Challenges in conditioning a stochastic geological model of a heterogeneous glacial aquifer to a comprehensive soft dataset.

Julian Koch[1,2,*], Xin He[2], Karsten Høgh Jensen[1], Jens Christian Refsgaard[2]

[1]Department of Geosciences and Natural Resource Management, University of Copenhagen

[2]Department of Hydrology, Geological Survey of Denmark and Greenland

Submitted to:

HESS – Hydrology and Earth System Sciences

[*]Corresponding Author Address:

Julian Koch

Department of Hydrology

Geological Survey of Denmark and Greenland (GEUS)

Øster Voldgade 10

Copenhagen, Denmark

E-mail: juko@geus.dk

Telephone: +45 38142768

Abstract

In traditional hydrogeological investigations, one geological model is often used based on subjective interpretations and sparse data availability. This deterministic approach usually does not account for any uncertainties. Stochastic simulation methods address this problem and can capture the geological structure uncertainty. In this study the geostatistical software TProGS is utilized to simulate an ensemble of realizations for a binary (sand/clay) hydrofacies model in the Norsminde catchment, Denmark. TProGS can incorporate soft data, which represent the associated level of uncertainty. High density (20m x 20m x 2m) airborne geophysical data (SkyTEM) and categorized borehole data are utilized to define the model of spatial variability in horizontal and vertical direction, respectively and both are used for soft conditioning the TProGS simulations. The category probabilities for the SkyTEM dataset are derived from a histogram probability matching method, where resistivity is paired with the corresponding lithology from the categorized borehole data. This study integrates two distinct datasources into the stochastic modeling process that represent two extremes of the conditioning density spectrum; sparse borehole data and abundant SkyTEM data. Conditioning to vast soft data causes overconditioning; triggered by incorporating spatially correlated data in the modelling process. This is addressed by a work around utilizing a sampling/decimation of the dataset. In the case of abundant conditioning data it is shown that TProGS is capable of reproducing non-stationary trends. The stochastic realizations are validated by five performance criteria: (1) Sand proportion, (2) mean length, (3) geobody connectivity, (4) facies probability distribution and (5) facies probability – resistivity bias. As conclusion, a stochastically generated set of realizations soft conditioned to 200m moving sampling of geophysical data performs most satisfying when balancing the five performance criteria. The ensemble can be used in subsequent hydrogeological flow modeling to address the predictive uncertainty originated from the geological structure uncertainty.

Key words: Geological model, stochastic simulation, geophysical data, soft conditioning, model performance, TProGS

1. Introduction

Constraints in accurate and realistic solute transport modeling in hydrogeology are caused by the difficulty of characterizing the geological structure. The subsurface heterogeneity heavily influences the distribution of contaminants in the groundwater system. The scale of heterogeneity is often smaller than the data availability (e.g. borehole spacing). In traditional hydrogeological studies, one geological model is built based on the best comprehensive knowledge from often sparse borehole data and subjective interpretations. This can lead to alleged correct results, for instance when addressing the water balance on catchment scale, but will often prove to be inadequate for simulations beyond general flows and heads, e.g. contaminant transport modeling. Therefore, it is proposed by numerous studies that the uncertainty on the geological conceptualization is crucial when assessing uncertainties on flow paths (Neuman, 2003; Bredehoeft, 2005 ; Hojberg and Refsgaard, 2005; Troldborg et al., 2007; Seifert et al., 2008). One of the strategies often recommended for characterizing geological uncertainty and assessing its impact on hydrological predictive uncertainty is the use of multiple geological models (Renard, 2007; Refsgaard et al., 2012).

In this respect geostatistical tools such as two-point statistics e.g. TProGS (Carle and Fogg, 1996; Carle et al., 1998) and multipoint statistics (MPS) (Caers and Zhang, 2002; Caers, 2003; Journel, 2004; Strebelle, 2002) are powerful tools as they enable the generation of multiple equally plausible realizations of geological facies structure. This study targets the realistic description of heterogeneity in a geological model by introducing diverse data into the stochastic modeling process to generate a set of equally plausible realizations of the subsurface using geostatistics (Refsgaard et al., 2006).

In geostatistical applications field observations can constrain the simulation as soft or hard conditioning. “Hard conditioning” forces the realizations to honor data completely whereas “soft conditioning” honors the data partly with respect to the uncertainty of the observation (Falivene et al., 2007). This feature is essential because it enables the user to associate uncertainties to the conditioning dataset that can be of either subjective or objective nature. Incorporating a comprehensive and continuous soft conditioning

datasets to a stochastic simulation such as TProGS is challenging. Alabert (1987) published an early study on the implications of using sparse soft conditioning data to a stochastic simulation. The analysis shows that soft conditioning significantly increases the quality of the realizations. The same was also observed by McKenna and Poeter (1995) where soft data from geophysical measurements could significantly improve the geostatistical simulation. In the past years, highly sophisticated geophysical methods and advanced computational power allow stochastic simulations that are conditioned to a vast auxiliary dataset. This poses new challenges to the data handling and to the simulation techniques.

