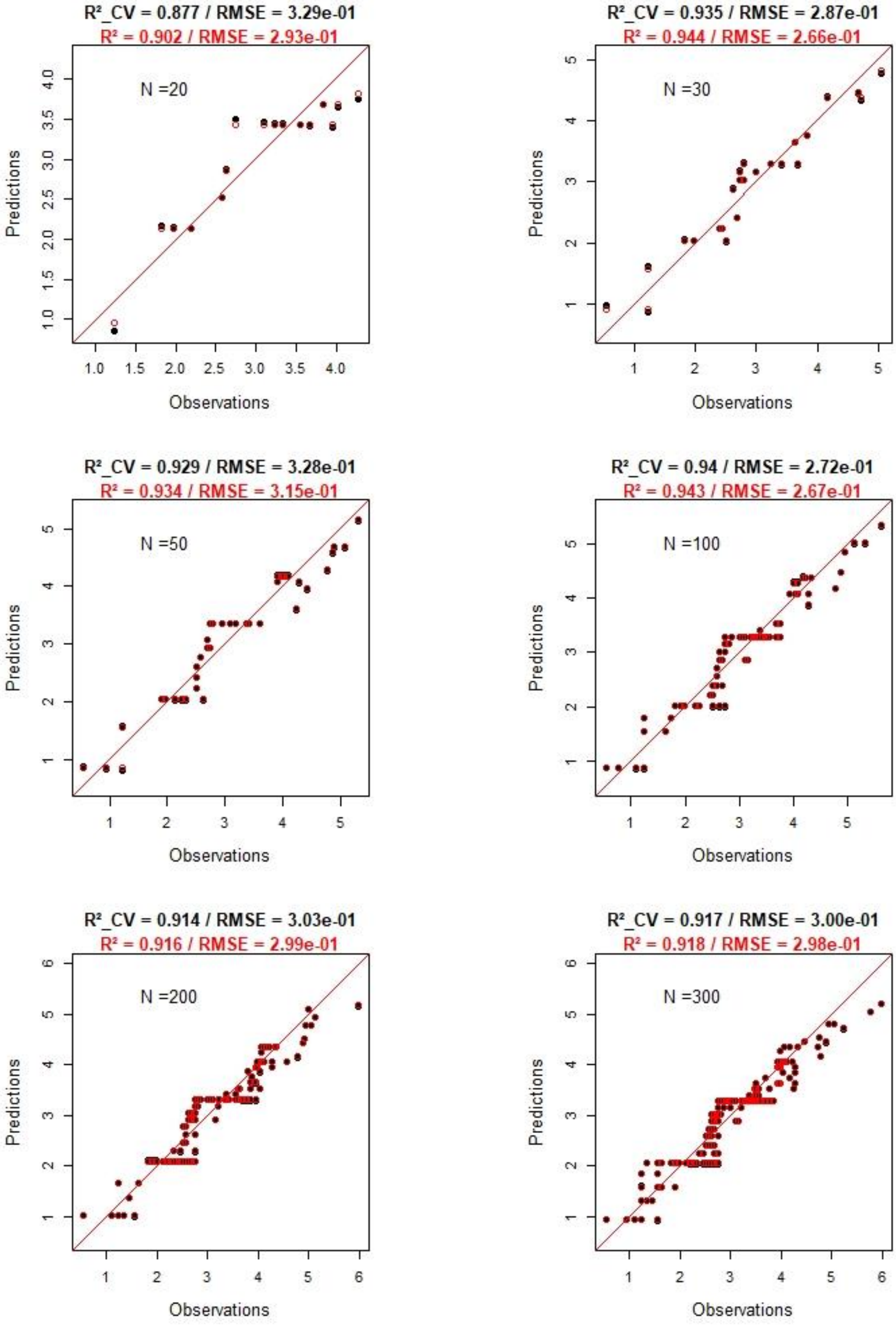
Figure 22: Scatter plots of predicted and observed values for the mass variable, at time step T = 10 and different sample size of the regression model; in red, points that served to construct the regression linear; in black, points used for the LOOCV. For each sample size N, two coefficients are estimated: the coefficient of determination of the regression model (R²) and the root mean squared error (RMSE), for both regression model (in red) and the LOOCV (in black).

Figure 23 to Figure 25 provide the different regression models constructed for the surface variable for different sample size N at, respectively, time step T=2, T=5, and T=10 years. Unlike the mass, the linear relationship is less obvious for the first time steps (for instance T=2) and small sample size (for instance N=20), we clearly see the discontinuous evolution of the surface indicator noticed previously. The strength of the linear relationship is stronger for the latest time steps.

Despite the “step effect”, the data points approximately follow the first bisector. The criteria R^2 and ${R_{CV}}^2$ are relatively high (minimum value of 0.877) and the RMSE’s are relatively low (no larger than 0.377), which confirm, also for the surface variable, the satisfactory approximation and predictive quality of the regression model. As for the mass, even if these validation criteria are very good, we can observe that R^2 (RMSE’s) does not necessarily increase (decrease) depending on the sample size N. For the same reasons, we perform a bootstrap procedure (results are shown in Table 7): RMSE and ${R_{CV}}^2$ are between, respectively, 0.911 and 0.978, and 0.285 and 0.322. We can also observe that logically ${R_{CV}}^2$ (RMSE’s) increase (decrease) depending on the sample size N but also depending on the time step (stronger strength of the linear relationship for the latest time steps)

Surface	***T = 2***		***T = 5***		***T = 10***	
N (sample size)	**R²_CV**	**RMSE**	**R²_CV**	**RMSE**	**R²_CV**	**RMSE**
20	0.911	0.301	0.961	0.298	0.974	0.322
30	0.917	0.295	0.963	0.294	0.975	0.317
50	0.923	0.292	0.965	0.291	0.977	0.312
100	0.925	0.29	0.966	0.288	0.978	0.31
200	0.927	0.289	0.967	0.285	0.978	0.308
300	0.927	0.288	0.967	0.286	0.978	0.308

Table 7: R^2_CV and RMSE derived from the leave-one-out cross validation procedure; with 2000 bootstrap samples of regression models constructed with N learning set size; at time step T=2, T=5 and T=10.



R²_CV = 0.877 / RMSE = 3.29e-01
R² = 0.902 / RMSE = 2.93e-01

R²_CV = 0.935 / RMSE = 2.87e-01
R² = 0.944 / RMSE = 2.66e-01

R²_CV = 0.929 / RMSE = 3.28e-01
R² = 0.934 / RMSE = 3.15e-01

R²_CV = 0.94 / RMSE = 2.72e-01
R² = 0.943 / RMSE = 2.67e-01

R²_CV = 0.914 / RMSE = 3.03e-01
R² = 0.916 / RMSE = 2.99e-01

R²_CV = 0.917 / RMSE = 3.00e-01
R² = 0.918 / RMSE = 2.98e-01

Figure 23: Scatter plots of predicted and observed values for the surface variable, at time step T = 2 and different sample size of the regression model; in red, points that served to construct the regression linear; in black, points used LOOCV. For each sample size N, two coefficients are estimated: the coefficient of determination of the regression model (R²) and the root mean squared error (RMSE), for both regression model (in red) and the LOOCV (in black).

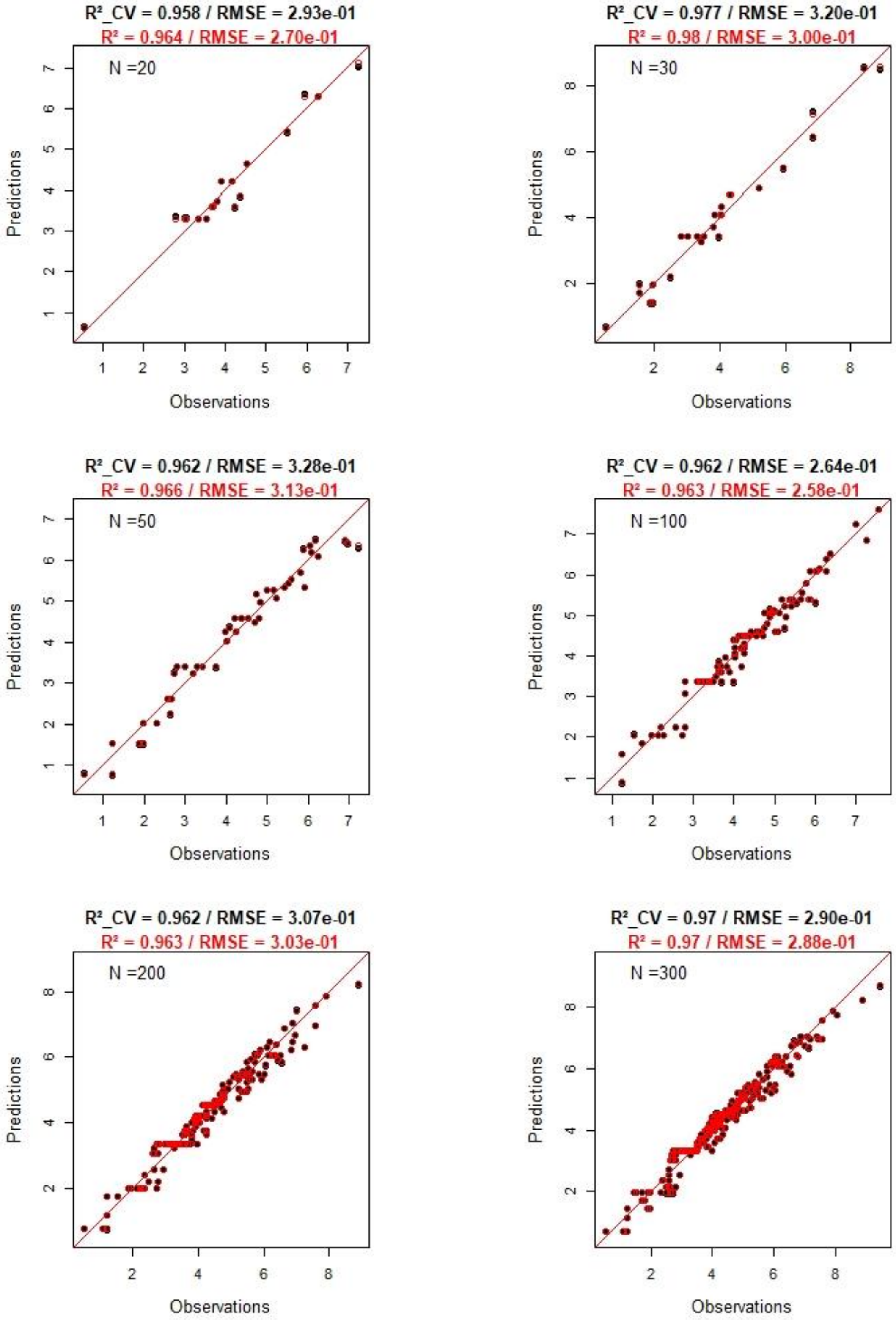


Figure 24: Scatter plots of predicted and observed values for the surface variable, at time step T = 5 and different sample size of the regression model; in red, points that served to construct the regression linear; in black, points used for the LOOCV. For each sample size N, two coefficients are estimated: the coefficient of determination of the regression model (R²) and the root mean squared error (RMSE), for both regression model (in red) and the LOOCV (in black).

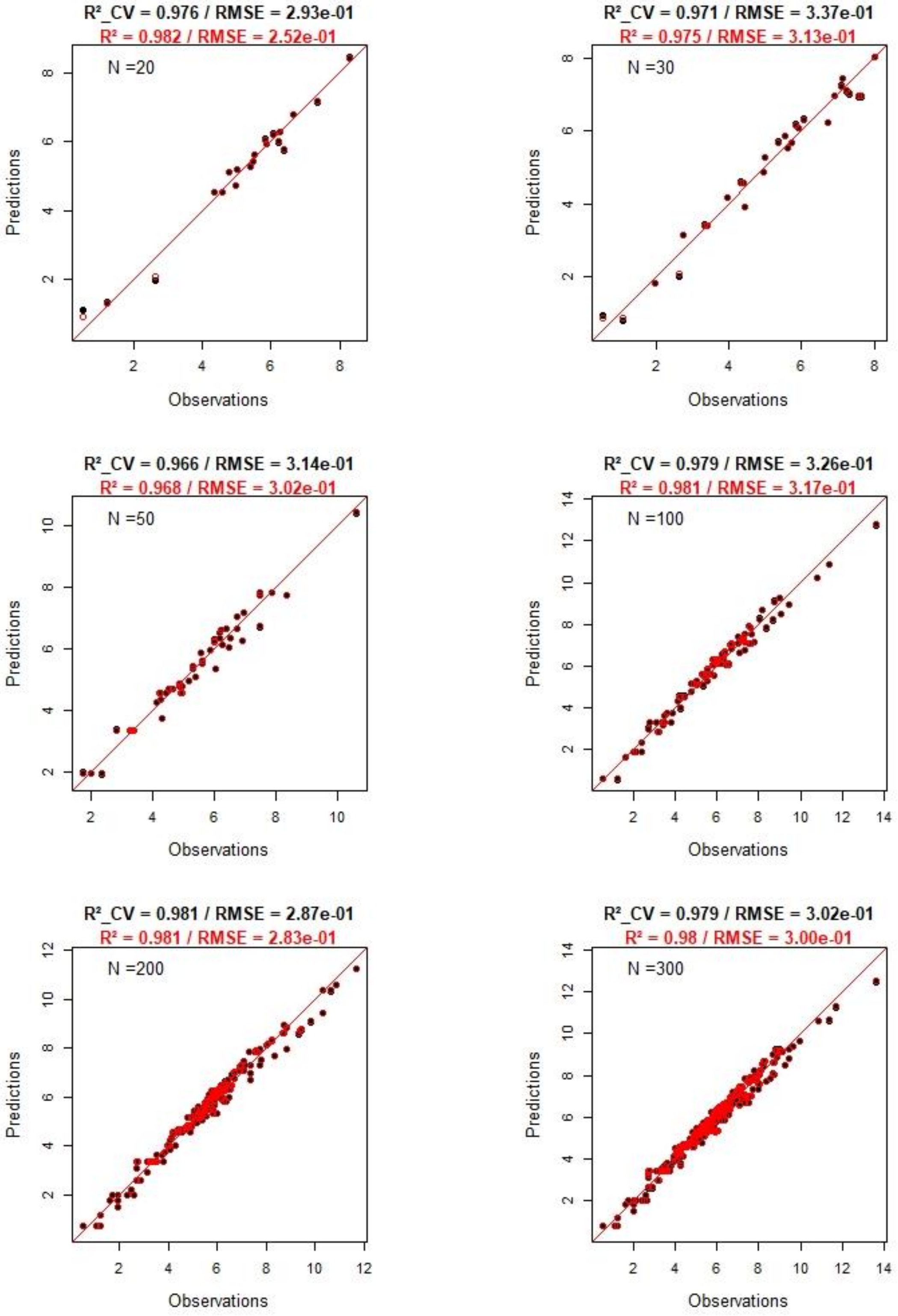


Figure 25: Scatter plots of predicted and observed values for the surface variable, at time step T = 10 and different sample size of the regression model; in red, points that served to construct the regression linear; in black, points used for LOOCV. For each sample size N, two coefficients are estimated: the coefficient of determination of the regression model (R²) and the root mean squared error (RMSE), for both regression model (in red) and the LOOCV (in black).

After having estimated the quality indicators of the regression model built between coarse-mesh (proxy) and refined-mesh model (exact) for the two capacity estimator (mass or surface), we now focus on their associated confidence and prediction intervals. Confidence and prediction intervals can also be derived from the linear regression procedure: in our case for a given coarse-mesh capacity estimator value, the confidence interval provides a range for the mean value of the refined-mesh capacity estimator: the computation of confidence interval implies the use of the standard error of the fitting, accounting for the uncertainty due to the sampling. The prediction interval provides a range for the value of the refined-mesh capacity itself and therefore also accounts, in the standard error of prediction, for the variability of the different samples around the estimated mean. The prediction interval is therefore larger than the confidence interval.

The formula for confidence and prediction intervals are the following:

Confidence interval: $\text{estimation of the mean value } \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)},$

Prediction interval: $\text{estimation of the mean value } \pm t_{\alpha/2,n-2}\sqrt{MSE\left(1+\frac{1}{n}+\frac{(x_k-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)},$

with $MSE = RMSE^2$ (mean square error), $t_{\alpha/2,n-2}$ is the t-value with n-2 degrees of freedom.

Figure 26 to Figure 28 present, for different learning set sizes N and different time step T, the constructed regression models of the mass variable and their 95% associated confidence (in blue) and prediction (in red) intervals. Same results are presented for the surface variable in Figure 29 to Figure 31:

- At a given time step T, larger sample sizes N will decrease the uncertainty due to the sampling, and result logically in narrower confidence and prediction intervals.
- A diminution of the confidence interval width is also observed when the time steps increase; this is particularly visible for the surface variable for which confidence and prediction intervals are larger for the smallest time steps. This is line with the better quality of the linear model observed for the latest time steps.
- It can also be observed that confidence and prediction intervals are relatively narrower for the mass variable than for the surface one, which is in line with the difference of quality of the linear models built for these two variables observed previously.

We recall that for methodological purposes, we have at disposal, together with the 540 results of simulations run with the proxy model, the same results for the simulations run with the exact model. Therefore (it would not be possible in a study with a dynamic model requiring very large computational time), we can also estimate confidence intervals (to represent the sampling error) with the bootstrap technique. The results are presented for different learning set sizes N, at different time step T, on Figure 32 to Figure 34 for the mass variable , and on Figure 35 to Figure 37 for the surface variable. Here also, we can observe that the higher the sample size N, or the larger the time step T is, the narrower the confidence intervals. Moreover, the 95% confidence intervals estimates for the mass variable are narrower than those obtained for the surface variable : at a given time step (T=2 and T=10), 95% confidence intervals are compared for different sample sizes N, respectively, for the mass on Figure 38, and for the surface on Figure 39.

Thanks to the different tests on the linear regression model, and on the quality estimation procedure, we suggest that the use of the results of the simulations run with 30 realizations of input parameters are sufficient to obtain a satisfactory error model for our application test case, allowing a good approximation and prediction quality as well as relatively narrow confidence and prediction intervals.

In the following section of the report, an error model is therefore constructed based on a subset of N=30 of proxy and exact responses, meaning that only 30 exact simulations are run.

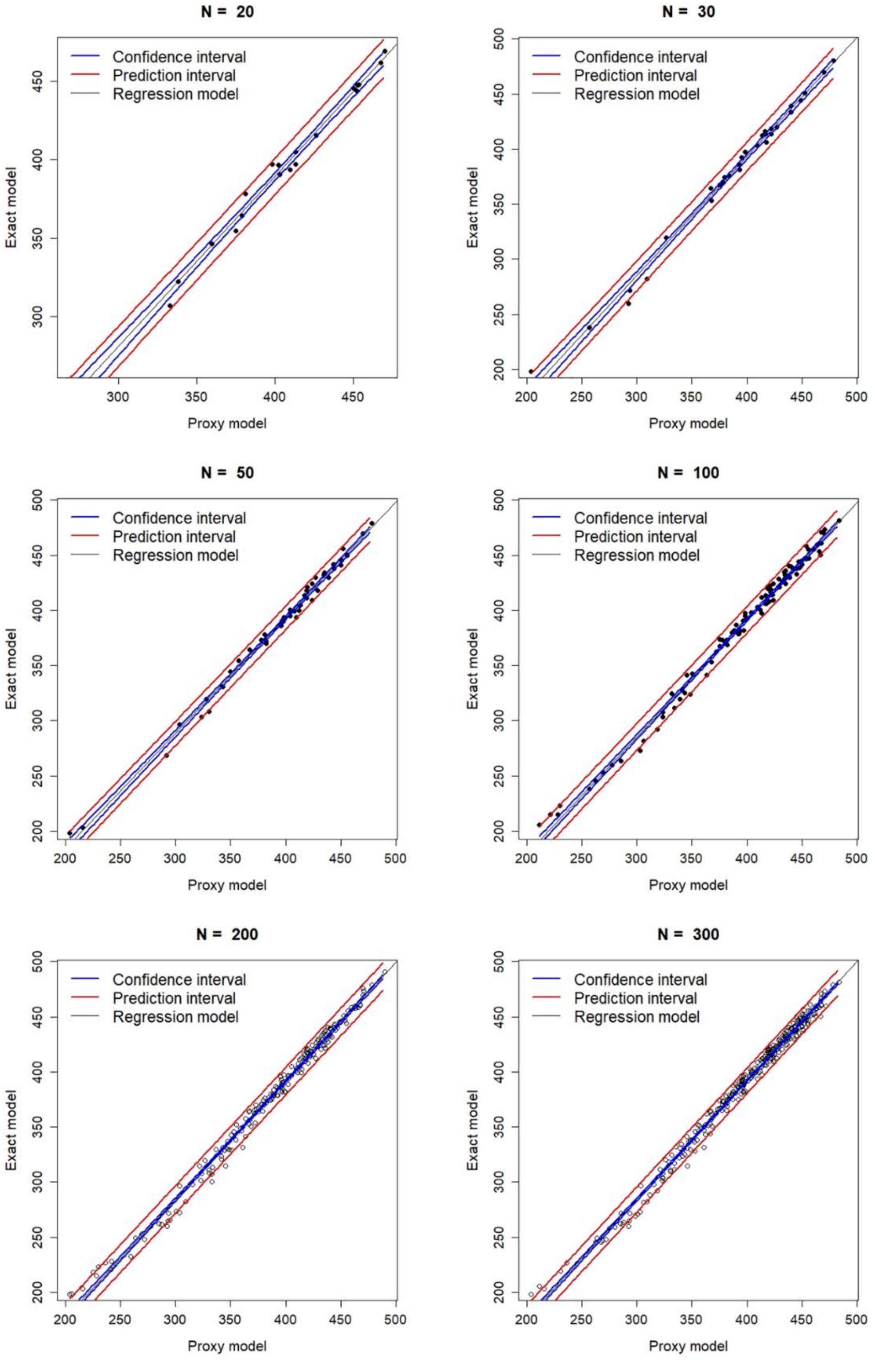


Figure 26: Regression model constructed between proxy and exact models for the mass variable, with different sample sizes N at time step T = 2. Regression model is le the black line and 95% confidence and prediction intervals are represented, respectively, by blue and red lines.