Chugunova and Hu (2008) present a study where continuous auxiliary data is introduced directly, without classification to a MPS simulation data in addition to the general training image. MPS requires a site specific training image that represents the geological structure accordingly, which is often the main source of uncertainty in MPS simulations. The above mentioned MPS studies conduct mostly 2D simulations, partly on synthetic data. The training image is the backbone of the MPS method and it has been acknowledged by dell'Arciprete et al. (2012) and He et al. (2013) that reliable 3D training images are difficult to acquire.

Alternative methods to integrate vast auxiliary information (e.g. geophysics) into the modeling process and at the same time force local accuracy are collocated cokriging or cosimulation techniques (Babak and Deutsch, 2009). Here a linear relationship between the auxiliary variable and the target variable is built in a model of cross covariance. The essentially linear relationship is often too restrictive and does not represent the complex physical processes. Mariethoz et al. (2009b) present a prospective method that extends the collocated simulation method by using a model of spatial variability of the target variable and a joint probability density distribution to depict the conditional distribution of the target variable and the auxiliary variable at any location.

The method of anchored distributions (MAD) (Rubin et al., 2010) is a suitable approach for the inverse modeling of spatial random fields with conditioning to local auxiliary information. Structural parameters such as global trends and geostatistical attributes are considered in a conditional simulation. The

conditioning is undertaken by anchored distributions which statistically represent the relationship between any data and the target variable.

The truncated plurigaussian simulation method (Mariethoz et al., 2009a) generates a Gaussian field for the target and the auxiliary variable using variogram statistics. These Gaussian fields are truncated to produce categorical variables that represent the hydrofacies. The truncation is controlled by threshold values that can be defined in a "lithotype rule" that represents the general geological concept. It is a very flexible method, because conceptual understandings are easily incorporated, but non-stationarity and especially directional depended lithotype rules are difficult to incorporate.

TProGS is a well-established stochastic modeling tool for 3D applications and it has been successfully applied to simulate highly heterogeneous subsurface systems by constraining the simulation to borehole data (Carle et al., 1998; Fleckenstein et al., 2006). Weissmann et al. (1999), Weissmann and Fogg (1999) and Ye and Khaleel (2008) use additional spatial information obtained from soil surveys, sequence stratigraphy and soil moisture, respectively for accessing the complex lateral sedimentary variability and thus improving the quality of the model in terms of spatial variability. It has not been tested whether TProGS, is capable of handling abundant soft conditioning data. Moreover, the risk that a cell-by-cell soft constraining may cause an overconditioning of the simulation has not been fully investigated. Overconditioing is defined by the authors as an effect triggered by dense and spatial correlated conditioning data that produces an altered picture of observable uncertainties. Therefore the self-consistency of the stochastic simulation is questioned, because soft constraining should be treated accordingly during the simulation.

Recent studies by Lee et al. (2007) and dell'Arciprete et al. (2012) highlight that TProGS is compatible with other geostatistical methods like, multi-point statistics, sequential Gaussian simulations and variogram statistics (Gringarten and Deutsch, 2001). The distinct strength of TProGS is the simple and direct incorporation of explicit facies manifestations like mean length, proportion and (asymmetric) juxtapositional tendencies of the facies.

Geophysical datasets are valuable information in many hydrogeological investigations. It can considerably improve the conceptual understanding of a facies or hydraulic conductivity distribution and identify non-stationary trends. However, the integration of geophysical data and lithological borehole descriptions is often difficult. A recent study by Emery and Parra (2013) presents an approach to combine borehole data and seismic measurements in a geostatistical simulation to generate multiple realizations of porosity. Hubbard and Rubin (2000) review three methods that allow hydrogeological parameter estimation based on geophysical data. The three methods link seismic, ground penetrating radar (GPR) and tomographic data with sparse borehole data to support the hydrogeological description of the study site. Our study integrates high resolution airborne geophysical data with borehole data to build a probabilistic classification of the subsurface at site. The geophysical data are collected by SkyTEM, an airborne transient electromagnetic method (TEM) that has been used extensively in Denmark for the purpose of groundwater mapping (Jorgensen et al., 2003b; Sorensen and Auken, 2004; Auken et al., 2009. This study utilizes a method that translates SkyTEM observation data into facies probability which enables associating the geophysical data with softness, according to the level of uncertainty. Very few studies have integrated high resolution airborne geophysical data in a stochastic modeling process (Gunnink and Siemon, 2009; He et al., 2013).

Most stochastic studies only make relatively simple validations of how well the simulations are able to reproduce known geostatistical properties. Carle (1997) and Carle et al. (1998) investigate the goodness of fit between the simulated and the defined spatial variability. The geobody connectivity is used by dell'Arciprete et al. (2012) to compare results originated from two- and multipoint geostatistics. Chugunova and Hu (2008) make a simple visual comparison between the auxiliary variable fracture density and stochastic realizations of the simulated fracture media. A more advanced validation is conducted in Mariethoz et al. (2009b) where simulated variograms and histograms are compared with reference data for the simulation of synthetic examples. In spite of these few studies that have addressed the validation issue, no guidance on which performance criteria to use and how to conduct a systematical validation of a stochastic simulation has been reported so far.