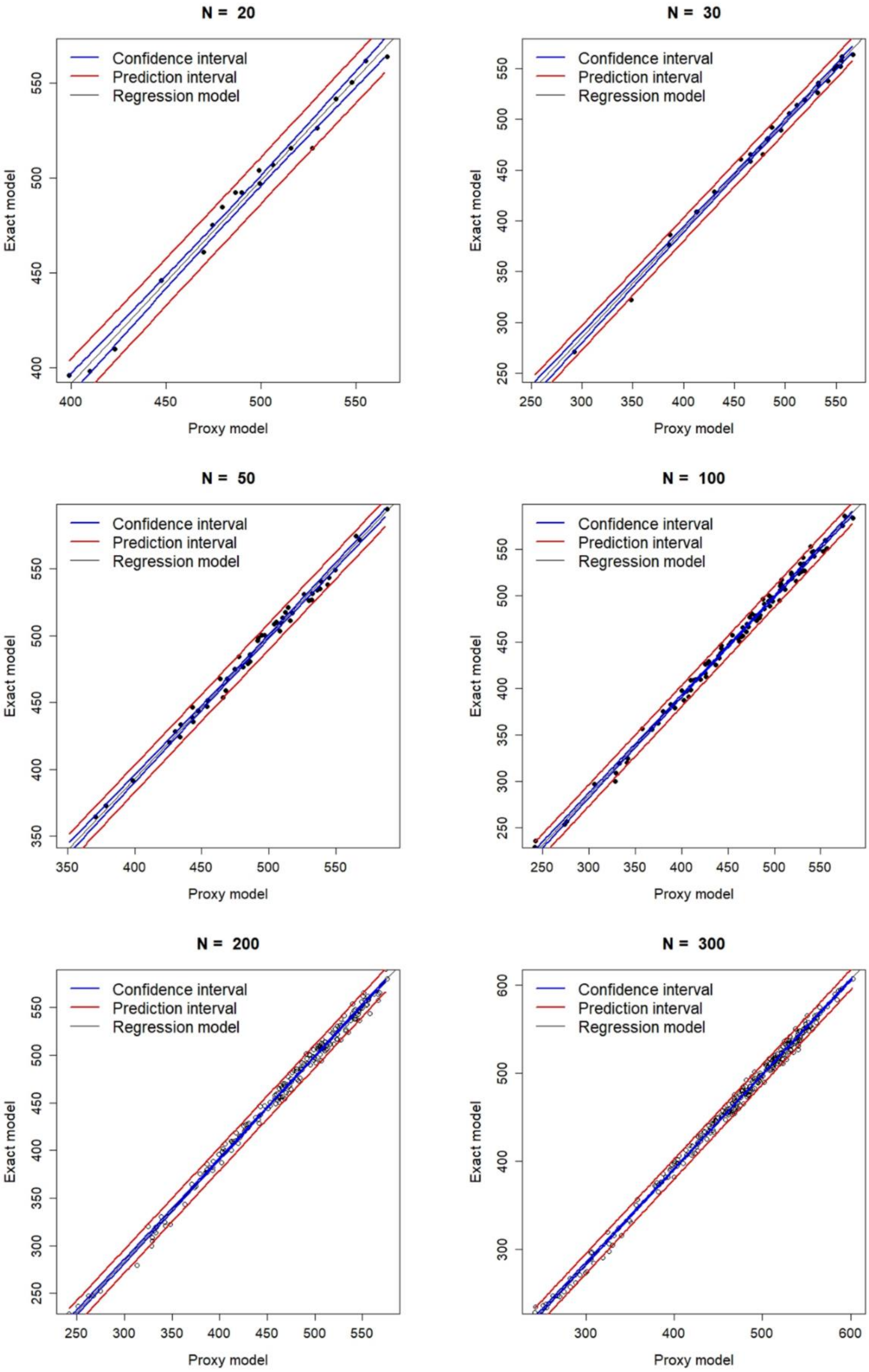


Figure 27: Regression model constructed between proxy and exact models for the mass variable, with different sample sizes N at time step T = 5. Regression model is le the black line and 95% confidence and prediction intervals are represented, respectively, by blue and red lines

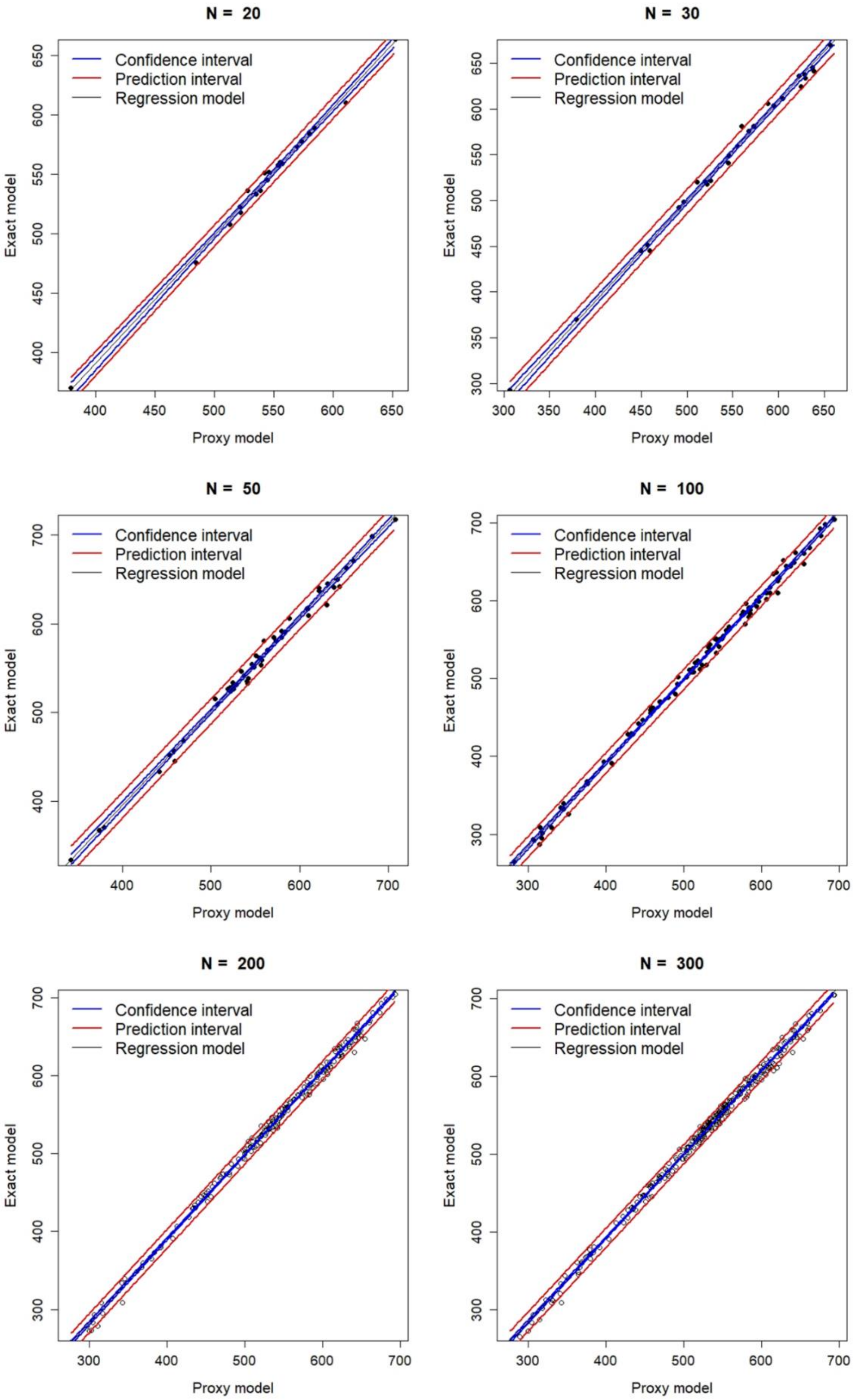


Figure 28: Regression model constructed between proxy and exact models for the mass variable, with different sample sizes N at time step T = 10. Regression model is le the black line and 95% confidence and prediction intervals are represented, respectively, by blue and red lines.

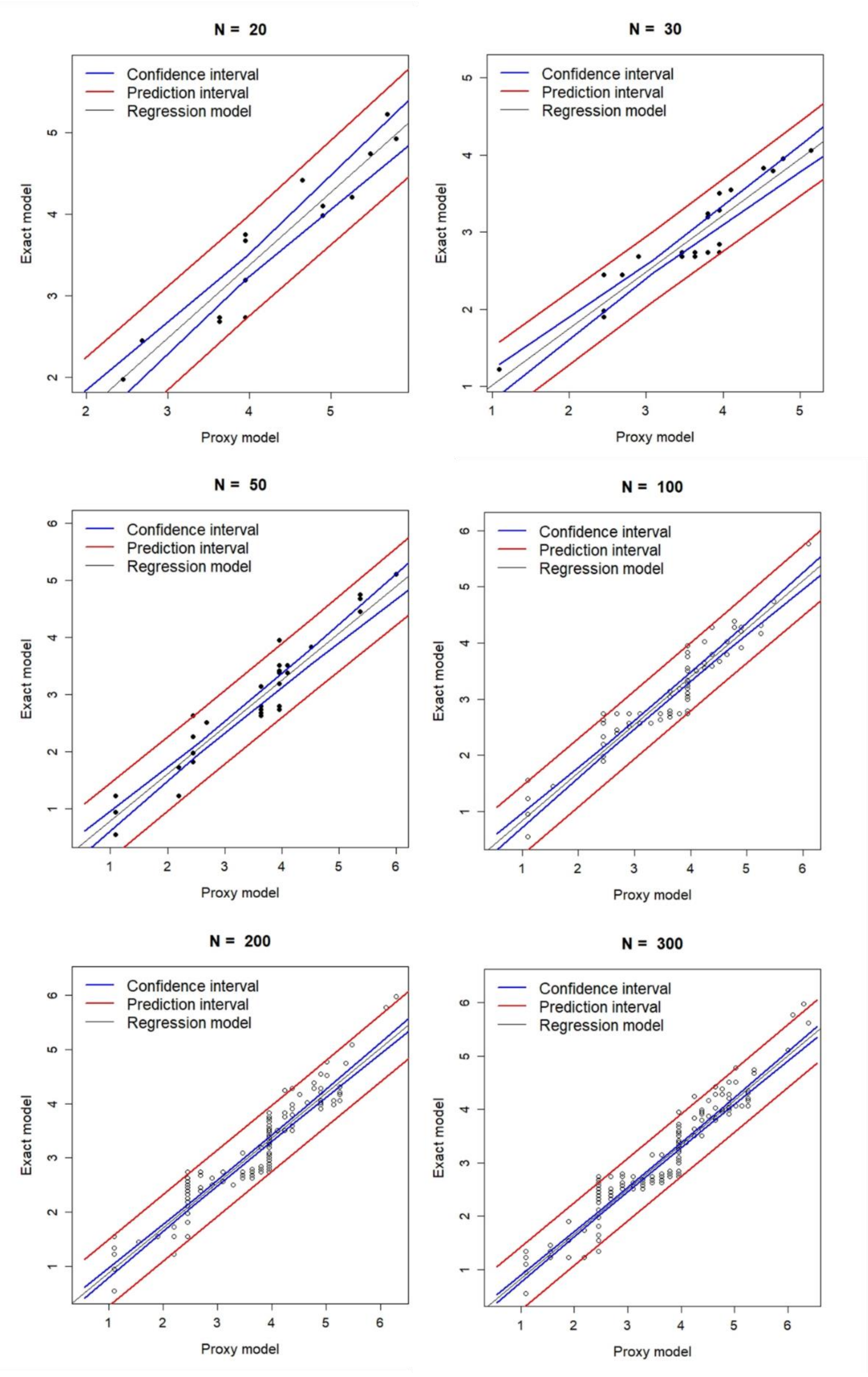


Figure 29: Regression model constructed between proxy and exact models for the surface variable, with different sample sizes N at time step T = 2. Regression model is le the black line and 95% confidence and prediction intervals are represented, respectively, by blue and red lines.

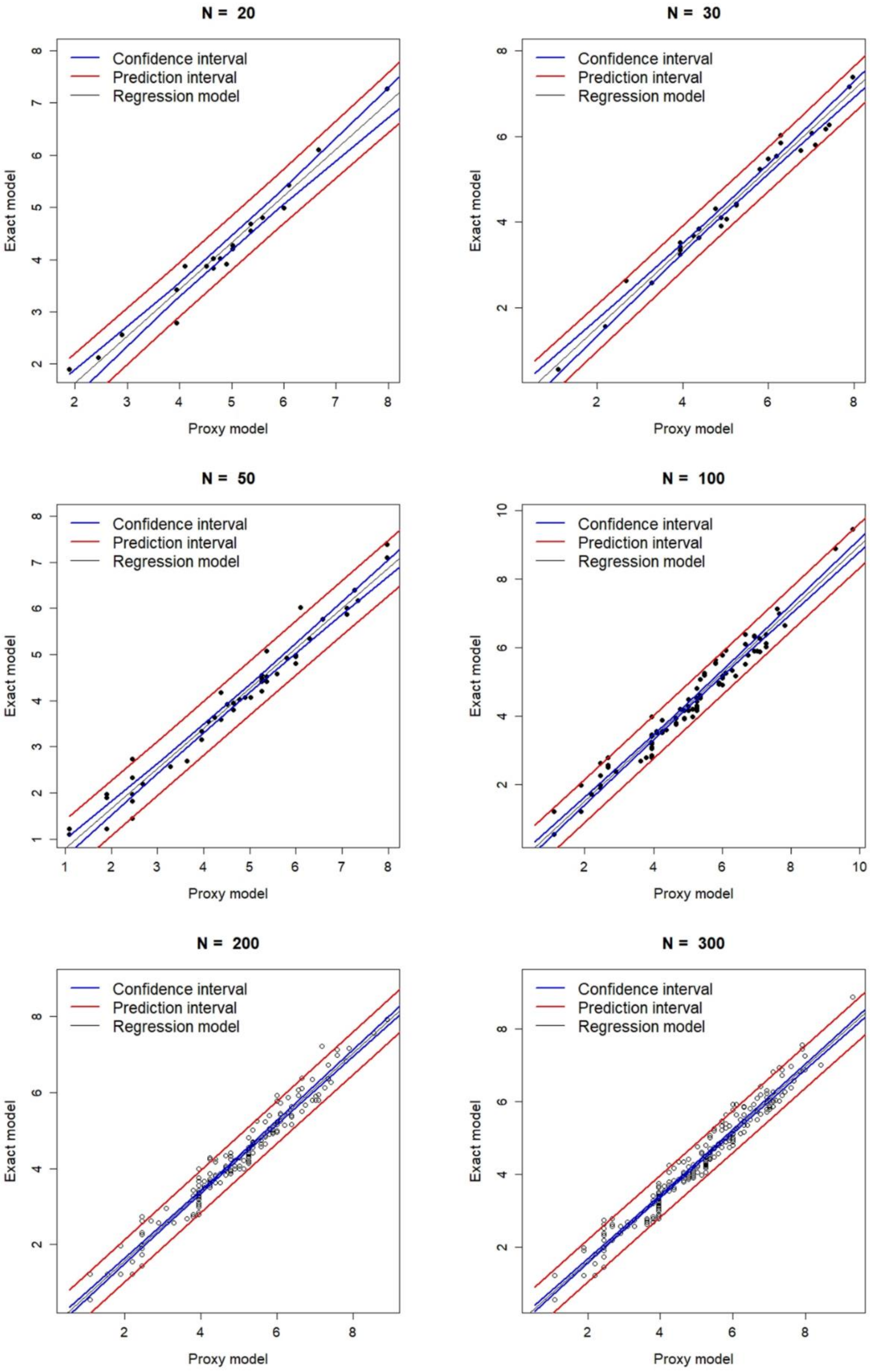


Figure 30: Regression model constructed between proxy and exact models for the surface variable, with different sample sizes N at time step T = 5. Regression model is le the black line and 95% confidence and prediction intervals are represented, respectively, by blue and red lines.

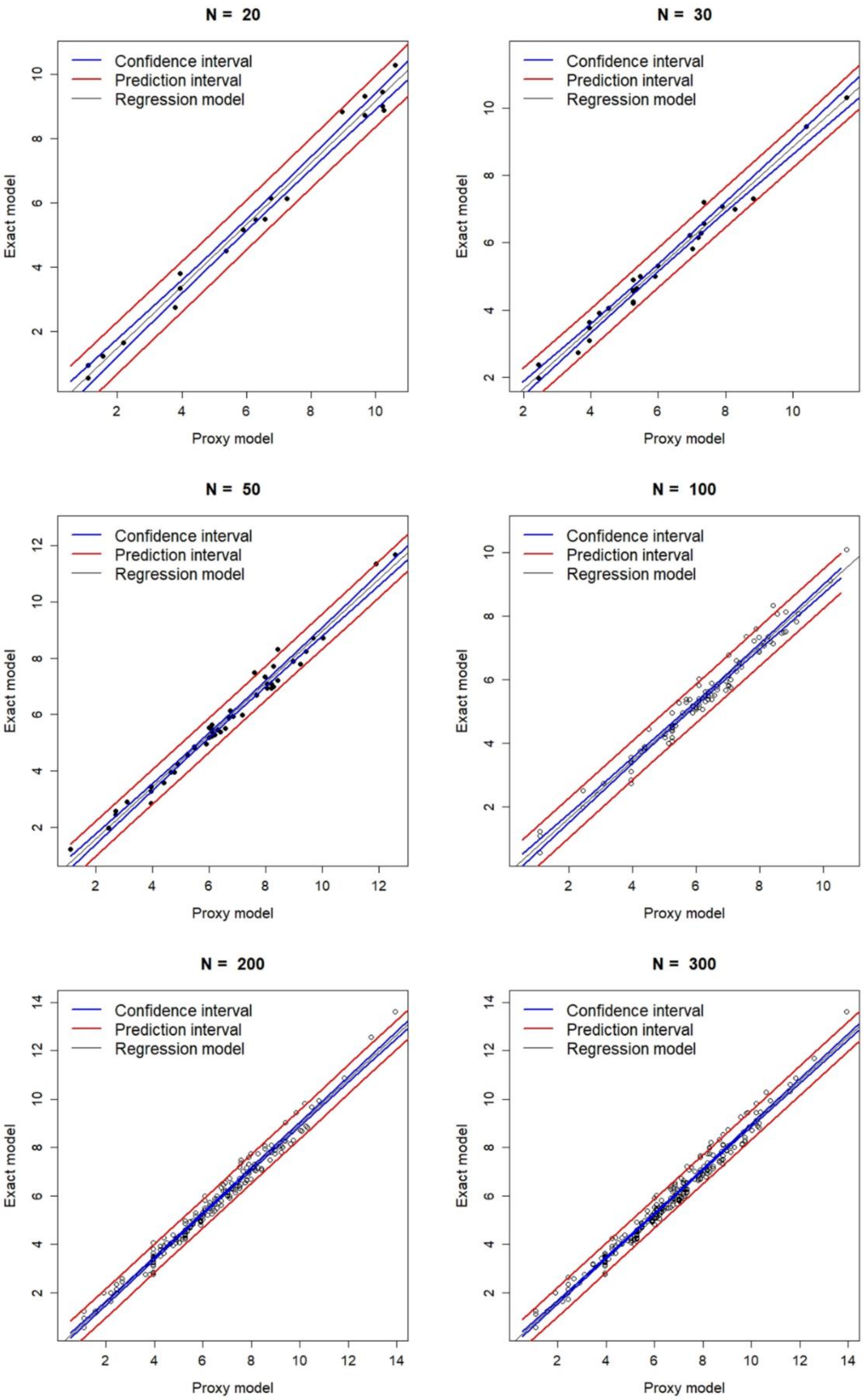


Figure 31: Regression model constructed between proxy and exact models for the surface variable, with different sample sizes N at time step T = 10. Regression model is le the black line and 95% confidence and prediction intervals are represented, respectively, by blue and red lines.

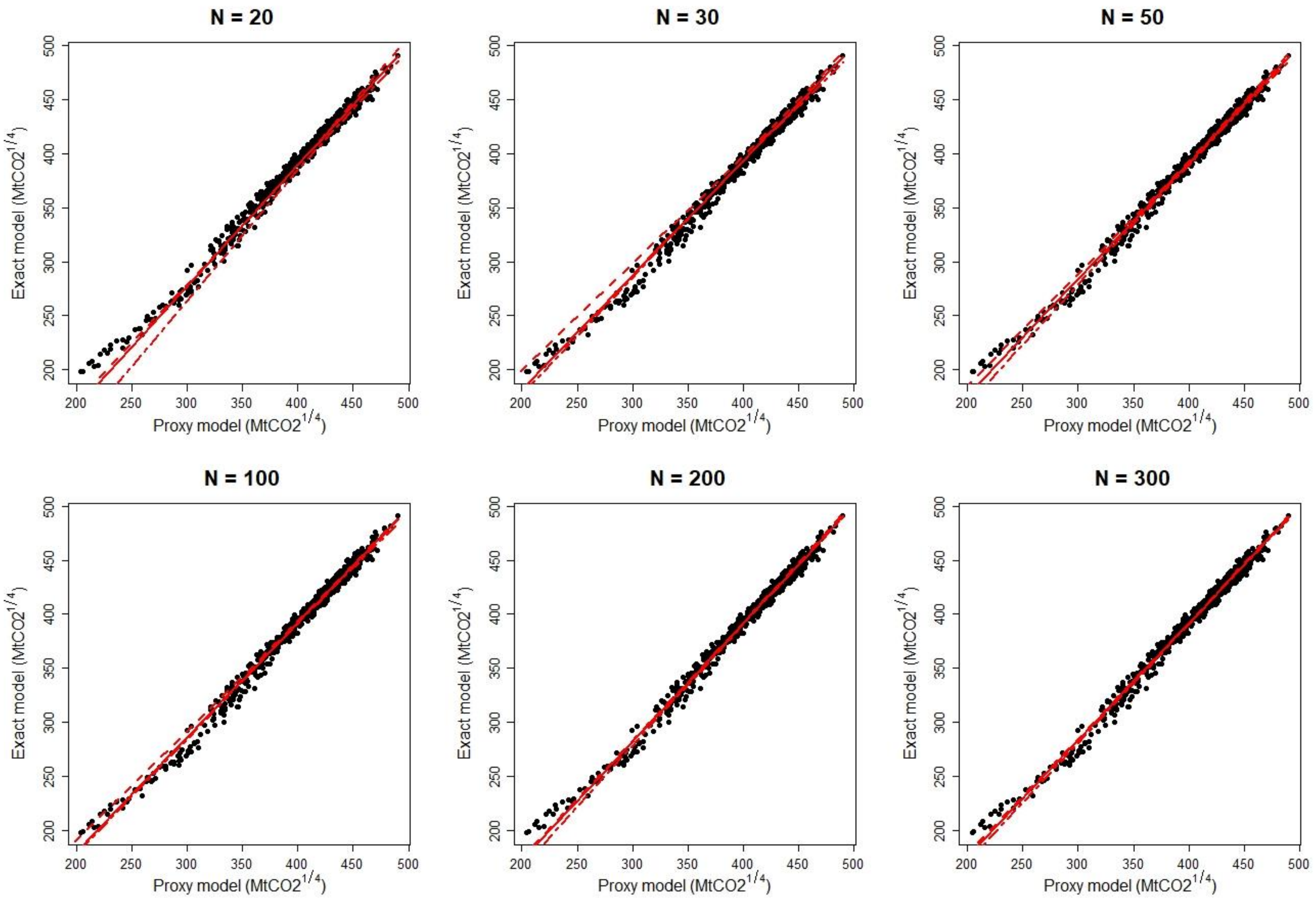


Figure 32: Regression model between proxy and exact simulations constructed for the mass variable and 95% confidence interval associated, estimated with 2000 bootstrapped coefficients at different sample size N and time step T = 2.

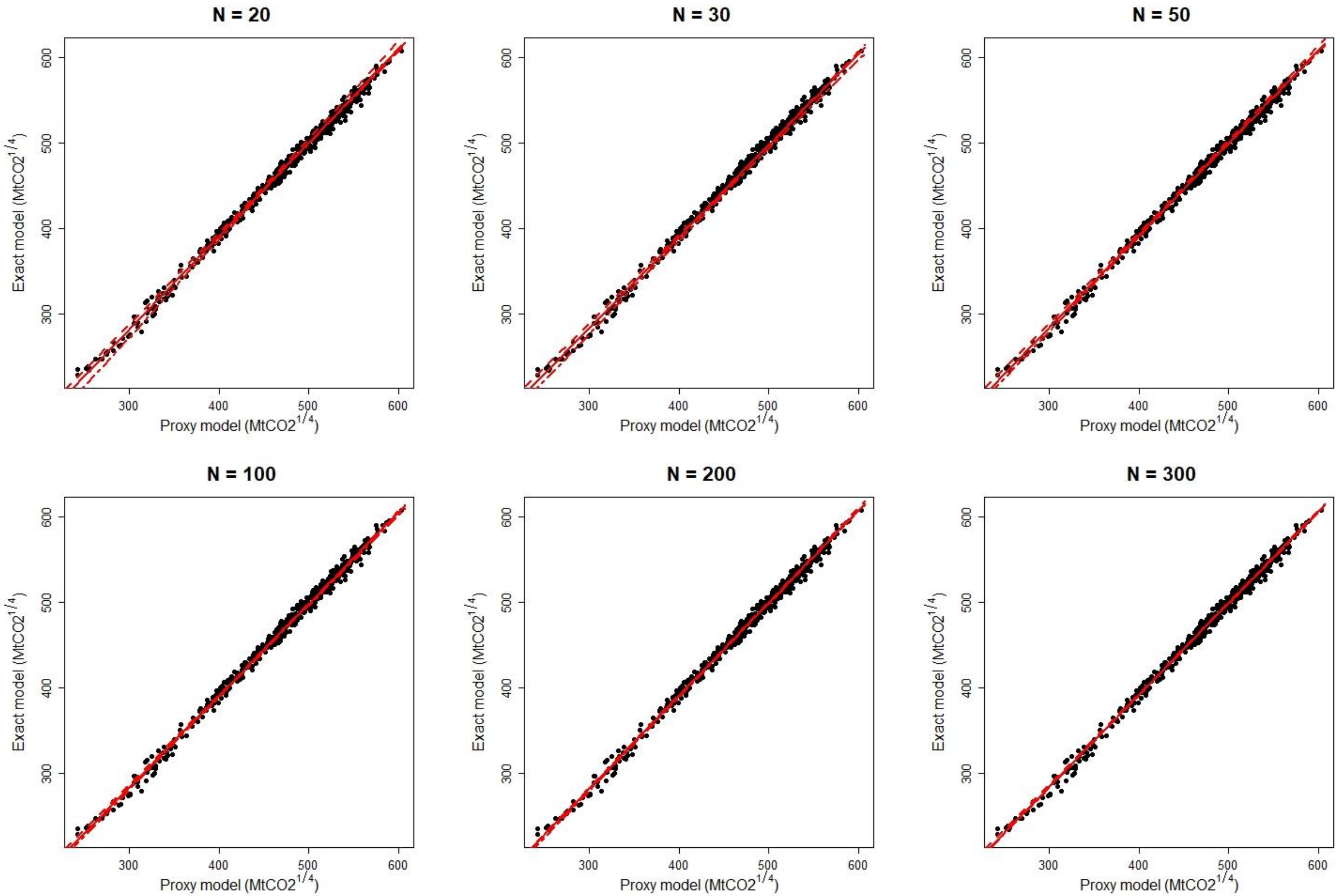


Figure 33: Regression model between proxy and exact simulations constructed for the mass variable and 95% confidence interval associated, estimated with 2000 bootstrapped coefficients at different sample size N and time step T = 5.

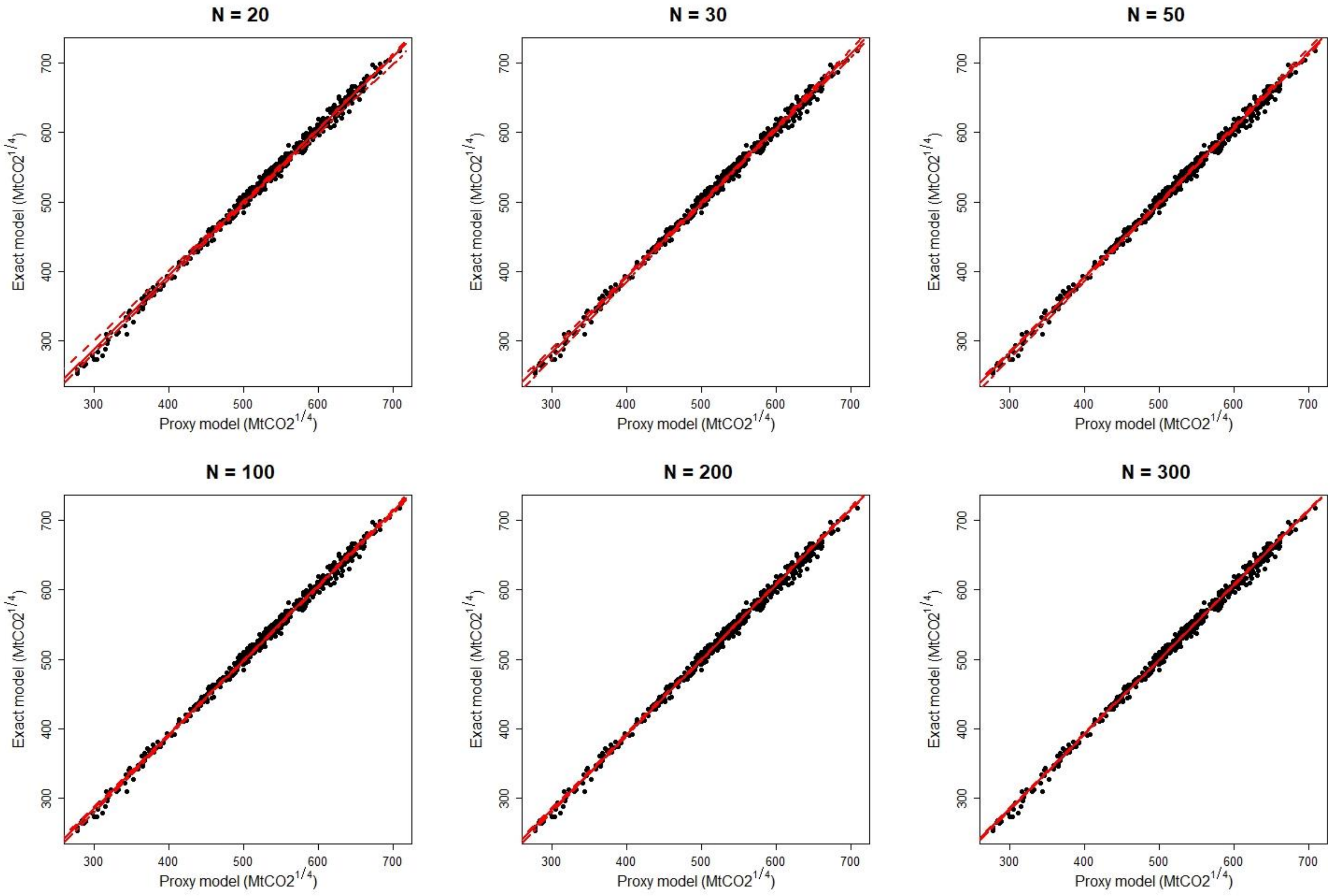


Figure 34: Regression model between proxy and exact simulations constructed for the mass variable and 95% confidence interval associated, estimated with 2000 bootstrapped coefficients at different sample size N and time step T = 10.

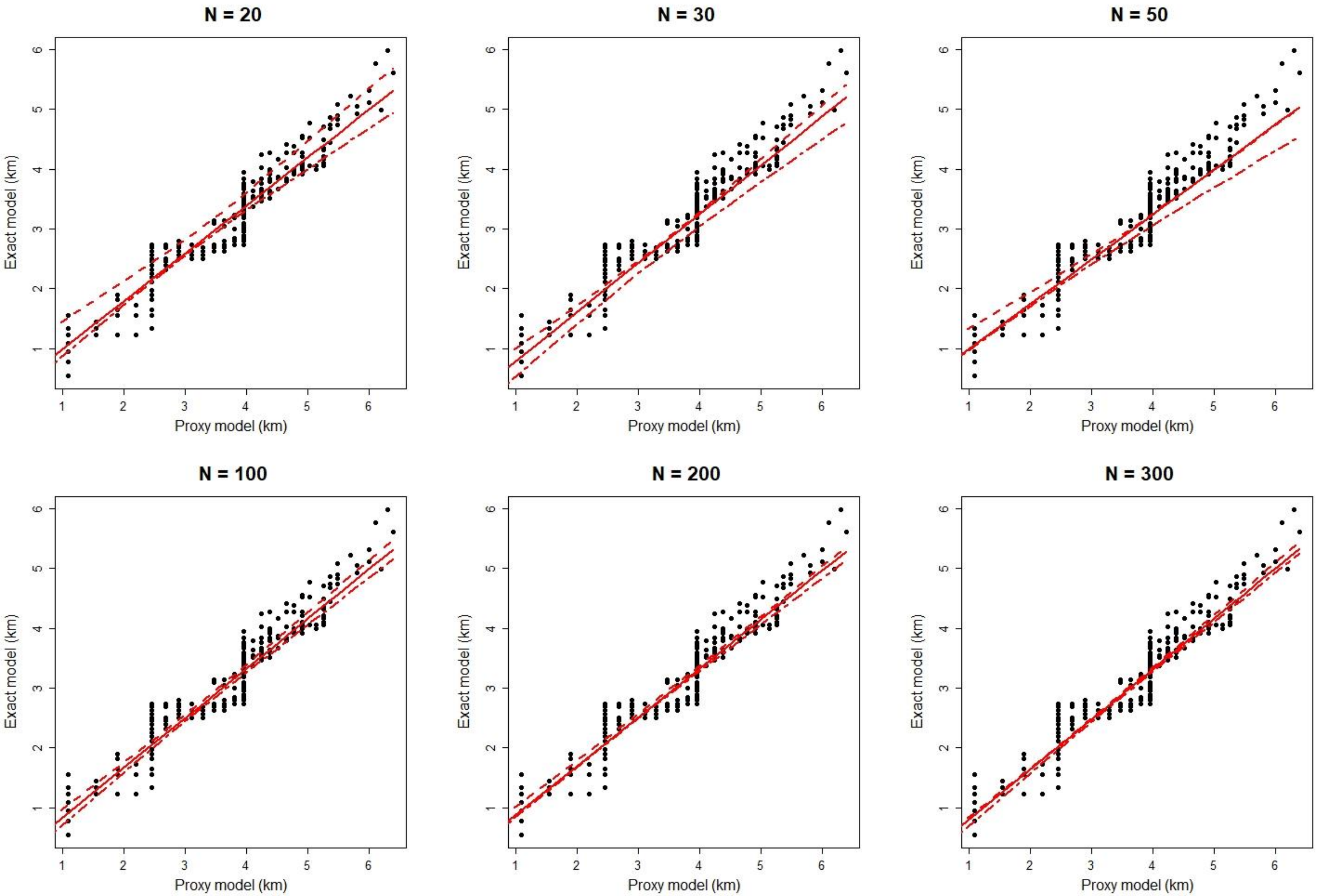


Figure 35: Regression model between proxy and exact simulations constructed for the surface variable and 95% confidence interval associated, estimated with 2000 bootstrapped coefficients at different sample size N and time step T = 2.

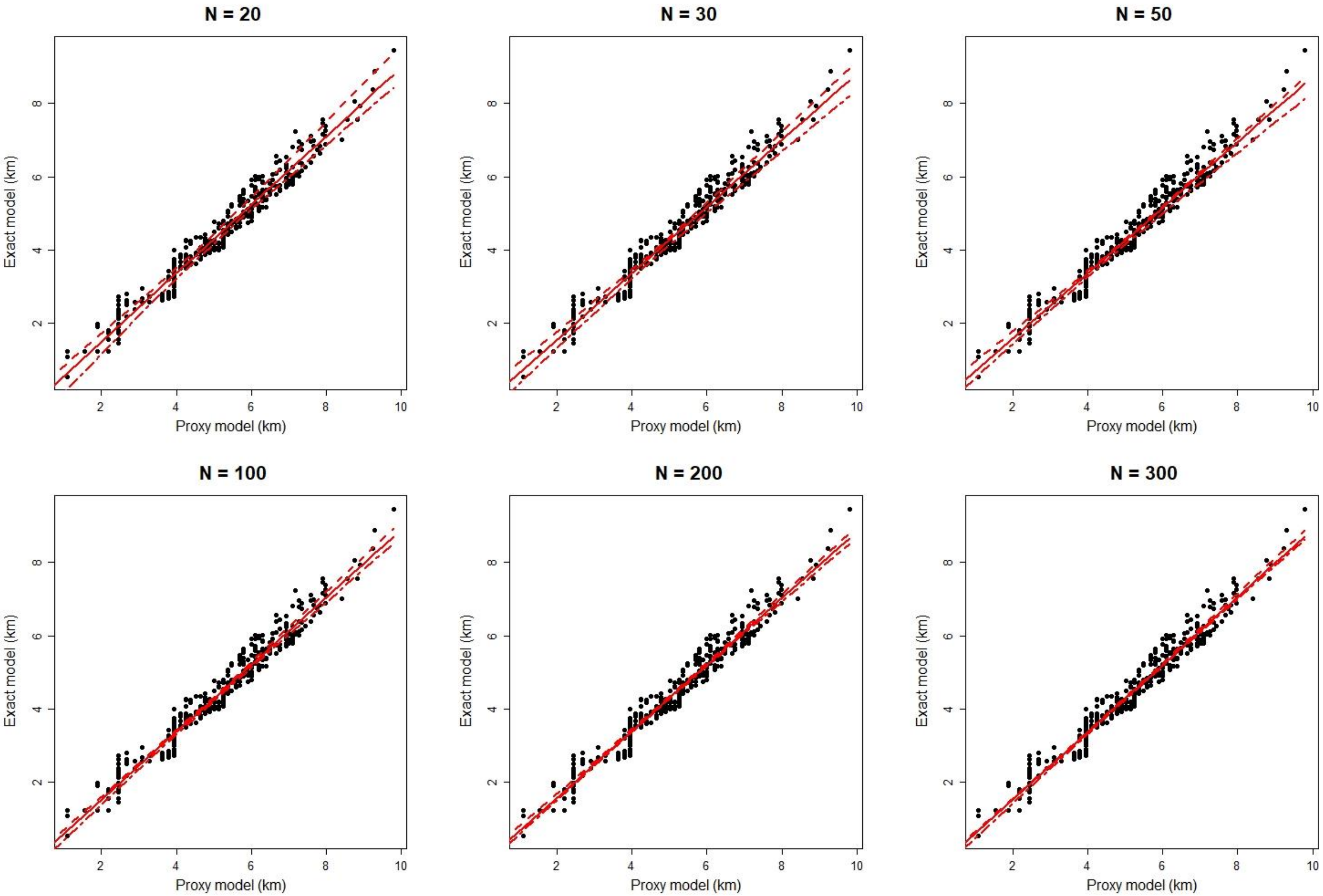


Figure 36: Regression model between proxy and exact simulations constructed for the surface variable and 95% confidence interval associated, estimated with 2000 bootstrapped coefficients at different sample size N and time step T = 5.