It should be noted that we in line with Refsgaard and Henriksen (2004) do not use the term model validation in a universal manner, but in a site specific context where a model validation is limited to the variables for which it has been tested as well as to the level of accuracy obtained during the validation tests.

The objectives of this study are: (1) to set up TProGS for a study site based on lithological borehole data and high resolution airborne geophysical data and investigate the effect of the two distinct conditioning datasets to the simulation; (2) to assess the problem of overconditioing in a stochastic simulation; (3) to ensure that non-stationary trends are simulated accordingly by TProGS; and (4) to identify and test a set of performance criteria for stochastic simulations that allow the validation against geostatistical properties derived from field data. The results of the present study are intended for application in a hydrological modeling context (Refsgaard et al., 2014).

2. Study Site

Figure 1 shows the 101 km^2 Norsminde catchment, located on the east coast of Jutland south of Aarhus. The topography allows a separation between an elevated western part, with changing terrain and a maximum elevation of 100 m and a flat and low elevated eastern part, where the coastline represents the eastern boundary. Glacial morphologies, namely moraine landscapes are predominant in most of the catchment. The geological stratigraphy indicated by borehole logs encompasses Paleogene and Neogene marine sediments underlying a heterogeneous stratigraphy of Pleistocene glacial deposits. The Paleogene sediments are characterized by very fine-grained impermeable marl and clay. Above the Neogene sequence shows sandy formations encased by a clay-dominated environment with Miocene marine sediments. The entire Miocene sequence varies in thickness up to 40 m and the sandy formations reach thicknesses of more than 10 m. The Miocene sequence is only present in the western part of the catchment where the stochastic modeling is conducted and forms the lower boundary of the simulation domain. Thus, only the upper Pleistocene glacial sequence is modeled. The glacial deposits in the western part of the catchment contain both sandy and clayey sediments, where clay is predominant. Borehole logs indicate that the Pleistocene clay spans from glaciolacustrine clay to clay till. Within the clay environment, the sandy units are allocated in small units and vary between gravel, meltwater-sand and sandy tills. The total thickness of glacial sediments varies between 10 and 40 m with heterogeneous distributions of the mostly glaciofluvial sand features between less than a meter and 20 m in thickness. The subject to the stochastic modeling, the delineated Pleistocene glacial sequence in the western part, provides interesting challenges like distinct heterogeneity and a diverse terrain.

3. Data

Two different sources of data, namely lithological borehole data and airborne based geophysical data (SkyTEM) are used, where the former is utilized to describe the vertical sand and clay variability and the latter for assessing the lateral direction.

3.1. Borehole Data

The borehole dataset contains 112 borehole logs with varying depths. The descriptions in the borehole reports are converted to a categorical binary (sand/clay) system at 5 cm vertical discretization. Further each borehole's uncertainty is validated He et al. (2014). The uncertainty assessment allows defining individual trust scores and thus the definition of how much each borehole should constrain the conditional simulation in the form of soft data. Drilling method, age, purpose of drilling, among others are used as variables to ensure a systematic approach to validate the uncertainty of each individual borehole. The boreholes are grouped into four quality groups with 100%, 95%, 90% and 85% as trust scores. The classified borehole dataset states an overall sand proportion of 30%.

3.2. Geophysical Data

The geophysical dataset comprises resistivity data from SkyTEM helicopter surveys. The SkyTEM method has been extensively used for subsurface mapping in Denmark (Jorgensen et al., 2003a; Jorgensen et al., 2005), where it has proven to be a successful tool for hydrogeological investigations. SkyTEM data have the advantage of a high spatial resolution in the top 20 to 30m and at large spatial coverage. However, some studies rise concern about the accuracy of interpretations of deep soundings (Andersen et al., 2013). In the Norsminde catchment data were collected at 2000 flight km containing over 100,000 sounding points. The distance between the flight lines is between 50 and 100 m. The dataset is processed with a spatially constrained inversion algorithm (Schamper and Auken, 2012) giving a 3D distribution of the underground resistivity. The sounding data were interpolated to a 20m x 20m x 2m grid domain by using 3D kriging as the interpolation method. The gridded resistivity data can be utilized as a proxy for lithology, as high and low resistivity cells indicate a high probability of sand and clay,

198 respectively. Bowling et al. (2007) conduct a detailed study on the relationship between sediment and
199 resistivity at a field site. Resistivity is linked to grainsize distribution and is used to delineate mayor
200 geological structures. A strong positive correlation between gravel content in the lithology and resistivity
201 is observed.

202 The SkyTEM dataset covers approximately 85% of the delineated glacial sequence. Figure 2 shows the
203 spatial variation of the median resistivity for a 4- and a 16- subarea grid. Higher median resistivity values
204 are located in the southern part of the glacial sequence. This indicates a greater sand proportion in the
205 given areas. The conclusion of the spatial pattern in Figure 2 is that stationarity cannot be attested to the
206 glacial sequence. This will have implications for the stochastic simulation.