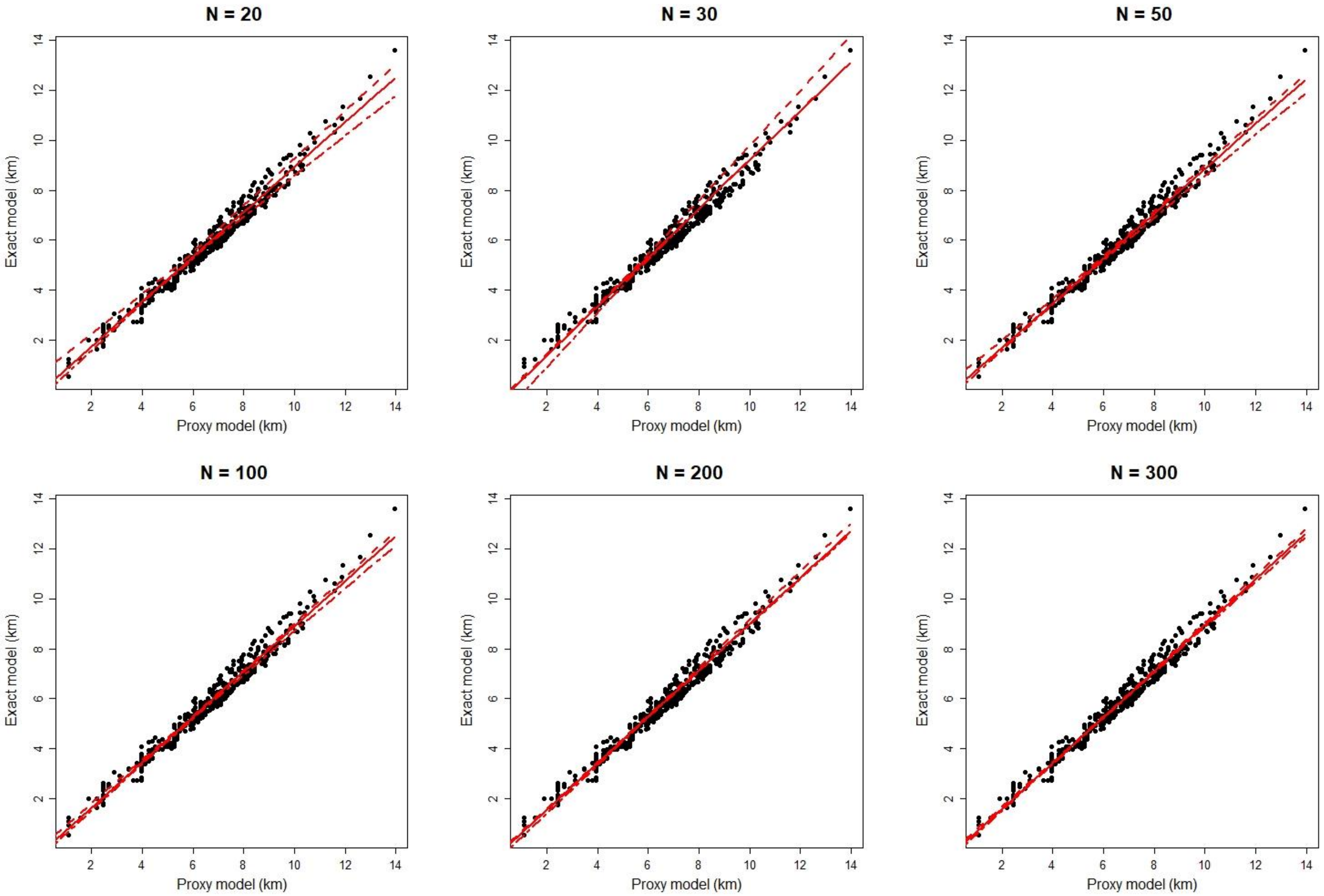


Figure 37: Regression model between proxy and exact simulations constructed for the surface variable and 95% confidence interval associated, estimated with 2000 bootstrapped coefficients at different sample size N and time step T = 10.

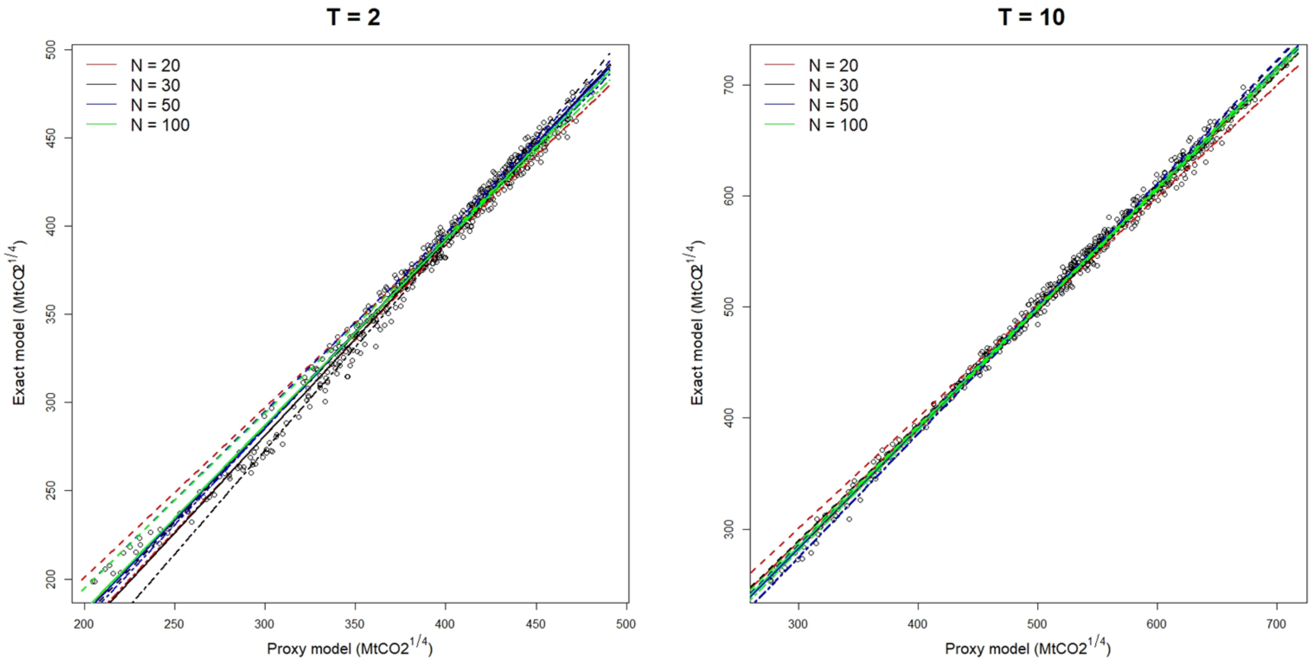


Figure 38: Regression model between proxy and exact simulations constructed for the mass double square root and different sample size N with their 95% confidence interval in dash line, 2000 bootstrapped coefficients, at time step t = 2 on the left and T = 10 on the right.

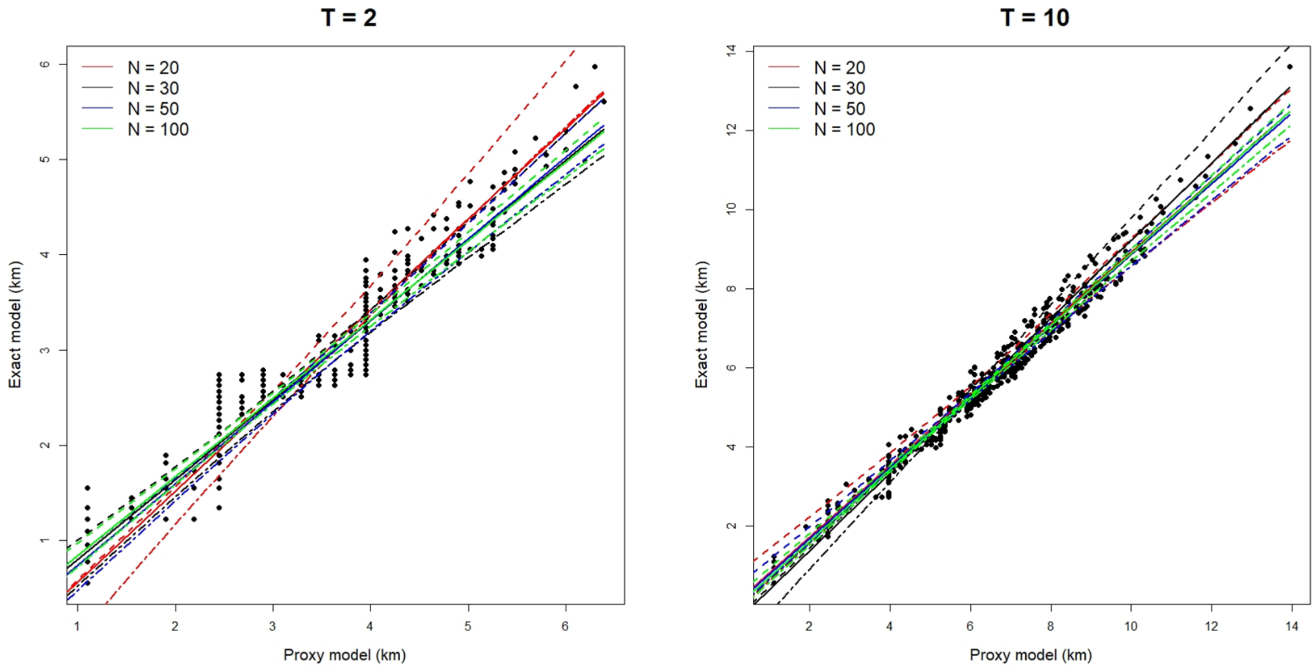


Figure 39: Regression model between proxy and exact simulations constructed for the surface square root and different sample size N with their 95% confidence interval in dash line, 2000 bootstrapped coefficients, at time step t = 2 on the left and T = 10 on the right.

6.3.3 Uncertainty propagation (UQ): estimation of quantiles of the capacity estimators

During the previous step of the methodology, an error model has been constructed with the objective of "correcting" the proxy model (dynamic model built from a coarse grid) responses in order to predict exact model (dynamic model built from a refined grid, in horizontal and vertical directions) responses. The chosen error model is based on a learning set of 30 realizations.

Once this error model has been constructed, it is used to predict the responses expected from the exact model for a large number of realizations of the proxy model (540 simulations): the uncertainty propagation can be performed with this large number of predictions rather than with a large number of responses of the exact model that would have required a very significant computational time.

At the end of this process, the mass and surface variables:

- have been computed from the 540 simulations run with the low-fidelity model
- have been estimated from a correction applied with the error-model to the 540 simulations run with the low-fidelity model

In addition, we recall that for methodological purposes, these two variables have also been computed from 540 simulations run with the "exact" model, which will enable an assessment of the performance of the error model.

To compare the distributions of the mass and surface variables computed from the proxy model and the exact model, and predicted with the error model, the empirical cumulative distribution function are plotted, as well as QQ plots are pltted: the results are provided on Figure 40 and Figure 41 for three time steps (T = 2, 5, 10)

Moreover, as we want to better the statistical estimation, having an important sample of exact responses is an advantage for comparison, in our case we have 540 simulations run with the refined mesh (exact model).

We notice, in most situations, a very good match between the distribution obtained with the exact model and that obtained with the error model. Interestingly, a good match is obtained even when significant differences between the proxy and exact model responses exist: for instance, the mass variable at time step T=2 is underestimated by the proxy model; this underestimation is well compensated by the error model, leading to a good match between the predictions and the calculations made with the exact model. At time step T=10, the proxy model tends to overestimate the mass variable; here again, despite this deviation between the proxy and the exact model, the application of the error model leads to very satisfying predictions. The interest of the use of the error model on proxy model results to "correct" them is even more visible for the surface variable. The differences between the proxy and the exact model are significant, but the error model manages to greatly improve the proxy model responses, leading to very satisfying predictions.

As mentioned above, in the introductory section for instance, we are interested, at the end of the uncertainty propagation process, in assessing low, high and best estimates of two CO_2 capacity estimators times series, under the form of quantiles (P5, P50 and P95).

Sample quantiles, i.e. estimations of the population quantiles from a given set of observation is a common issue (see for instance Hyndman and Fan, 1996), relatively closely linked to the so-called plotting position formula (PPF) to be used in a quantile plot. Several rules can be followed for computing sample quantile:

- First of all, discontinuous quantile functions have been proposed: the quantile function is therefore a step function with different alternatives existing to account for each jump, or for the placement of each step.
- Continuous quantile functions have also been proposed: all these functions are made of linear interpolation between plotting positions. The different existing PPF (see for instance a review in Rosbjerg et al., 1992) lead to a relative diversity in terms of continuous quantile functions:
    - Some formula are based on the sample cumulative frequency; the plotting position proposed for rank k are for instance $p_k = \frac{k-1}{n}$ or $p_k = \frac{k}{n}$ (California formula), or $p_k = \frac{k-0.5}{n}$ (Hazen formula).

- Others are based on the distribution of the sample cumulative frequency P_k; if the mean of P_k is used as plotting position, $p_k = \overline{p_k} = \frac{k}{n+1}$ (Weibull formula), if the mean is used, $p_k = \widehat{p_k} \approx \frac{k-0.31}{n+0.38}$ (Beard formula), and if the mode is used $p_k = \widetilde{p_k} = \frac{k-1}{n-1}$ (modal formula).
- Finally some formula are based on the distribution of the order statistics. Each distribution requires its own formula. For example, for a normal distribution of the order statistics, $p_k = \frac{k-3/8}{n+1/4}$.

The algorithm that allows a continuous sample quantile with the modal formula has been used in this study. Figure 42 and Figure 43 present the 5th, 50th and 95th percentiles and their 95% confidence interval (using bootstrap method), plotted against the sample size dimension, respectively, for the mass and the surface. Note that the quantiles are plotted as a function of the sample size to verify that the set of 540 simulations was sufficient for obtaining stable quantiles. As for the whole empirical distribution shown previously, the quantiles are estimated with the proxy model responses, with the exact model responses (for methodological purposes) and with the predictions obtained with the error model.

All the figures show a convergence of the computed quantiles obtained for 200 samples in average. As observed with the whole distribution plots, the gain obtained with the application of the error model is more visible when large deviations exist between the proxy responses and the exact model ones:

- for the mass variable, the most visible improvements are visible for the P05 and P95 notably at time step T=10;
- for the surface variable, the proxy model over-estimates all quantiles whatever the time step and an excellent estimate is obtained with the error model. The error-model therefore appears to be able to correct the approximate responses to predict quantiles close to the exact ones, despite the fact that, for that variable, the low-fidelity model significantly deviates from the exact ones.

Barplots representing the estimated quantiles with the full set of samples have also been produced to show, in a more synthetic way, the interest of applying the error model and to focus on confidence intervals. For these barpolots, we focus on the most important quantiles regarding CO_2 capacity estimation in a risk perspective: the P05 for the mass variable (risk of injecting a too low quantity of CO_2) and the P95 for the surface variable (risk of a too large footprint). These quantiles together with their confidence interval are shown on Figure 44 and Figure 45 for two time steps.

These figures show that the error model lead to a real improvement of the quantile estimation, leading to predicted quantiles very close to the quantiles computed with the exact model. One interesting observation is that while the quantile prediction seems accurate, the confidence interval regarding those quantiles can sometimes be inferior to the one computed with the exact simulations, leading to a too "optimistic" estimation.

The obtained results show the procedure ability to help deriving estimations of quantiles from a limited number of simulations. Another approach for estimating high or low quantiles, very much used for instance for nuclear safety assessment, is the Wilks' method. This method provides, for a limited amount of simulations, a majoring value of high quantiles (or minoring value of low quantiles), with a level of confidence β, i.e. an upper limit $\hat{q}_{\alpha,sup}$ that satisfies: $P\left(q_\alpha \leq \hat{q}_{\alpha,sup}\right) = \beta$, with q_α the αth percentile. This upper limit is taken equal to the $j(\alpha,\beta,n)$th value (rank) of the sample ordered in ascending order, with $j(\alpha,\beta,n)$ defined as the smallest integer satisfying the following equation:

$$\sum_{i=1}^{j(\alpha,\beta,n)-1} C_n^i \alpha^i (1-\alpha)^{n-i} > \beta$$

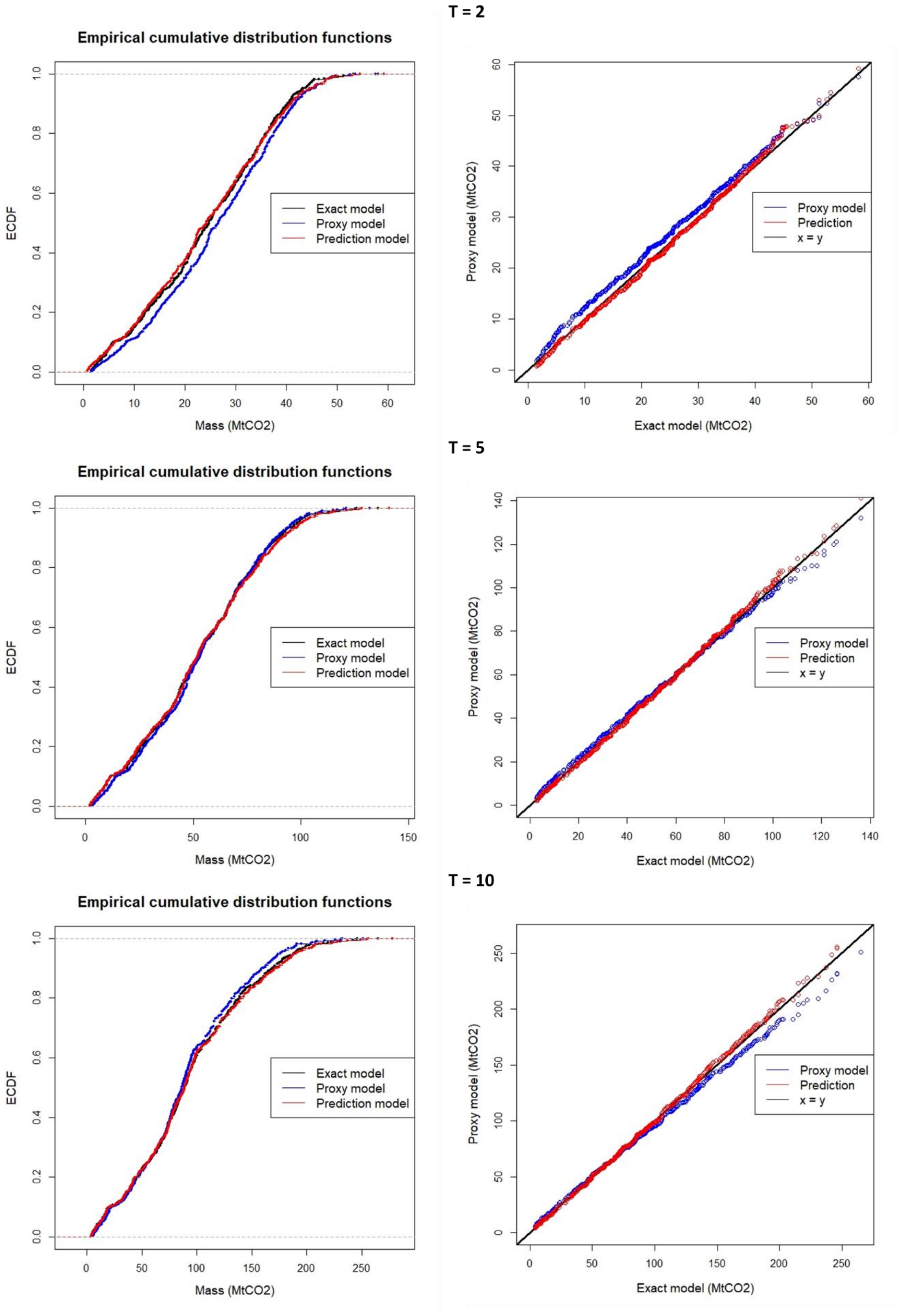
Table 4 provides, for a level of confidence of 95% and for different values of n, the value of the rank $j(\alpha,\beta,n)$ for determining the 95th percentile.

95th percentile	
n	$j(0.95, 0.95, n)$
59	59
93	92
124	122

Table 8: Value of the rank $j(\alpha, \beta, n)$ for determining the 95th percentile with a 95% level of confidence and for different sample size.

Note that, conversely, the Wilks' method can also be used to determine the number of simulations necessary to compute the upper limit of a quantile for given α, β and for a given rank $j(\alpha, \beta, n)$.

Therefore, coming back to our case study, this formula can provide a minimum number of simulations to be run for obtaining an upper limit of different quantiles: the estimation of an upper value for P95 with Wilks' formula requires at least 59 simulations (as shown in Table 8). This minimum number of simulations is higher than the 30 simulations run for applying the UQ procedure presented in this report.



T = 2

T = 5

T = 10

Figure 40: On the left: empirical cumulative distribution estimation for the mass at different time step; on the right: Q-Q plot of the mass at different time step T.

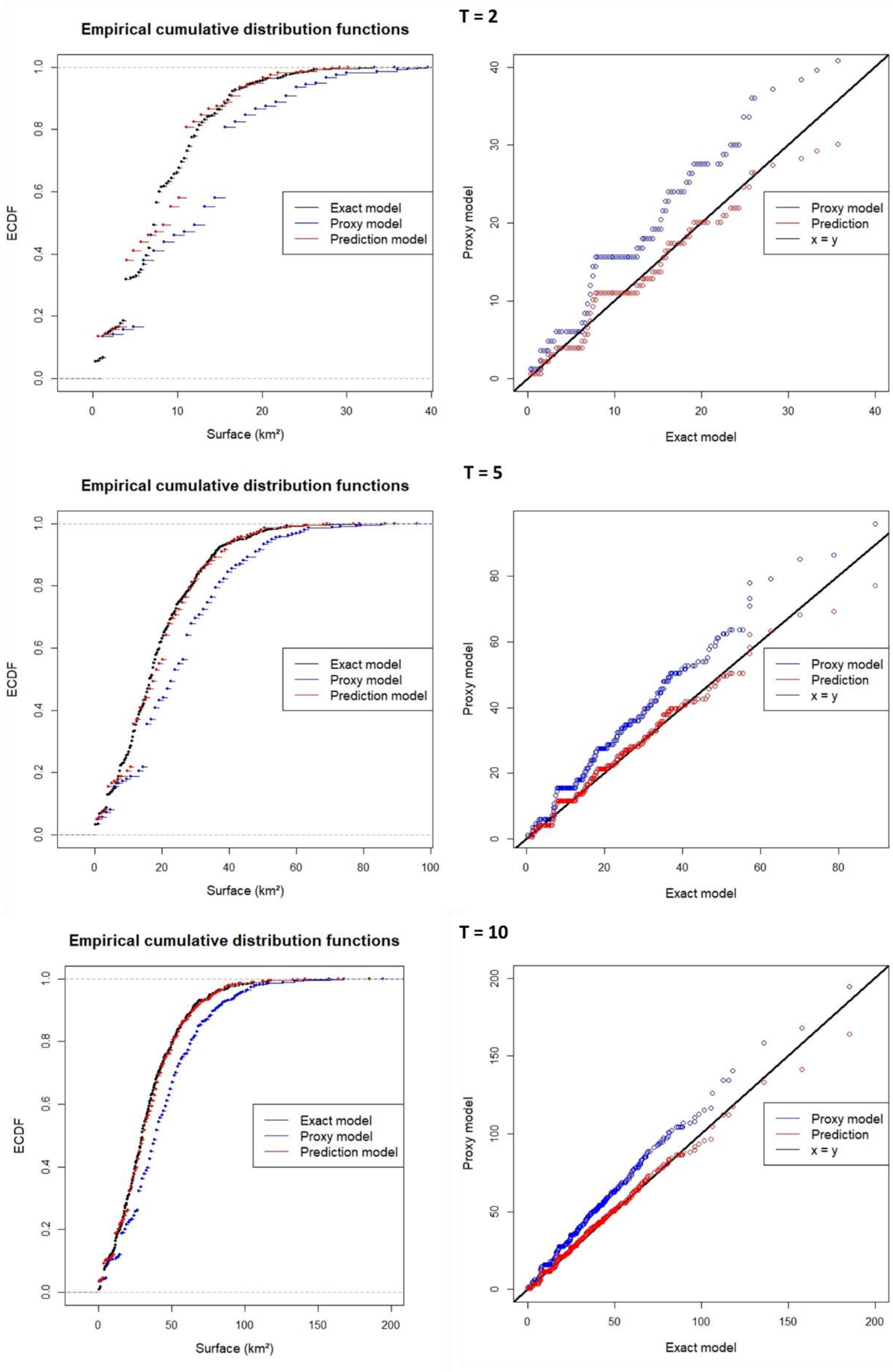


Figure 41: On the left: empirical cumulative distribution estimation for the surface at different time step T; on the right: Q-Q plot of the surface at different time step T.

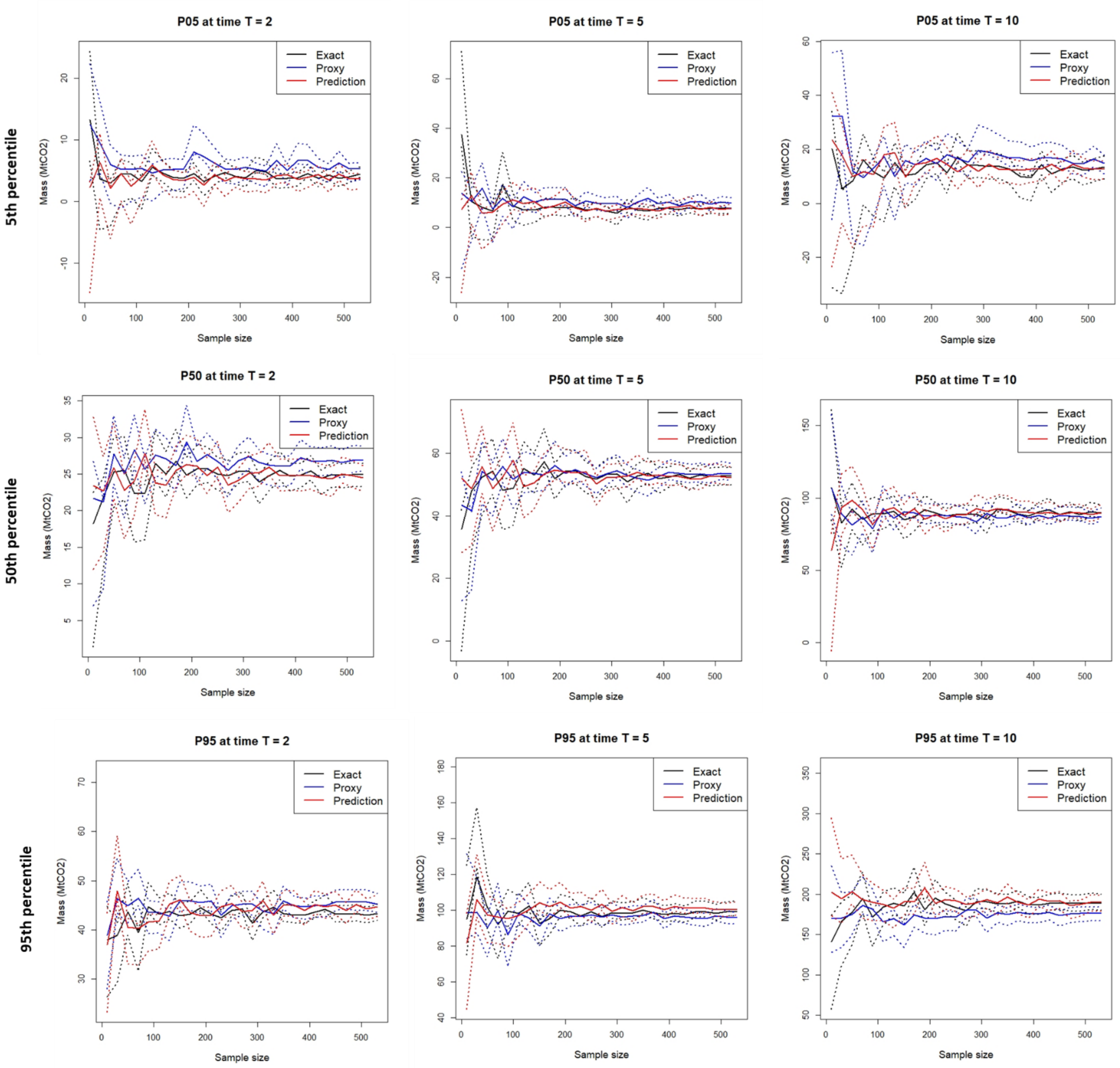


Figure 42: Convergence of the quantile estimation (5th, 50th, and 95th percentile) for the mass as a function of sample size; at different time step T.

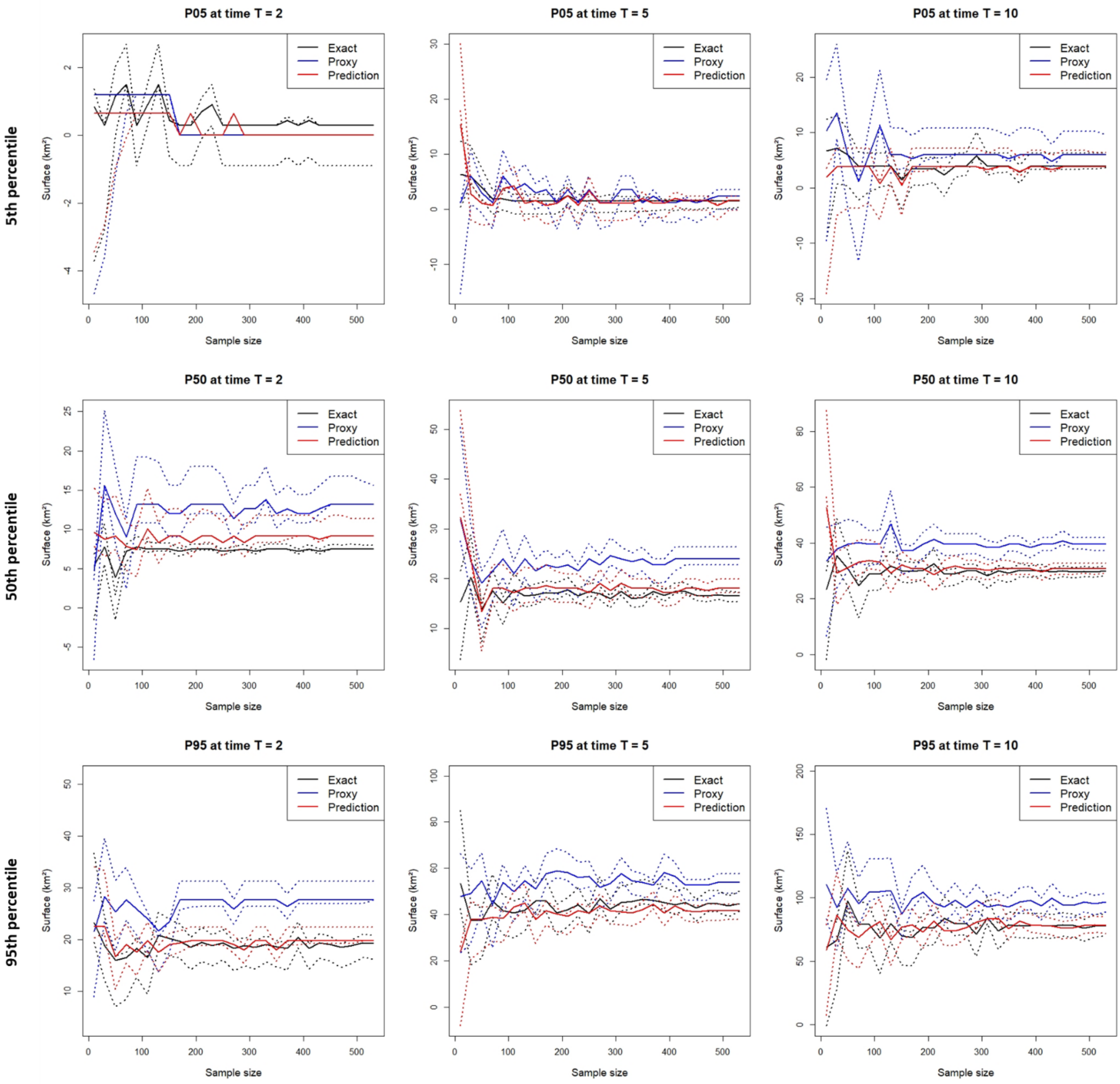


Figure 43: Convergence of the quantile estimation (5th, 50th, and 95th percentile) for the surface as a function of sample size; at different time step T.

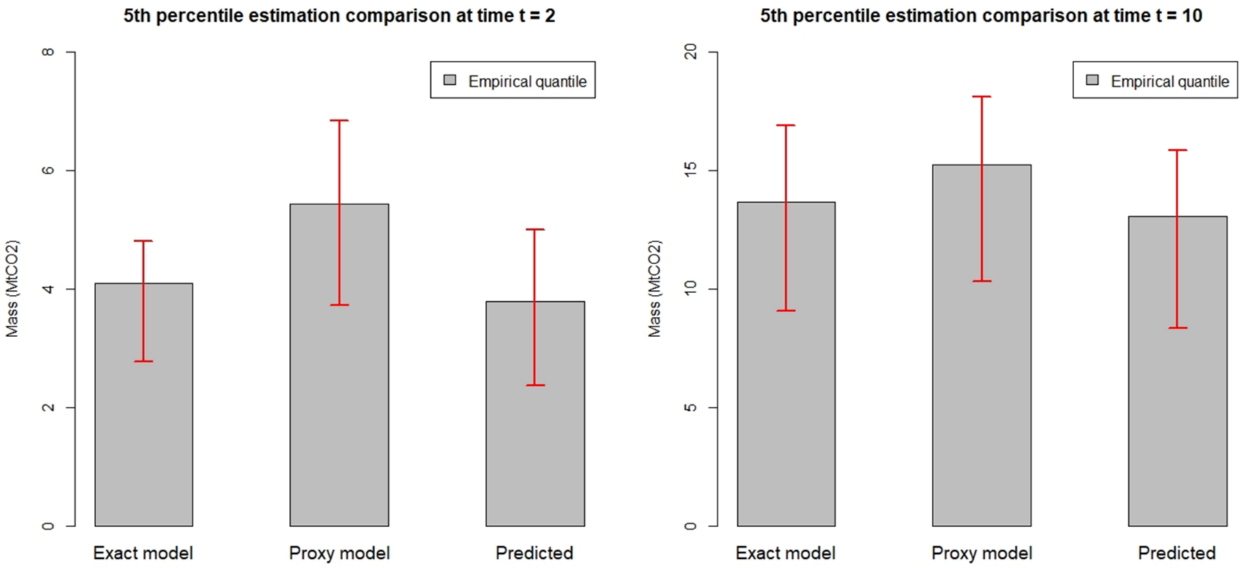


Figure 44: 5th percentile estimation of the mass with their 95% confidence (tolerance) intervals for the whole sample of realizations (i.e. 540 simulation responses with the proxy model for the “Proxy model” and “Predicted” bar plots; 540 simulation responses with the exact model for the “Exact model” bar plots); at different time step T.

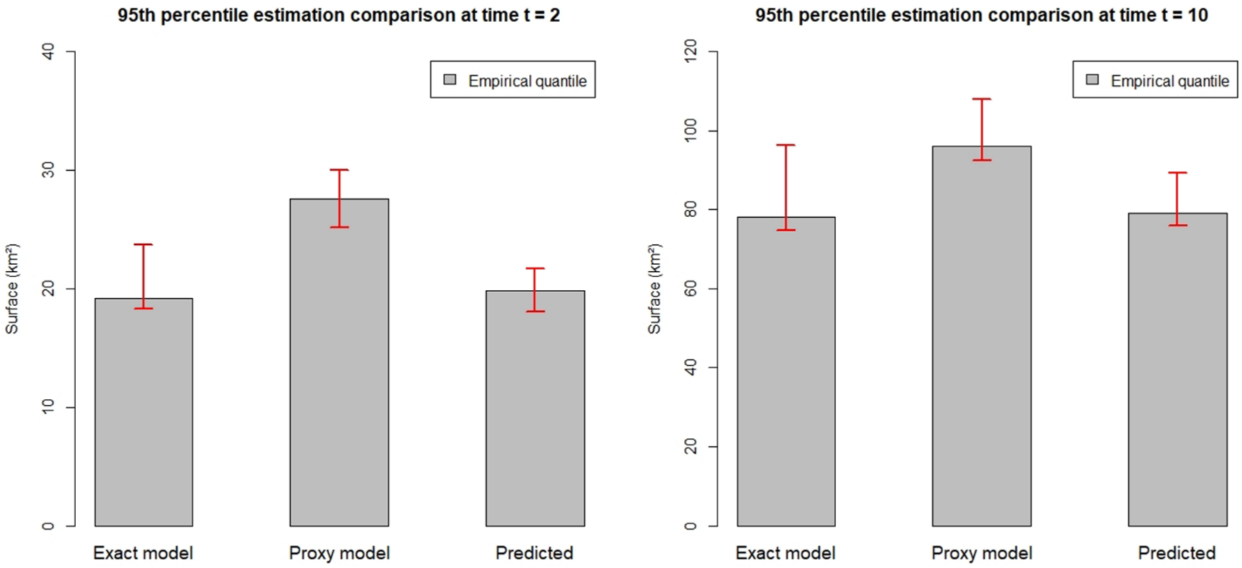


Figure 45: 95th percentile estimation of the surface with their 95% confidence (tolerance) intervals associated for the whole sample of realizations (i.e. 540 simulation responses with the proxy model for the “Proxy model” and “Predicted” bar plots; 540 simulation responses with the exact model for the “Exact model” bar plots); at different time step T.

6.4 Application of the UQ procedure on time series

In the previous section, the UQ procedure has been applied to each time steps, i.e. a new error model (regression) has been built for each variable (mass and surface) and each time step.

In this section, the same procedure is applied directly on the time series. The first step of the approach is to apply Principal Component Analysis (denoted PCA) on these time series, with the purpose of diminishing the dimension of the response spaces (while maximizing the retained information) and therefore the size of the regression problem. Then, regression models are built on the components coefficient (as much as number of dimensions kept in PCA Npca). Once constructed on a limited learning set of realizations, this error model can be used to predict exact responses from a large set of proxy simulation responses, projecting proxy responses on the PCA basis first.

Comparing to the scalar procedure which requires the construction of a regression model at each time step, the gain of the time series procedure is that it requires only Npca regression models to predict exact responses at any time step.

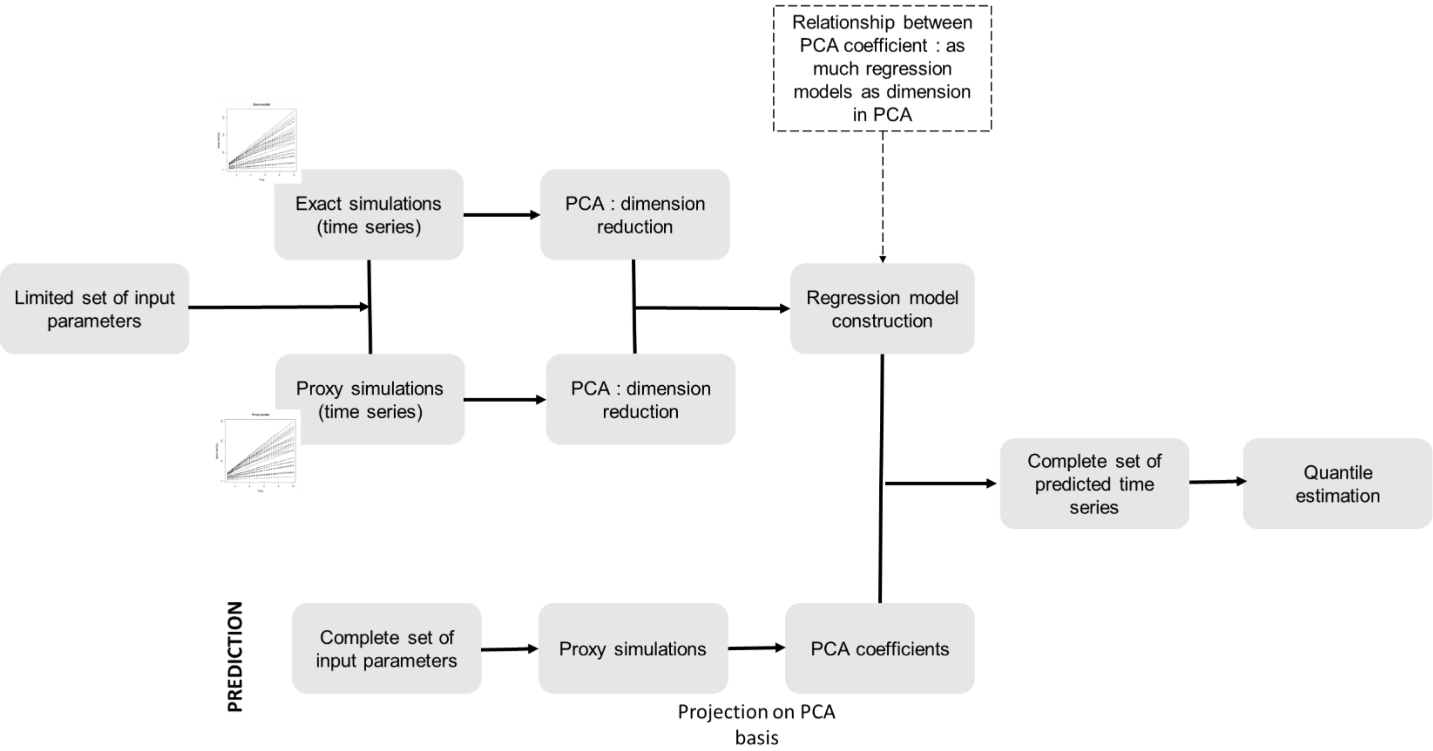

Relationship between PCA coefficient : as much regression models as dimension in PCA
Exact simulations (time series)
PCA : dimension reduction
Limited set of input parameters
Regression model construction
Proxy simulations (time series)
PCA : dimension reduction
Complete set of predicted time series
Quantile estimation
PREDICTION
Complete set of input parameters
Proxy simulations
PCA coefficients
Projection on PCA basis

Figure 46: Methology illustration.