207 The exact sand proportion can be derived by introducing a cut off value that divides the SkyTEM dataset
208 into a sand and a clay fraction. Jorgensen et al. (2003b) estimate resistivity thresholds to differentiate
209 between sediments in buried valleys in Denmark. Accordingly, glacial sand has a resistivity greater than
210 60 Ωm whereas clayey till sediments are placed between 25 and 50 Ωm and thus the exact cut off value
211 varies between study sites.

212 ### 3.3. Data integration

213 Figure 3 underlines some of the associated problems of the data integration of geophysical SkyTEM data
214 and borehole descriptions. The lithological information from the borehole interprets thin layers of
215 meltwater sand confined by clay in the top few meters. The SkyTEM data with a vertical resolution of
216 two meters cannot capture this small scale variability. This supports to use geophysical data only for the
217 lateral model of spatial variability and to incorporate the fine descriptions from the borehole data for the
218 vertical model of spatial variability.

219 He et al. (2014) developed a method to manually calibrate the cut off value by comparing borehole with
220 SkyTEM data at different spatial domains with the aim to reduce the deviation in sand proportion
221 between the two data types. It is assumed that the deviation has to be minimized at domains with a high
222 borehole density where the boreholes are assumed to best represent the domain conditions. It is shown

that a borehole density of 2 per km^2 reduces the representative error and that 46 Ωm as cut off value reduces the deviation in sand proportion between the two datasets. The calibrated cut off value yields a sand proportion of 23%.

Further He et al. (2014) developed a histogram probability matching (HPM) method that enables a direct translation from resistivity into facies probability. Resistivity is paired with the lithological borehole description at the coinciding cell. The data pairs are grouped in 10 Ωm bins and for each bin the sand/clay fraction is first calculated and then plotted as a histogram. 3^{rd} order polynomial curve fitting is applied to the histogram and the manually calibrated cut off value is superimposed to the fitted curve (Figure 4). The shape of the curve reflects the lumped uncertainties from both datasets. The flatness of the transition zone, around 50%, sand probability indicates a high uncertainty for the corresponding resistivity values. There are many sources of uncertainty that will affect the relationship between electrical conductivity and facies information. The HPM-method lumps various sources and the shape of the fitted curve reflects those, especially the width of the transition zone. He et al. (2014) discussed the prevalent uncertainties: first, borehole descriptions are not accurate, and classification of borehole lithology is subjective. Second, there are uncertainties on the resistivity data due to the resolution of the physics itself, the geophysics instruments, field measurements and signal processing (inversion). Third, there is no unique relationship between resistivity and lithology, and the curve can therefore be fitted in various ways. Last, there are uncertainties related to the scale of aggregation, since the borehole data and geophysical data have different resolutions and hence different supporting scales. The HPM-method used in this study is purely based on spatial correlations and is not build up on physical relationships. The main limitation is that it is site specific and cannot be applied to other catchments.

4. Methods

4.1. TProGS – Transition Probability Geostatistical Software

The geostatistical software TProGS is applied in this study. It is based on the transition probability (TP) approach (Carle and Fogg, 1996; Carle et al., 1998). Continuous Markov Chain models (MCM) are used to represent the model of spatial variability (Krumbein and Dacey, 1969; Carle and Fogg, 1997; Ritzi, 2000). TProGS allows for the simulation of multiple realizations by utilizing a sequential indicator simulation (SIS) (Seifert and Jensen, 1999) and by performing simulated quenching (Deutsch and Cockerham, 1994; Carle, 1997). These two steps are mutually dependent and they make sure that the realizations honor local conditioning data as well as the defined model of spatial variability.

The major advantage of TProGS is that fundamental observable attributes are parameterized in the modelling process: volumetric fractions (proportions), mean lengths (thickness and lateral extent) and (asymmetric) juxtapositional tendencies. These attributes can be assessed by data analysis and geological interpretations and control the shape of the MCM model. The facies proportion is related to the asymptotic limit of the MCM. The mean length is indicated on a plot of auto-transition probabilities as the intersection of the tangent at the origin with the x-axis. Asymmetric juxtapositional tendencies are of interest when simulating a system with at least three categories and can thus be neglected in this study. TProGS computes the realizations of the geology in two uncoupled, but mutually dependent steps. An initial configuration of facies distribution is produced by the SIS algorithm (Deutsch and Journel, 1992). Secondly, the initial configuration is reshuffled by the simulation quenching optimization algorithm (Deutsch and Cockerham, 1994). The TProGS simulation domain of this study is discretized into 20m x 20m x 2m cells on a 450 x 600 x 40 cell grid. The horizontal transition probabilities (TP) are based on SkyTEM data, that is categorized by a cut off value of 46 Ωm and the vertical extent is purely based on borehole data.