6.4.1 *Choice and validation of the error model*

As previously, error model have to be built for the two capacity estimators, the total injected mass and the spread of the CO_2 plume. Figure 47 and Figure 48 allow the visualization of, respectively, the mass time series and surface time series, for the proxy and exact model on the limited learning set (random subset of 30 realizations).

The PCA is applied independently to both learning sets of proxy and exact curves. If all the components are considered, no approximation is made and the time series are represented exactly. The approximation depends on the number of dimensions kept in PCA. Figure 49 and Figure 50 present the RMSE calculated for, respectively, the mass and the surface variables, for different Npca. Logically the RMSE decrease when Npca gets higher. In terms of variance captured by each components, the first three compoeenents appear to capture more than 99.9% of the total variance of the mass variable (computed either with the proxy model responses or with the exact model ones), and more than 99.15% for the surface variable. PCA ability to reproduce time series is shown in Figure 51 and Figure 52, for a randomly chosen time series for Npca=1 and 3. Even though only one components seems to behave relatively correctly, mass and the surface, we chose to keep three components for maximizing the retained information at each time step.

The linear models between the three PCA coefficient couples are then built from a limited set of responses obtained with the exact and proxy models. These error models will allow, as in section 6.3, the prediction of surface and mass responses from a large set of responses obtained with the proxy model.

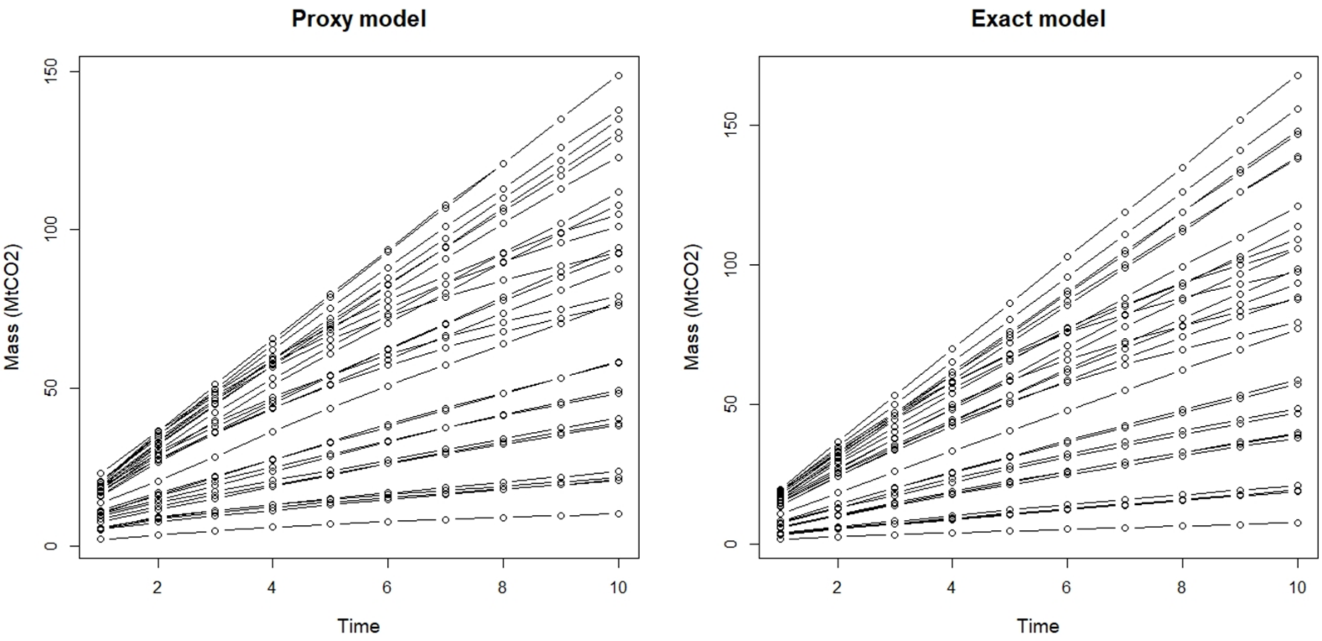


Figure 47: Time series of the total injected mass for the learning set of 30 simulations; on the left for the proxy model; on the right for the exact model.

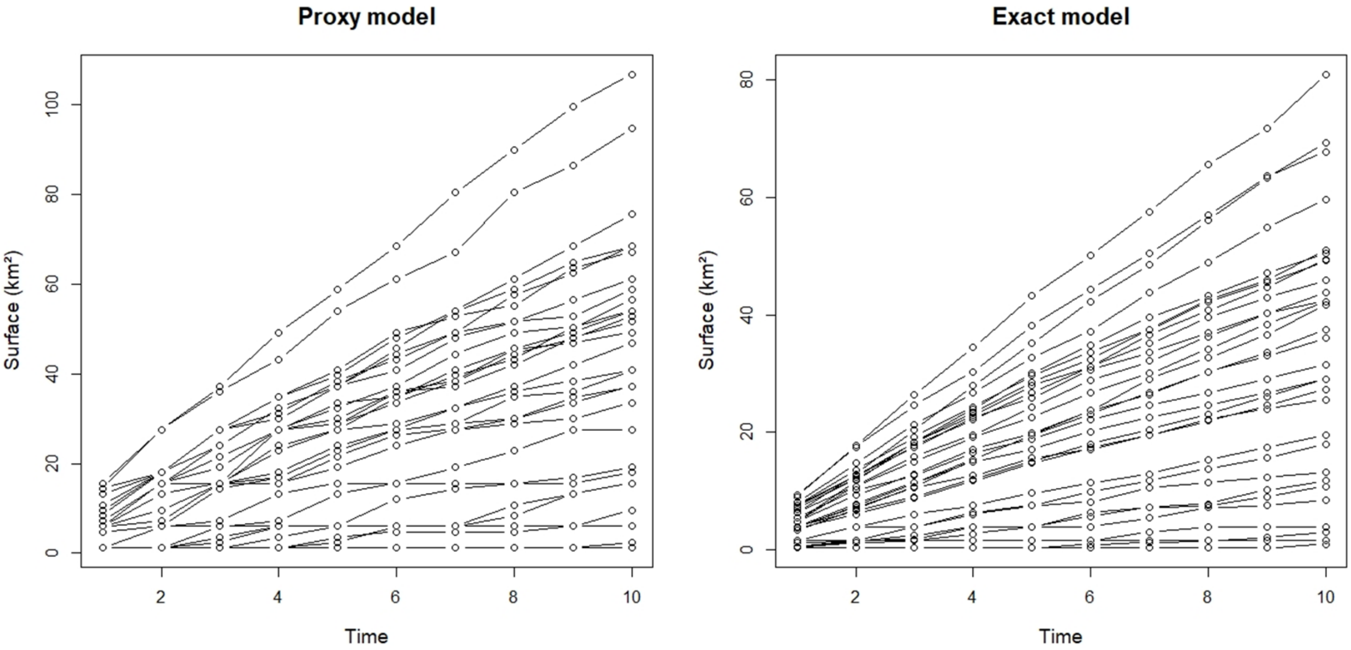


Figure 48: Time series of the CO_2 plume surface for the learning set of 30 simulations; on the left for the proxy model; on the right for the exact model.

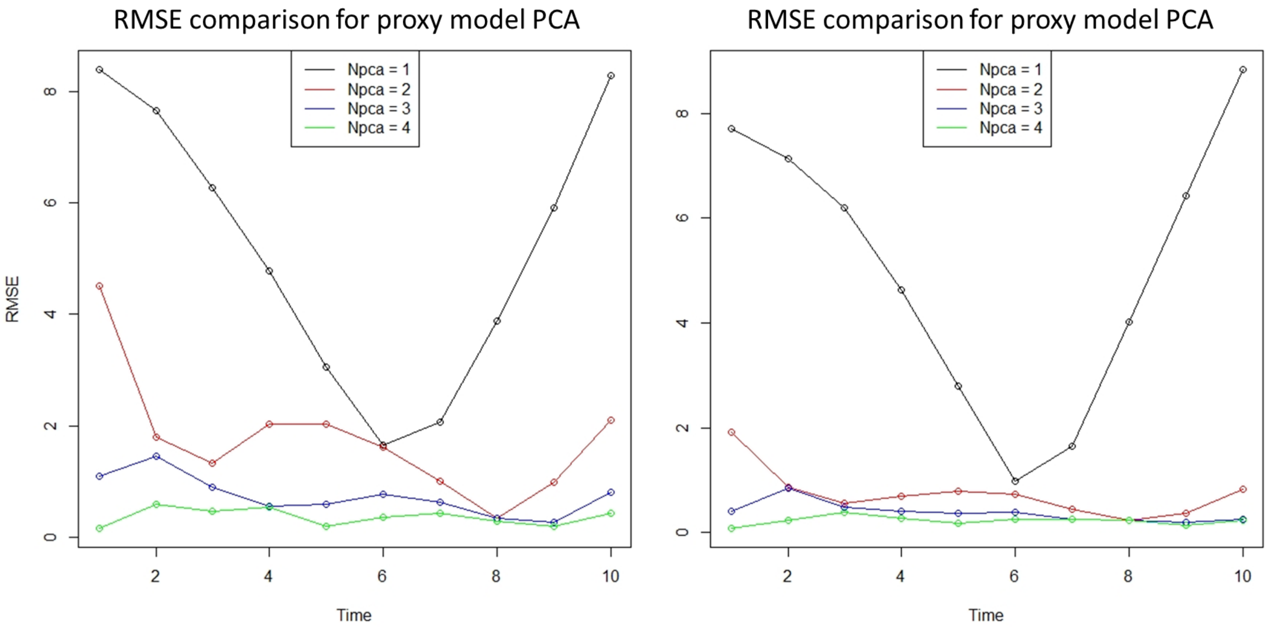


Figure 49: RMSE estimation for the mass time series, respectively, for the proxy model on the left and for the exact model on the right; at different dimensions kept in PCA (Npca).

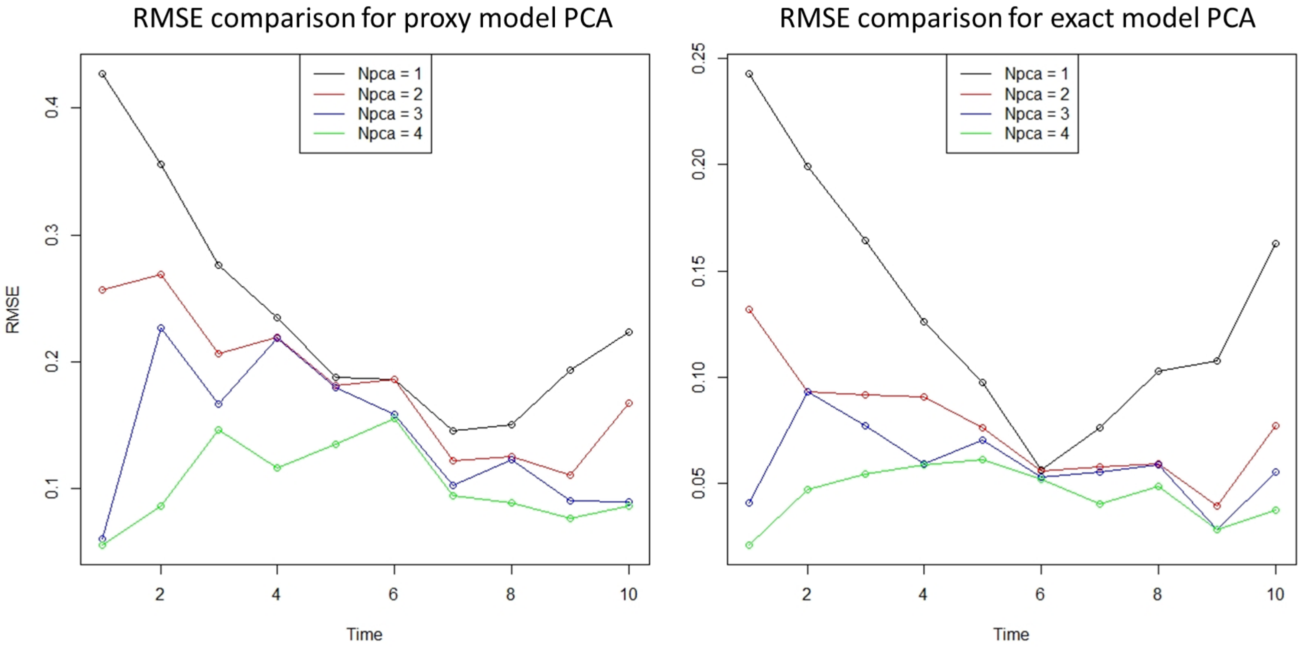


Figure 50: RMSE estimation for the surface time series, respectively, for the proxy model on the left and for the exact model on the right; at different dimensions kept in PCA (Npca).

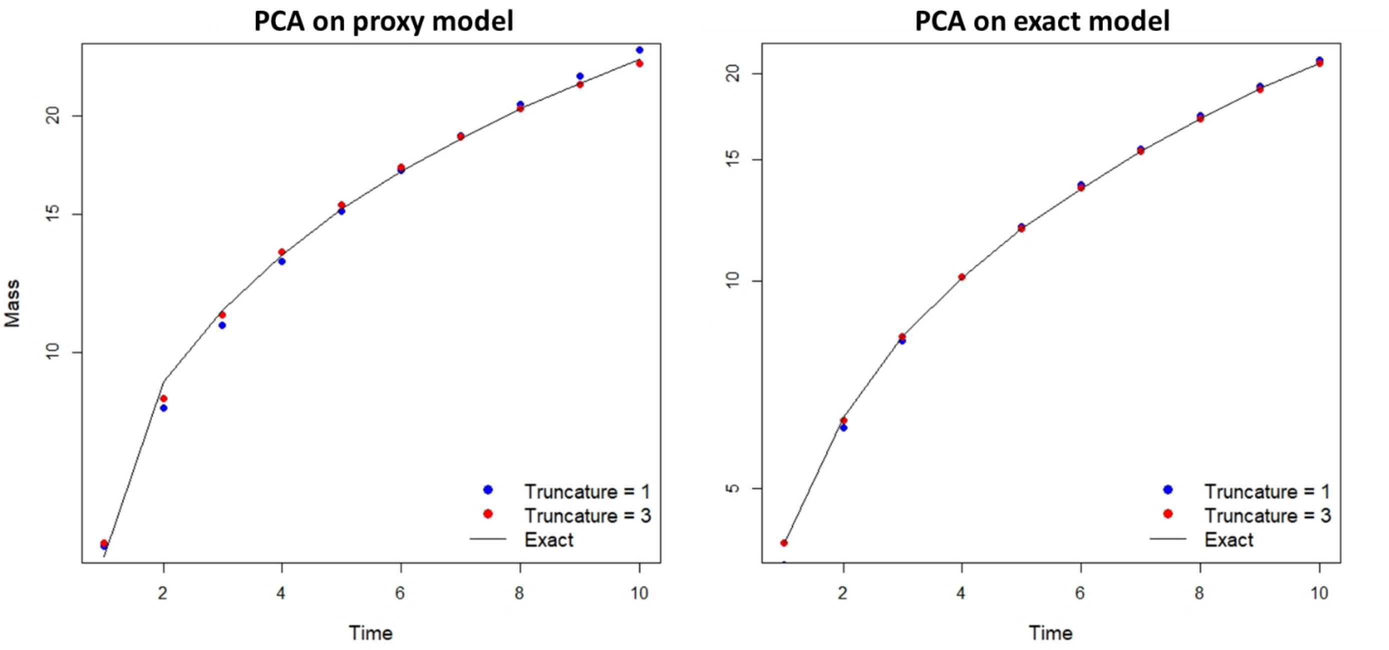


Figure 51: Mass time series reconstructed by PCA for Npca = 1 and 3; for the proxy model on the left and the exact model on the right.

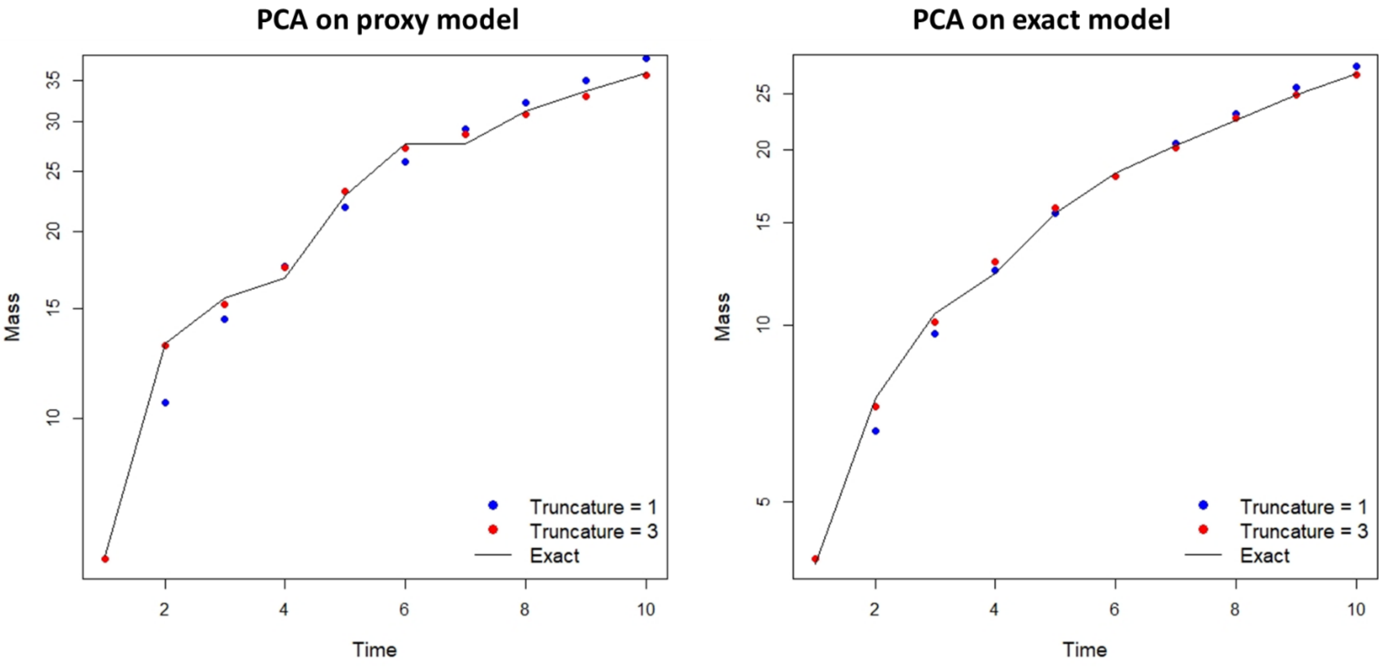


Figure 52: Surface time series reconstructed by PCA for Npca = 1 and 3; for the proxy model on the left and the exact model on the right.

6.4.2 Uncertainty propagation: quantiles estimation

As mentioned above, we are interested in the assessment of different quantiles for time-dependent capacity estimators. After having built error models on a limited learning set (N=30) of simulations run with the low fidelity and high fidelity models, predictions have been computed using the whole set of proxy responses (N=540) without requiring the simulations with the exact model for the entire set of input parameters.

The same quantiles than those presented in the barplots of section 6.3.3 (P05 for the mass variable and P95 for the surface variable) are shown as time series on Figure 53 and Figure 54. On these figures, the quantiles prediction computed with the error models are compared to 1) the quantile estimations computed with the proxy model responses and 2) the quantile estimations computed with the exact model responses (for methodological purposes). Similarly than for the scalar procedure, "predicted quantiles" are in good agreement with the "exact quantiles" even when deviations exist between the "proxy quantiles" and the "exact quantiles". The error models allow correcting the bias and aligning the "predicting quantile" time series along the "exact" ones.