4.2. Split Sample Test

The two incorporated conditioning datasets are very distinct and will affect the simulation in opposite ways: sparse borehole data allow large simulation freedom whereas dense SkyTEM data limit the simulation freedom. Naturally they will be combined in order to condition the simulation to the best combined knowledge of the system. However it is of interest to know how each individual conditioning dataset affects the simulation. In this context a split sample test can reveal valuable information: one simulation conditioning to purely borehole data and the other one conditioned to purely SkyTEM data. It will be tested how well the simulations conditioned to borehole data reproduce the high resistivity cells, where a high sand probability is evident and how well the simulations conditioned to SkyTEM data reproduce the locations with borehole information.

4.3. Moving Sampling

Most studies on stochastic modeling condition the simulation to sparse data. In this study a comprehensive cell-by-cell soft conditioning dataset is applied and it is anticipated that this may result in overconditioning. Decimating the conditioning dataset out is a very intuitive sampling approach to work around the problem of overconditioning. However, if the resampled conditioning dataset is too sparse, information from the original dataset might not be sufficiently accounted for. Thus the tradeoff between two extremes, too much and too few data is investigated. Opposed to the static sampling technique a moving sampling method is applied. *n* different location grids with the same distance between the samples for each chosen distance (100m, 200m, etc.), where each has an accumulated shift of the origin (+ sampling distance/n in X and Y direction). For the 100m moving sampling approach the first sampling grid has the origin (0,0) the second (20,20), the third (40,40), etc. For the TProGS application in this study five location grids are generated, which yields five independent soft conditioning datasets. Five realizations are computed for each soft dataset; giving a total of 25 realizations. In addition to the comparison between moving and static sampling, different sample densities are also be tested.

4.4. Sampling scenarios

In total, eight conditioning scenarios are tested in this study. For the split sample test two scenarios are used, namely purely borehole data ('*onlyBH*') and purely cell-by-cell SkyTEM data ('*onlySky20*'). In the following both datasources are combined to represent the best combined knowledge of the system. Further, static and moving sampling are applied: Borehole data and SkyTEM data sampled statically at 20m, 100m, 200m and 500m ('*BH-Sky20static*', '*BH-Sky100static*', '*BH-Sky200static*' and '*BH-Sky500static*', respectively). Moving sampling is tested at 100m and 200m sampling distance ('*BH-Sky100moving*'and '*BH-Sky200moving*', respectively).

4.5. Performance criteria

Five performance criteria are defined to evaluate an ensemble of realizations of the geology. They aim for validating the ensemble with respect to the TProGS input, namely the defined model of spatial variability (mean length and proportion) and the soft conditioning dataset. The five performance criteria test the self-consistency of TProGS and thus if all input parameters and data are treated accordingly. The glacial structure in the Norsminde catchment represents only approximately 20% of the entire TProGS simulation domain and deviations in simulated spatial statistics between the entire model domain and the simulation target are expected.

4.5.1. Sand proportion

The deviation between the mean simulated sand proportion and the defined sand proportion in the MCM can be calculated for a set of realizations. The focus should be on the target area only, the area that will be extracted from the rectangular model domain for subsequent applications. The analysis of the sand proportion is based on 25 realizations.

4.5.2. Mean length

The simulated mean length can be estimated by recalculating the TPs from the TProGS output for the target area only. The simulated TPs for a set of realizations can be averaged (10 realizations in this case)

315 and compared with the measured TPs to estimate the deviation in mean length between the predefined
316 and the mean simulated length.

317 4.5.3. Geobody connectivity

318 The degree of connectivity of permeable areas in the subsurface has major implications for flowlines and
319 particle ages. Renard and Allard (2013) conducted a methodology study on various static and dynamic
320 connectivity metrics. These metrics can be utilized as a comparison and interpretation indicator for
321 multiple stochastically generated realizations of the geology. The work by dell'Arciprete et al. (2012)
322 shows the successfully implementation of connectivity metrics to compare stochastic realizations
323 computed by two- and multi-point statistics. Giudici et al. (2011) underline that evidence of a single
324 "best" connectivity metric is still missing and further research is necessary in that field.

325 For this study two static connectivity metrics, θ and Γ, are selected. They refer to the first and second
326 geobody connectivity defined by Hovadik and Larue (2007). A geobody is defined as one connected 3D
327 cluster of sand. Hence it is a distinct sand feature that is confined by clay. The architectural elements are
328 interpreted on a 20m x 20m x 2m scale.

329 $$\theta = \frac{V_l}{\sum_{i=0}^{n} V_i} \qquad \text{Eq.1}$$

330 $$\Gamma = \frac{\sum_{i=0}^{n} (V_i)^2}{\left(\sum_{i=0}^{n} V_i\right)^2} \qquad \text{Eq.2}$$

331 where V_i is the volume of an individual geobody, n is the number of unconnected geobodies and V_l is the
332 volume of the largest occurring geobody. θ represents the ratio of the volume of the largest geobody to
333 the total volume. Denoted as Γ is the proportion of the pairs of cells that are connected among the entire
334 pairs. The two selected connectivity metrics originate from the percolation theory, which describes the
335 transition from many disconnected clusters to one large coherent cluster. This is mainly depending on the
336 facies proportion. As the proportion gradually increases it reaches a point where one big cluster appears.