The results of the application the UQ procedure on time series (including PCA) have been compared to the results of the procedure application on scalar presented in section 6.3: Figure 55 and Figure 56 provide the comparison of those two applications, regarding the estimation of the P05 for the mass variable and of the P95 for the surface variable. The results of the two applications are not exactly similar but both provide very satisfying estimates of the different quantiles over time.

To conclude, scalar and time series procedures are relatively close, this also raises the question of interest to sophisticate further the method. In our case study, 10 regression models for each variable had to be built when applying the scalar procedure against 3 models for the time series procedure. The gain in computational effort is, in our test case, not significant because of the few number of time steps (here 10), but may be more appreciable in case of time series with larger number of time steps. The main gain in computational efficiency appears to be in the application of the UQ procedure itself (that requires only a limited set of computationally intensive simulations, here a few tens) rather than in the choice to apply this procedure on each time steps or on the entire time series.

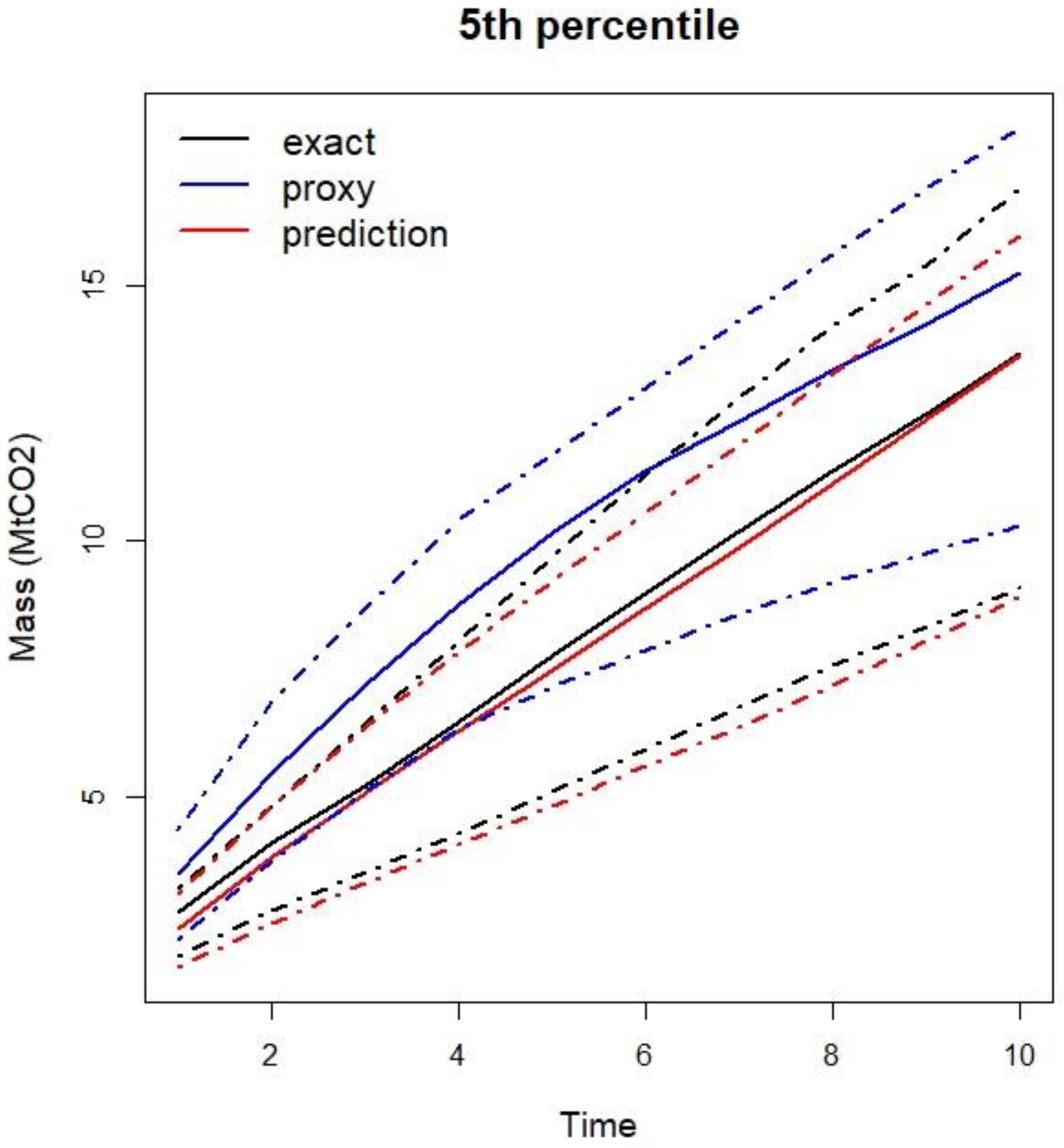


Figure 53: 5th mass percentile curves and their 95% confidence interval, obtained with the proxy and predicted models and compared to the reference quantile curves computed using the whole set of exact responses (solid black line).

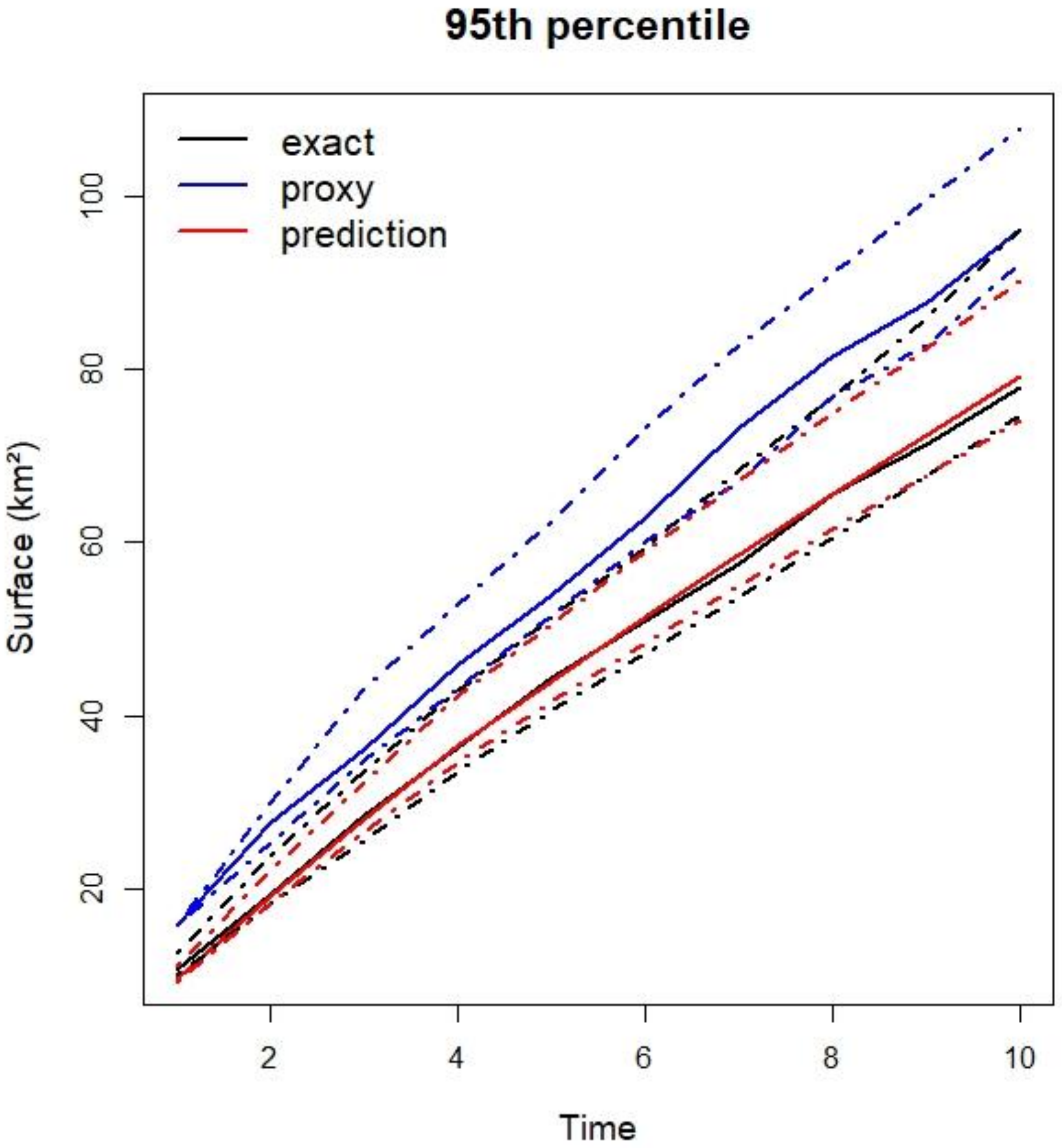


Figure 54: 95th surface percentile curves and their 95% confidence interval, obtained with the proxy and predicted models and compared to the reference quantile curves computed using the whole set of exact responses (solid black line).

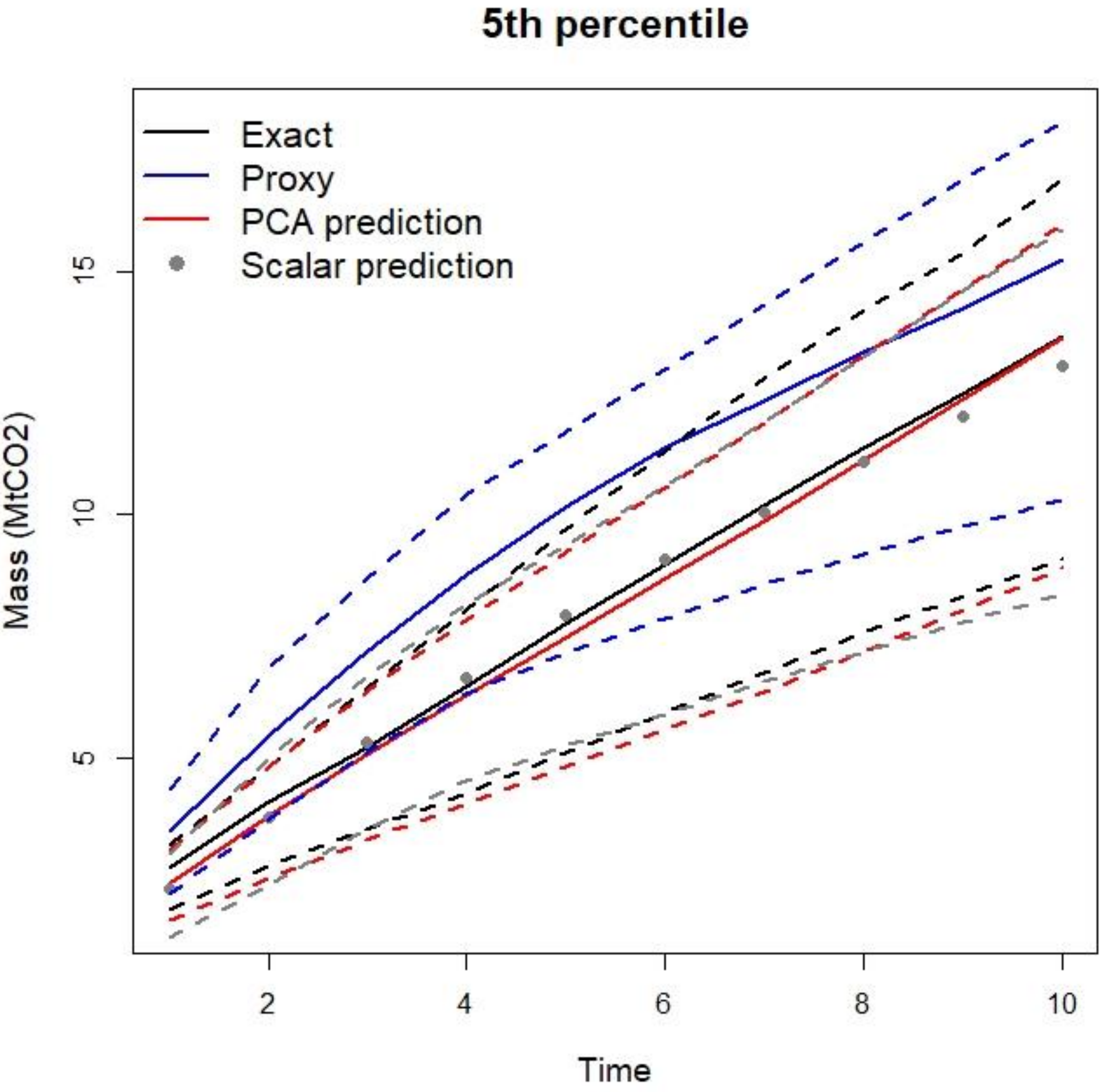


Figure 55: Model prediction comparison: 5th mass percentile curves and their 95% confidence interval, obtained with the proxy and two predicted models (with PCA and with scalar study) and compared to the reference quantile curves computed using the whole set of exact responses (solid black line).

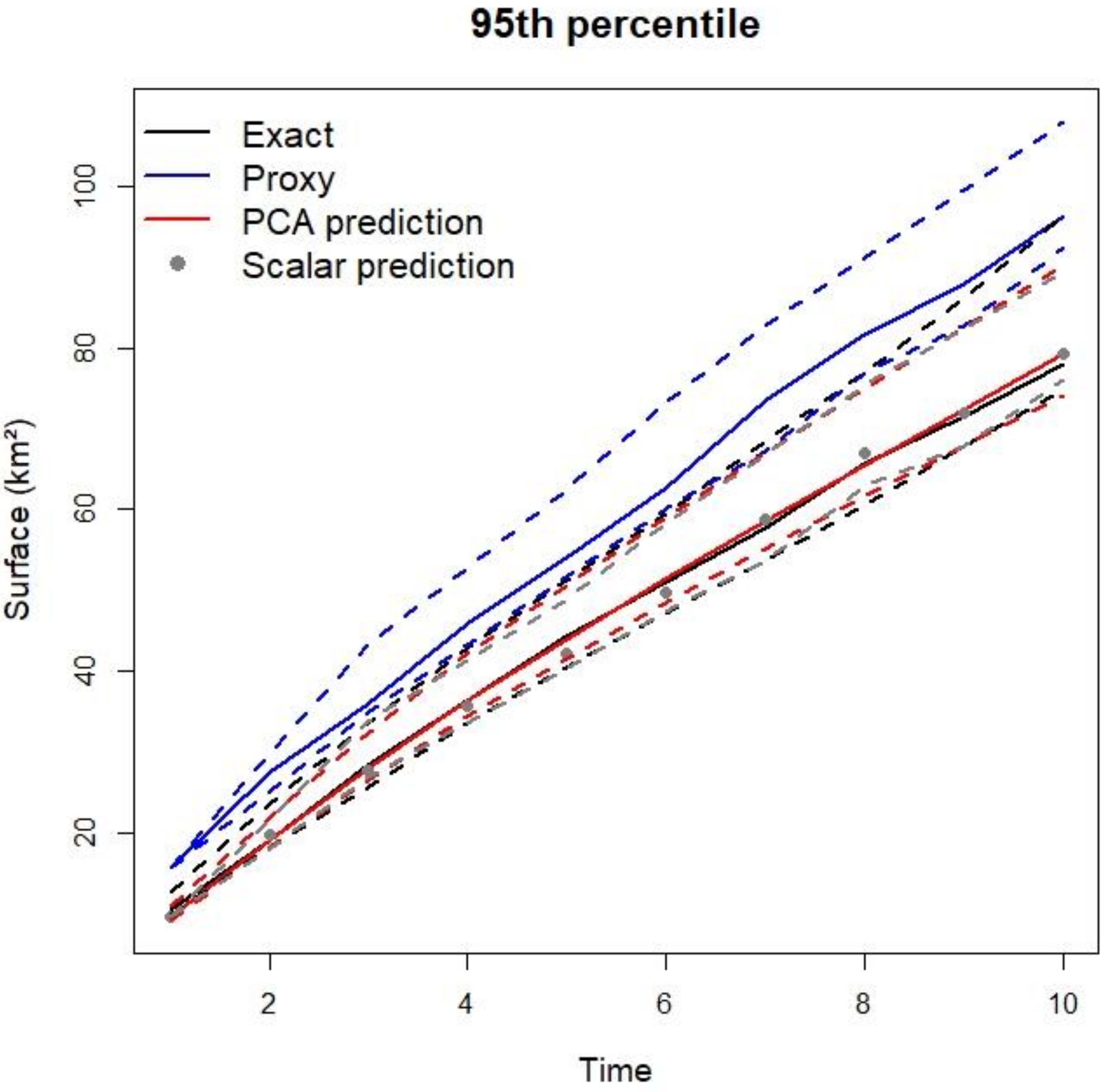


Figure 56: Model prediction comparison: 95th surface percentile curves and their 95% confidence interval, obtained with the proxy and two predicted models (with PCA and with scalar study) and compared to the reference quantile curves computed using the whole set of exact responses (solid black line).

7 General conclusion

Assessment of CO_2 storage capacity is associated to potentially large uncertainties, and reservoir simulations appear to be the most accurate way to perform such assessment (Bradshaw et al., 2007).

The objective of the present report was to quantify the reliability of storage capacities estimates performed with dynamic modelling. Such quantification is essential for an increased confidence of the different stakeholders (private and public) involved in a storage project, but is made difficult for several reasons:

- The procedure and choice of the indicator(s) for assessing dynamic capacities: even though a theoretical definition of capacity (storable quantity) can be given, neither guidelines nor (modelling) procedure seem to exist for unified and homogeneous practices regarding capacity estimation.
- The management of the uncertainties stemming from the dynamic modelling process: the variety of uncertainties stemming from the dynamic modelling construction workflow (uncertain input parameters due to limited number, poor representativeness, imprecision, simplification of physical processes, modelling assumptions, gridding, …) require uncertainty management techniques that handle different types of input variable (notably continuous and discrete).
- The computational time cost required to run refined dynamic models to ensure the most reliable capacity estimation: for a reliable capacity estimate, the reservoir simulations needs to offer the most accurate picture of the reality, which often implies the construction of a high complexity dynamic model. Such modelling is characterized by high computational costs, while a large number of simulations is needed for a robust uncertainty assessment.

Two facets of uncertainty management have been tackled in this study:

- importance ranking or sensitivity analysis (ENOS Task 2.1.2, step 1): The first objective was to develop a framework for getting a better insight in the role played by these different forms of uncertainties
- reliability quantification or uncertainty quantification (ENOS Task 2.1.2, step 2): The second objective is to develop methods for propagating the afore-mentioned uncertainties by estimating the quantiles P5, P50, P95 of the capacity estimates.

All the developments on importance ranking and uncertainty quantification in the context of long-running dynamic modelling are supported with the real case of onshore injection CO_2 in the lower Triassic sandstone formation based on a potential project in the Paris basin (France) as described by Manceau and Rohmer (2016).

According to the description of work of the ENOS project, the application of the proposed approaches should have been performed on a priori selected sites: Hontomin, Spain (fractured carbonate aquifer), and GeoEnergy Test Bed, UK (faulted Permo-Triassic sandstone aquifer).

Because of delay in the dynamic model provision, it has been decided in 2017 to apply the uncertainty analysis on a dynamic model developed in a previous project (FP7 ULTimateCO_2); this dynamic model was used for assessing the long term fate of injected CO_2 in deep saline aquifers (Manceau & Rohmer, 2016); the uncertainty analysis was therefore carried out to this end.

In 2018, no additional models was available, and therefore, it has been decided to also apply the UQ procedure on the FP7 ULTimateCO_2 dynamic model but to modify the initial objective of the model: the dynamic model was thus run with a capacity estimation aim, in order to better respond to the ENOS WP2 objective.

Importance ranking results:

Global sensitivity analysis (SA) is a powerful setting to provide valuable information by addressing the following questions: what sources of uncertainty contribute the most to the uncertainties in the flow simulation results? How to rank these sources of uncertainties? And how to set priorities for future investigations? Methods for SA have extensively been used for unravelling the role played by uncertain parameters. Most studies compute sensitivities based on one specific SA approach, although different SA methods may result in a different importance ranking. However, for analysing sensitivity to both parameter and model uncertainties, there is to the authors' best knowledge often no consensus on which methods are best applicable regarding the specificities of the situation. The objective of the work was to test the feasibility (and potentially extend the functionality) of the available methods/approaches for dealing with global sensitivity analysis with respect to parameter and model uncertainties. To counteract the computational cost issue associated with dynamic modelling, we further restrict the comparison analysis to methods, which do not require too many numerical simulations (of the order of a few hundred)

Four types of approaches are tested on a dataset of 1000 numerical reservoir flow simulations:

- DGSA: a distance-based generalized sensitivity analysis approach based on the regionalized sensitivity analysis method;
- PAWN: a density-based GSA (*aka* moment-indepedent);
- M-VBSA: a combination of variance-based GSA and metamodeling techniques adapted to situations using continuous and categorical variables;
- RF: a machine learning approach based on the random forest technique.