337 The percolation threshold is expected to be approximately 0.59 and 0.31 for a 2D and 3D grid,
338 respectively Hovadik and Larue (2007). Mean values of θ and Γ are computed based on 10 realizations.

339 #### 4.5.4. Facies probability distribution

340 The facies probability distribution reflects the inter variability among a set of realizations and can be
341 extracted from a probability map. Each cell in the probability map reflects the simulated category
342 probability within a set of realizations. The comparison between the distribution of the original soft
343 dataset, which constrains the simulation and the simulated facies probability distribution, allows
344 validating the performance of the simulation. Ideally the distribution of the original soft dataset is
345 reproduced by the simulation, which does not allow assumptions concerning the accuracy of the
346 allocation pattern of the simulated facies probability.

347 #### 4.5.5. Facies probability – resistivity bias

348 The validation of the facies probability – resistivity bias depicts if the simulated facies probability
349 corresponds to the fitted curve derived from the histogram probability matching method, and thereby test
350 whether the simulated facies probability is according to the resistivity pattern. The simulated facies
351 probability value is paired with the coinciding resistivity value of the gridded SkyTEM dataset. The pairs
352 are grouped in 5 Ωm bins and the median values of simulated facies probability can be plotted for each
353 bin. Further the RMSE can be calculated between the simulated facies probability and the fitted curve at
354 each bin in order to quantify the agreement.

355 ## 5. Results

356 ### 5.1. TProGS setup

357 The computed transition probabilities (TP) and the fitted Markov Chain model (MCM) for both
358 horizontal and vertical direction are given in Figure 5. A sand proportion of 23% and a mean length of a
359 sand lens of 5m and 500m for vertical and horizontal direction respectively yield MCMs that are in good

agreement with the measured TPs. Figure 3 indicates an increasing gradient in sand proportion from north to south. This non-stationary trend is also shown in Figure 5 where the additional sand-sand transition MCMs are plotted that fit measured TP data from the northern and southern subdomain; defined by 13%, 2m, 400m and 30%, 5m and 600m respectively. 25 realizations are generated based on the MCMs that are specified in Figure 5.

5.2. Split sample test

Two sets of 25 realizations are computed. The entire conditioning dataset is split into two parts, in order to analyze the effect of both extremes of the conditioning spectrum: Abundant data (*onlySky20*) and sparse data (*onlyBH*).

5.2.1. Visual comparison

Figure 6 presents two individual realizations (a) and (b) and the resulting probability maps (c) and (e) from both conditioning datasets at an elevation of 49m. Examining the individual realizations reveals that the spatial variability is much greater for the *onlyBH* scenario results. This is reasonable, because the amount of constraining data is also much less. This conclusion is supported by the probability maps. The probability map computed from the *onlySky20* conditioning scenario shows only little inter variability among the 25 realizations and resembles almost a binary sand and clay image. The *onlyBH* scenario simulates a probability map that shows high inter variability among the computed realizations, but the high probable sand areas do not coincide with the high resistivity areas in the SkyTEM data (d), because many large sand features are not captured by borehole data. On the other hand, some high probable sand features in the *onlyBH* scenario are not represented by the *onlySky20* scenario, because small sand features that are indicated by the borehole data are not detected by the SkyTEM survey.

5.2.2. Quantitative comparison

High resistivity areas are defined by a minimum resistivity value of 60 Ωm which is equivalent to 70% probability of sand occurrence based on the fitted histogram curve in Figure 4. The results of the split

sample test are given in Table 1. The *onlyBH* scenario allocates only 20.1% of the high resistivity cells accordingly. Also, only 74.3% of the cells, where the lithology in the borehole reports shows sand are simulated correspondingly. Some of the borehole data are treated as soft data, which enables the simulation to overwrite the lithological information, during the SIS and the simulated quenching. This will happen especially when sand lenses are very thin and vertically confined by clay. The *onlySky20* scenario simulates 44% of those cells accordingly and allocates almost all high resistivity cells as sand. However, almost 60% of the high resistivity cells are simulated with 100% sand probability. This is in poor agreement with field data, because the fitted histogram curve does not exceed sand probability values higher than 85% (Figure 4). The SkyTEM dataset indicates a large high resistivity cluster in the south-west at an elevation of 49 m (Figure 6), which is not detected at all by the borehole dataset, because there is only one borehole penetrating this area.

5.2.3. Local comparison

Figure 7 shows the vertical profile of one borehole (99.918) that penetrates the sand cluster and compares the simulation results from the *onlyBH* and *onlySky20* scenarios. The borehole has a trust score of 95%. While both datasets agree on the top layer being sandy and the occurrence of a thick clay layer below 75 m followed by a sand layer, they disagree on the location of the deeper sand layer. In the borehole data this sand deeper sand layer is detected at an elevation of 45m and below, whereas the SkyTEM dataset indicates sand occurrence approximately 8m higher; 53m and below. This discrepancy between 45m and 53 m has considerable implications for the simulation results at 49 m shown in Figure 6. However, one borehole alone will not be sufficient to substantially influence the simulation over large areas. Marginal amplification of the *onlyBH* scenario is noticeable at borehole 99.918. On the other hand, sand probabilities are clearly amplified in the *onlySky20* scenario; everything above 0.5 is amplified close to 1.0 and everything below 0.5 close to 0. The results from Table 1 and Figures 6 and 7 support the assumption of overconditioning caused by the comprehensive cell by cell soft conditioning.