Six uncertain input parameters have been considered (porosity, permeability, permeability anisotropy, regional hydraulic gradient, relative permeability and capillary pressure), two characterized by parameter uncertainties and four characterized by model uncertainties.

The results of the SA considering in turn each method (within-method analysis) have been evaluated by investigating the impact of the number of simulations as well as the robustness to the parametrisation of each method. Three criteria have been set-up:

1) Convergence of the sensitivity indices: it is reached if the values of the indices remain stable;
2) Convergence of ranking: it is achieved if the ordering between the parameters remains stable;
3) Convergence of screening: it is reached if the partitioning between non- and -influential parameters remains stable.

The comparison between methods has been done in two ways:

1) from a quantitative point of view, with the comparison of normalized sensitivity measures: After normalization of the sensitivity measures with respect to the maximum value reached for each method, and considering 250 simulations, a clear similarity appeared among most methods regarding the large importance of the relative permeability parameter. However, the importance ranking of the other parameters was less straightforward, and the normalized sensitivity measures could not be quantitatively compared, which has complicated the comparison.
2) from a practical point of view, with the comparison of the implementation requirements, and of the capacities of each approach. This comparison has allowed the establishment of a pros/cons table (Table 5) to be used when a sensitivity analysis is performed.
3)

<u>Reliability quantification results</u>

In our study, we propose to evaluate the quantiles (namely P05, P50 and P95) of capacity indicators (total injected CO_2 mass and the surface of the free CO_2 in contact to the caprock). Two difficulties are dealt with:

1) the computational burden related to the use of high fidelity dynamic models;
2) both indicators are time series.

We propose the following approach that relies on a limited set of "exact" simulations and a larger set of low fidelity simulations (proxy model). For the application to the case study:

- We chose as proxy model, the dynamic model built from a coarse grid (in horizontal and vertical direction), compared to the exact model, which is a model built from a refined grid.

- The quantities of interest are twofold: total injected mass, and spread of the CO_2 plume. The application is therefore carried out on these two model responses, from which the 3 quantiles (P05, P95 and P50) are estimated.

An error model is then constructed, based on that limited set of exact simulations, to unbias the proxy model responses; the proxy model is corrected with the error model to predict the exact model responses for the entire set of simulations.

As the quantities of interest are time series, two different approaches are tested:

- The UQ procedure is applied to each time steps, i.e. a new error model (regression) is constructed for t=1, 2, …, 10 years (application on scalar quantities);
- The UQ procedure is applied after that a Principal Component Analysis is applied on the simulation responses time series, with the purpose of diminishing the dimension of the response spaces and therefore the size of the regression problem (application on time series).

The application to the injection case in the lower Triassic sandstone shows that both procedures yield similar results (in terms of predictability performance). We note however that the gain in computational effort would be more appreciable in case of time series with larger number of time steps than in the considered case (here 10 time steps). The main gain in computational efficiency appears to be in the application of the UQ procedure itself (that requires only a limited set of computationally intensive simulations, here a few tens) rather than in the choice to apply this procedure on each time steps or on the entire time series.

Nota bene: though the delays did not let us apply the procedures on an ENOS test case, it should be underlined that the developments described for importance ranking and uncertainty quantification are generic and can be applied to any dynamic models dedicated to capacity estimates. Besides, we believe that the considered test case (of onshore injection CO_2 in the lower Triassic sandstone formation based on a potential project in the Paris basin (France) as described by Manceau and Rohmer, 2016) is sufficiently realistic (in terms of quality, quantity and types of data in early stages of CO_2 storage test site) to consider the described recommendations potentially meaningful for future (or on-going) onshore CO_2 storage projects.

8 References

Altmann, A., Tolosi, L., Sander, O. and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure, Bioinformatics 26(10): 1340–1347.

Bachu S, Bonijoly D, Bradshaw J, Burruss R, Holloway S, Christensen NP, Mathiassen OM (2007) CO_2 storage capacity estimation: Methodology and gaps. International Journal of Greenhouse Gas Control 1 :430–443

Breiman, L., 2001. Random forest. Mach. Learn. 45 (1), 5e32.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. CRC Press, (1984)

Doughty, C.: User's guide for hysteretic capillary pressure and relative permeability functions in iTOUGH2. In: Report LBNL-2483E. Lawrence Berkeley National Laboratory, Berkeley, CA, USA, (2009)

Fenwick, D., C. Scheidt, J. Caers Quantifying asymmetric parameter interactions in sensitivity analysis: application to reservoir modelling. Math. Geosci., 46 (4) (2014), pp. 493–511

Fenwick, D., Scheidt, C., Caers, J., 2014. Quantifying asymmetric parameter inter-actions in sensitivity analysis: application to reservoir modeling. Math.Geosci. 46 (4),493–511.

Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer, New York (2009)

Homma T, Saltelli A. Importance measures in global sensitivity analysis of model output. Reliability Engineering and System Safety 1996;52:1–17

Jin, M., G. Pickup, E. Mackay, A. Todd, M. Sohrabi, A. Monaghan, et al. Static and dynamic estimates of CO_2-storage capacity in two saline formations in the UK SPE J, 17 (2012), pp. 1108-1118

Kopp, A., Class, H., Helmig, R.: Investigations on CO_2 storage capacity in saline aquifers: part 1. Dimensional analysis of flow processes and reservoir characteristics. Int. J. Greenh. Gas Control (2009a), 3, pp. 263-276

Kopp, A., Class, H., Helmig, R.: Investigations on CO_2 storage capacity in saline aquifers, part 2: estimation of storage capacity coefficients. Int. J. Greenh. Gas Control (2009b), 3, pp. 277-287

Land, C. S.: Calculation of imbibition relative permeability for two and three-phase flow from rock properties, Soc.Pet. Eng. J., 8(2), 149–156 (1968)

Larkin, R.G.: Hydrodynamic trapping of CO_2 geosequestrated in saline aquifers. SPE 128205. 2010 Improved Oil Recovery Symposium held in Tulsa, Oklahoma, USA (2010)

Larkin, R.G.: Hydrodynamic trapping of CO_2 geosequestrated in saline aquifers. SPE 128205. 2010 Improved Oil Recovery Symposium held in Tulsa, Oklahoma, USA, 24-28 April 2010 (2010)

Lenhard, R.J., Parker, J.C.: A model for hysteretic constitutive relations governing multiphase flow—2. Permeability–saturation relations. Water Resour. Res. 23(12):2197–2205 (1987)

Liaw A., and M. Wiener. Classification and regression by randomForest. R News, 2:18–22, 2002.

Lin, Y., Zhang, H.: Component selection and smoothing in smoothing spline analysis of variance models. Annals of Statistics 34, 2272–2297 (2006)

Manceau, J. C., & Rohmer, J. (2016). Post-injection trapping of mobile CO_2 in deep aquifers: Assessing the importance of model and parameter uncertainties. Computational Geosciences, 20(6), 1251-1267.

Pappenberger, F., Beven, K.J.: Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water resources research, 42(5), W05302 (2006).

Park, J., Yang, G., Satija, A., Scheidt, C., & Caers, J. (2016). DGSA: A Matlab toolbox for distance-based generalized sensitivity analysis of geoscientific computer experiments. Computers & Geosciences, 97, 15-29.

Park, J., Yang, G., Satija, A., Scheidt, C., Caers, J., 2016. DGSA: a Matlab toolbox for distance-based generalized sensitivity analysis of geoscientific computer experiments. Comput. Geosci.. http://dx.doi.org/10.1016/j.cageo.2016.08.021.

Peña, E. A., & Slate, E. H. (2006). Global validation of linear model assumptions.Journal of the American Statistical Association, 101,341e354.http://dx.doi.org/10.1198/016214505000000637

Pianosi, F., & Wagener, T. (2015). A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environmental Modelling & Software, 67, 1-11.

Pianosi, F., Beven, K., Freer, J., Hall, J. W., Rougier, J., Stephenson, D. B., & Wagener, T. (2016). Sensitivity analysis of environmental models: A systematic review with practical workflow. Environmental Modelling & Software, 79, 214-232.

Pianosi, F., Beven, K., Freer, J.W., Hall, J., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ. Modelling& Softw. 79, 214-232.

Pianosi, F., Wagener, T., 2015. A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. Environ. Model. Softw. 67, 1-11.

Pickup, G. E. (2013). CO_2 storage capacity calculation using static and dynamic modelling. In J. Gluyas, & S. Mathias (Eds.), Geological storage of carbon dioxide (CO_2): Geoscience, technologies, environmental aspects and legal frameworks (pp. 26-44). (Woodhead Publishing series in energy; Vol. 54). Oxford: Woodhead Publishing Ltd.. https://doi.org/10.1533/9780857097279.1.26

Pruess, K., Oldenburg, C.M., Moridis, G.J.: TOUGH2 User's Guide, Version 2.0. LBNL Report LBNL-43134 (1999)

Pruess, K., Spycher, N.: ECO_2N – A Fluid property module for the TOUGH2 code for studies of CO_2 storage in saline aquifers. Energy Conversion and Management. 48 (6), 1761-1767 (2007)

Rohmer, J. (2014). Combining meta-modeling and categorical indicators for global sensitivity analysis of long-running flow simulators with spatially dependent inputs. Computational Geosciences, 18(2), 171-183.

Saltelli, A., 2002b. Sensitivity analysis for importance assessment. Risk Anal. 22, 579-590.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al., 2008. Global Sensitivity Analysis. The Primer. John Wiley & Sons, Chichester.

Scheidt C, Caers JK (2009a) Representing spatial uncertainty using distances and kernels. Math Geosci 41(4):397–419

Sobol', I.M.: Sensitivity estimates for non linear mathematical models, Mathematical Modelling and Computational Experiments 1, 407–414 (1993)

Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., & Xu, C. (2015). Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. Journal of hydrology, 523, 739-757.

Spear RC, Hornberger GM (1980) Eutrophication in peel inlet—II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res 14(1):43–49

Spear, R.C., G.M. Hornberger Eutrophication in peel inlet – II. Identification of critical uncertainties via generalized sensitivity analysis Water Res., 14 (1980), pp. 43–49

Storlie, C. B., Swiler, L. P., Helton, J. C., & Sallaberry, C. J. (2009). Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. Reliability Engineering & System Safety, 94(11), 1735-1763.

Storlie, C.B., Bondell, H.D., Reich, B.J., Zhang, H.H.: Surface estimation, variable selection, and the nonparametric oracle property. Statistica Sinica 21, 679–705 (2010)

Storlie, C.B., Reich, B.J., Helton, J.C., Swiler, L.P., Sallaberry, C.J., Analysis ofcomputationally demanding models with continuous and categorical inputs. Reliability Engineering and System Safety 113, 30-41 (2013)

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Biases in random forest variable importance measures: illustration, sources and a solution. BMC Bioinfor. 8 (1), 25.

van Genuchten, M.T.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44, 892–898 (1980)

Wei, P., Lu, Z., Song, J. (2015). Variable importance analysis: A comprehensive review. Reliability Engineering & System Safety, Volume 142, October 2015, Pages 399-432

Young, P.C., Spear, R.C., Hornberger, G.M., 1978. Modeling badly defined systems: some further thoughts. In: Proceedings SIMSIG Conference, Canberra, pp. 24-32.

Zhou Q., J.T. Birkholzer, C.-F. Tsang, J. Rutqvist A method for quick assessment of CO_2 storage capacity in closed and semi-closed saline formations Int. J. Greenhouse Gas Control, 2 (2008), pp. 626-639

Zulqarnain, M., Zeidouni, M., Hughes, R. G. (2017). Static and Dynamic CO_2 Storage Capacity Estimates of a Potential CO_2 Geological Sequestration Site in Louisiana Chemical Corridor. Carbon Management Technology Conference, July 17-20, 2017, Houston, TX. doi:10.7122/486020-MS.

9 Appendices

9.1 M-VBSA: Metamodel-aided Variance-Based global Sensitivity Analysis

VBSA relies on the Sobol' indices (Sobol' 1993), which enables to perform the importance ranking depending on its objective (Saltelli et al. 2008): 1) the first-order Sobol' index (Si) of a given uncertain parameter, referred to as its main effect, corresponds to the expected proportion of the total variance of the model output (i.e. representing the uncertainty in the model output) that would be removed in average if one were able to fix the true value of the given uncertain parameter. This statistical measure is useful for prioritizing the input parameters in terms of effort to implement to decrease the output uncertainty range. 2) the total effect (St,i) corresponds to the expected proportion of the total variance of the model output if one were able to fix all the uncertain input parameters but the parameter i. This additional measure is useful to highlight the non-influential input parameters that can be fixed and therefore to simplify the model.

The implementation of VBSA in case of large scale flow modeling is hindered by two major limitations. First, VBSA is computationally intensive and requires a large number of model runs, of the order of 1,000 to 10,000 (see a review of existing algorithms in Saltelli et al. 2008), which is impractical using a large scale flow model with CPU time typically reaching a few hours. To overcome this computation burden, a possible solution relies on the use of meta-modeling techniques (aka response surface, or surrogate model, or statistical emulator), which basically consists in replacing the numerical model by a mathematical approximation referred to as meta-model. This corresponds to a function constructed using a few computer experiments (i.e. a limited number of time consuming simulations), and aims at reproducing the behaviour of the "true" model in the domain of model input parameters and at predicting the model responses with a negligible computation time. In this manner, any approach relying on intensive multiple simulations, such as VBSA, is made achievable at a reasonable computation time cost. The second problem is related to the nature of the model input parameters, which may either be continuous or categorical (scenarios like the type of relative permeability law). Accounting for categorical variables in meta-models has only been addressed recently (see the review provided by Storlie et al. 2013) especially in the case when the model is highly non-linear. In the present study, we focus on the ACOSSO-type metamodel, which has proven satisfactory performance results for CO_2 reservoir modelling (Manceau and Rohmer 2016; Rohmer 2014) for other underground uses like the Yucca Mountain repository for high-level radioactive waste (Storlie et al., 2013).

9.2 RF: Random forest

Random forest (denoted RF) is a non-parametric regression technique based on a combination (ensemble) of tree predictors (using regression tree as described by Breiman et al. 1984), which makes it "naturally" adapted to the processing of categorical input variables. Each tree in the ensemble (i.e. a forest) is built based on the principle of recursive partitioning, where the parameter space is recursively partitioned into a set of rectangular areas based on splitting rule (using for instance the decrease of the mean square error MSE). The partition is completed when the number of element within each node has reached a minimum value (a commonly adopted rule is 5 for regression, see e.g., Liaw and Wiener, 2002). Then, a constant value of the response variable is predicted within each area using the average value over the elements inside the partition. Compared to standard regression tree, randomness is introduced at two levels: i. each tree is constructed using a different bootstrap sample of the data; ii. each node is split using the best among a subset of predictors randomly chosen at that node.

Random forest model assigns a variable importance measure VIM to each input variable to reflect its influence on the quantity of interest. Different indicators of variable importance have been proposed in the literature (see a comprehensive review by Wei et al., 2015), namely Gini VIM and permutation-based VIM (denoted PVIM). Since previous studies have shown that Gini VIM tends to be biased and tends to overestimate the importance of the categorical variables with larger levels (Strobl et al., 2007). The principle of PVIM relies on the idea that the most important input parameter should have the largest decrease on the predictability of the RF model if its value was randomly perturbed.

Though the permutation-based approach can be useful to highlight influential variables using the value of PVIM as a ranking criterion, identifying negligible variables remains to be hindered by the lack of straightforward threshold to discriminate non- from important variables. Approaches based on hypothesis testing and repeated computation of RF model have arisen in the literature (Altmann et al. 2010), which provide a significance p-value to decide the informativeness (or not) of each input variable on a more objective ground. This proceeds by randomly permuting the response variable to break any associations between the response variable and all input variables. The data generated in this way is then used to construct a new RF model and to compute the PVIM for the input variables. This is re-conducted a given number *S* of times (say *S*=100 times). The *S* PVIM are used to compute the p-value for the variable as the fraction of S PVIM that are greater than the PVIM.

9.3 DGSA: Distance-based Generalised Sensitivity Analysis

The third category of method is originally based on Regional Sensitivity Analysis RSA and has recently been extended (method DGSA) to account for high-dimensional outputs and the influence of (asymmetric) interactions among the inputs. In the following, we present in turn both methods.

9.3.1 RSA: Regional (or generalized) Sensitivity Analysis

The RSA method (Young et al., 1978; Spear and Hornberger, 1980) consists in sorting the different sets of input parameters into two different groups according to the model's behaviour. The separation is made between one "behavioural" set, and one "non-behavioural" set. It can be based on the distinction between expected versus unexpected outcomes, or on a given output performance threshold. For each input parameter, the difference of the marginal cumulative probability distribution functions (denoted CDF) of the two sets can be used as a sensitivity index, i.e. as a measure of the impact of the input parameters on the model performance. The Kolmogorov-Smirnov (K-S) statistic is often used for the evaluation of that difference, and a two-sample K-S test may evaluate the statistical significance of the difference.

$$S_i = \max_{x_i} \left| \widehat{F_i}^{B} - \widehat{F_i}^{NB} \right|$$

Where x_i is the i^th^ input parameter, and $\widehat{F_i}^{B}$ and $\widehat{F_i}^{NB}$ are the empirical CDF of respectively of x_i considering respectively the behavioural and non-behavioural sets of output parameters.

The computation of the RSA sensitivity index, relying on input CDFs, is possible for different types of inputs, continuous and categorical. However, it cannot extract any information on the influence of interactions among input parameters on the model outputs. Moreover, the division into groups may be natural if the evaluation model is an objective function or if the threshold is not an arbitrary and subjective choice (e.g. imposed by regulators). In other situations, this choice may be questionable as it influences the sensitivity analysis results: RSA does not allow to extract the influence of one parameter within one group (Saltelli, 2002). A way to overcome this difficulty is to divide the outcomes in several groups equally separated, even though the method becomes, in that case, only visual (Pianosi et al., 2016).

9.3.2 DGSA: Distance-Based Generalized Sensitivity Analysis

Fenwick et al. (2014) extended the RSA concepts to account for 1) high-dimensional outputs (e.g. those or reservoir simulators), and 2) the influence of (asymmetric) interactions among the inputs. Relying as the RSA approach on inputs CDF, the DGSA method can be applied with different types of input parameters.

First Fenwick et al. (2014) generalized the RSA to multiple classes classified according to a chosen distance. An approach for distance-based classification is proposed by Scheidt and Caers (2009) consisting in choosing an appropriate distance function in reducing the dimension of the space of uncertainty and in performing the outputs classification using clustering techniques.

The number of classes depends on the objective of the analysis: small number of classes highlights parameters inducing large outputs changes, while smaller variations can be detected with a more significant number of classes. However, according to Fenwick et al. (2014) a class should contain not less than 10 simulations.

Similarly than for the RSA approach, a sensitivity index for a given parameter is defined based on the distance (area between the curves) between the empirical CDF (considering all the simulations) and the class-conditional empirical CDFs. A statistical test completes the analysis to judge whether a given parameter is significant or not regarding the afore-described distance.

9.4 PAWN: Density-based Global Sensitivity Analysis

The PAWN approach (presented in Pianosi and Wagener, 2015) relies on the emprirical CDF concept for assessing sensitivity. However, contrary to the RSA technique that quantifies the difference between input CDFs, the PAWN approach focuses on the variation of the output CDF when fixing an input. In that sense, the PAWN method is therefore closer to other density based GSA (see a recent review by Borgonovo and Plischke 2016). Since the method uses the CDF conditional on the value of the input parameter, this method is adapted to discrete indicator variables like the ones assigned to a set of modelling choices.

The following numerical procedure is proposed by Pianosi and Wagener (2015). The Kolmogorov-Smirnov (K-S) statistic is used to compute the difference between conditional and unconditional output CDFs. It is approximated as:

$$KS(x_i) = \max_y \left| \hat{F}(y) - \hat{F}(y|x_i = x_i^*) \right|$$

where y is the output; x_i is the considered input parameter and x_i^* is the fixed value assumed for x_i. A global sensitivity indicator S_i is proposed by summarizing the value of KS over a broad range of value for x_i, i.e. $S_i = \underset{x_i}{\mathrm{stat}}(KS(x_i))$, where stat is a statistic such as the maximum or the median. S_i is helpful for factor prioritisation but also for factor fixing (contrary than the RSA sensitivity index). The procedure proposed for factor fixing consists in performing KS two-sample statistical test at a given confidence level for each parameter and each conditioning value x_i. If the null hypothesis (no difference between $\hat{F}(y)$ and $\hat{F}(y|x_i)$) is never rejected for each conditioning value, then Pianosi et al. (2015) propose to consider that parameter as not influential.

This deliverable is prepared as a part of ENOS project

More information about the project could be found at http://www.enos-project.eu

Be nice to the world!

Please consider to use and distribute this document electronically.

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 653718