5.3. Overconditioing

The observed problem of overconditioing is caused by spatially correlated data which are incorporated into the modeling process. A very intuitive approach to work around the problem of overconditioning is decimating the SkyTEM dataset by only sampling part of it. This will only be necessary in horizontal direction because the correlation length of the data is much less in the vertical direction. There is a tradeoff between the correctly simulated facies probability and the accuracy of the spatial allocation pattern. To illustrate this tradeoff three resampled conditioning scenarios are compiled: 100m, 200m and 500m sampling distance in X- and Y-direction and at the same time also including the boreholes for conditioning. For each of the three conditioning scenarios (*BH-Sky100static*, *BH-Sky200static* and *BH-Sky500static*, respectively) 25 realizations are computed and the probability maps for sand are presented in Figure 8. The simulated probability maps of the *BH-Sky100static* and *BH-Sky200static* conditioning scenarios are visually almost identical. Therefore only the latter is shown (d) and the image reflects already a higher variability than the results by the *BH-Sky20static* scenario (c). Reducing the conditioning data density increases the uncertainty of sand or clay. But at the same time the accuracy of correctly locating sand or clay units decreases, because the *BH-Sky500static* scenario (e) shows high probable sand areas which are not indicated by the original dataset (b). If for instance a high resistivity cell embedded in low resistivity cells is sampled for the conditioning, this cell may generate a sand lens in the out thinned conditioning scenario but would be limited by the neighboring cells in the *BH-Sky20static* scenario. The moving sampling method can improve the spatial coverage of the conditioning datasets and thus improve the quality of a set of realizations.

Again, the high resistivity cells are investigated to analyze if the bigger sand lenses are simulated correctly by the different conditioning datasets (Table 2). It is evident that the percentage of cells at the extreme ends of the simulated sand probability falls drastically after decimating the soft data out. The 100m distance scenarios still allocates more than 80% of the high resistivity correctly. On the other hand, the *BH-Sky500static* performs poorly, by only simulating 32.7% of the high resistivity cells correctly. It is also evident that the differences between static and moving sampling are small with regard to the correct allocation of the higher resistivity cells.

435 ### 5.4. Performance criteria

436 For further validation of the different sampling distances (20m, 100m, 200m and 500m) and sampling
437 schemes (static and moving) the five identified performance criteria will be applied to quantify the quality
438 of the simulations.

439 #### 5.4.1. Sand proportion

440 Table 3 shows the defined sand proportions of the delineated glacial structure. In order to investigate non-
441 stationarities the model domain is additionally subdivided into north and south. The SkyTEM dataset
442 indicates a higher sand fraction in the southern part compared to the north, 30% and 13% respectively.
443 The simulated sand proportions for the *BH-Sky20static* scenario show a good agreement with the defined
444 values. Larger deviations are evident for the *BH-Sky200moving* scenario. Both conditioning scenarios are
445 capable of reproducing the non-stationarity of the system, in regard to the sand proportion. The sand
446 proportions are somewhat overestimated for *BH-Sky200moving* scenario, and much less for the *BH-*
447 *Sky20static* scenario. Also the overestimation of simulated sand proportion in the northern subarea is
448 larger than in the southern subarea.

449 #### 5.4.2. Mean length

450 The comparison of the early (first lag = 100m) measured and simulated TPs for the sand-sand transitions
451 in X- and Y-direction allows to validate how well the lateral mean length is simulated by TProGS. Figure
452 9 comprises the measured TPs in horizontal direction, the fitted MCM and the computed mean TPs for
453 the *BH-Sky20static* scenario and *BH-Sky200moving* scenario, based on 10 realizations, for the total and
454 the sub-domains. The effect of overconditioning is very evident, as the computed mean TPs based on 20m
455 sampling conditioning data purely represent the original measured TP values. Since no simulation
456 freedom is present, the MCM cannot control the output. On the contrary, the *BH-Sky200moving* scenario
457 computes mean TPs that are more independent from the original data and rather follow the defined MCM.
458 The mean length of a sand lens can be derived by the steepness of the tangent where the lag approaches
459 zero. In general, the TP at lag 0 and 100 m are simulated too low; indicating that the simulated mean size

460 of a sand lens is too small. This is more prominent in results by the *BH-Sky200moving* scenario. It is
461 evident that the non-stationarity of the mean length of a sand lens is represented accordingly, although it
462 is undersimulated at all domains.

463 5.4.3. Geobody connectivity

464 For the categorized SkyTEM data θ and Γ are computed as 98.7% and 99.3%, respectively. This shows
465 values close to unity and should not be seen as a real reference, rather as a benchmark, because the
466 extreme low variability picture does not account for any uncertainties. The TProGS simulations based on
467 the two conditioning scenarios both undersimulate the connectivity metrics. The *BH-Sky20static* scenario
468 yields negative deviations of 2.1% and 1.1%, respectively and the *BH-Sky200moving* scenario 2.8% and
469 1.4%, respectively. The results indicate that θ and Γ show a similar behavior, where Γ appears to be
470 decreasingly greater as the proportion increases. Values close to unity and the very small deviations are in
471 good agreement with the general percolation theory, which sets the percolation threshold to
472 approximately 30% for 3D grids (Hovadik and Larue, 2007).

473 5.4.4. Facies probability distribution

474 Figure 10 shows the probability distribution for all discussed conditioning scenarios, with static (a) and
475 moving (b) sampling, with 25 realizations in each set. The original soft data distribution has its maximum
476 at approximately 20% and less than 5% are with either 0% or 100% sand probability. The *BH-Sky20static*
477 scenario simulates approximately 70% of the cells with zero change and thus has an extremely poor fit
478 with the soft dataset and the overconditioning is very prominent. It appears that overconditioning
479 amplifies the conditioning values to the extremes (e.g. 0.6 is simulated as 1.0 and 0.4 as 0.0, Figure 7).
480 The *BH-Sky500static* scenario reproduces the probabilities from the original soft dataset well, with only
481 approximately 10% zero change cells. However, the allocation pattern shows small resemblance with the
482 original dataset (Figure 8, (b)). *BH-Sky100static* scenario gives an intermediate solution, as the
483 probability is better reproduced than with the *BH-Sky20static* scenario, but still, more than 20% of the
484 cells are simulated as purely either sand or clay within the ensemble. Nevertheless, the *BH-Sky100static*

scenario is dense enough to capture the full variability of the system, as indicated by the original SkyTEM dataset. Additionally the results of the *BH-Sky200static* scenario are plotted in (a). The number of zero variability cells is decreased to approximately 20% and the maximum at 20% sand probability is close to the original soft dataset. Figure 10, (b) compares the static with the moving sampling approach for the 100m and 200m distance scenarios. The simulated facies probability distribution shows no differences for the static and moving 100m distance scenarios. However, at 200m sampling distance, the two sampling techniques are distinguishable, as the moving sampling yields fewer zero variability cells than the static sampling.

5.4.5. Facies probability – resistivity bias

The results are given in Figure 11 for the static sampling (a) and the moving sampling approach (b). The strong amplification of the resulting probabilities originating from the *BH-Sky20static* scenario is obvious in (a). The *BH-Sky500static* scenario performs poorly, especially in high resistivity areas, because those areas are not sufficiently covered by the 500m sampling distance. A better fit is represented by the *BH-Sky100static* scenario, because the amplification is much lower than for the *BH-Sky20static* scenario, especially for high resistivity areas. On the other hand, low resistivity areas are more amplified than high resistivity areas. The *BH-Sky200static* scenario gives a satisfying fit with the original fitted curve, especially in high resistivity areas, which indicates that the high probable sand cells are mostly allocated correctlty by the model. The simulated facies uncertainty for the low resisitivty cells is rather amplified by the *BH-Sky200static* scenario. Figure 11, (b) investigates the simulation differences caused by the static and moving sampling approach. The behaviour is similar to Figure 10, (b), because the differences for the 100m distance scenarios are marginal, while the *BH-Sky200moving* scenario generates a slightly lower facies probability – resisitivity bias than the *BH-Sky200static* scenario. The RMSEs between the fitted curve (Figure 4) and the simulations show that the *BH-Sky200moving* and *BH-Sky200static* sampling conditioing scenarios perform best, both with a RMSE of 0.06. Comparable are the *BH-Sky100moving* and *BH-Sky100static* sampling conditioing scenarios with a RMSE of 0.09 and 0.08, respectivley. The *BH-Sky20static* scenario performs poorest with a RMSE of 0.2.

6. Discussion

6.1. Choice of geostatistical method

The choice of the stochastic method for this study is application driven (Refsgaard et al., 2014). In the Norsminde catchment, it is evident from both borehole and geophysical data that the glacial sequence contains till clay and sand lenses distributed in extremely irregular patterns that are non-stationary. Without dense conditioning data the heterogeneous and non-stationary structures will not be simulated correctly. TProGS among other two-point statistics enables soft conditioning, where the soft information represents the associated level of uncertainty of an observation. The other distinct strength of TProGS is the easy incorporation of observable geological attributes when defining the Markov Chain models. In multi-point statistics (MPS) the definition of a reliable 3D training image is challenging, especially when simulating extremely irregular patterns (Honarkhah and Caers, 2012). Defining a MPS training image for the Norsminde catchment is peculiar, because it could only be based on interpreted SkyTEM data; with inflated length scales in the vertical direction. This makes the model of spatial variability in TProGS more reliable and objective, because it is based on measured transition probabilities and not on an interpreted training image. Further the transition probabilities are based on the data type we trust best: borehole data in the vertical- and SkyTEM data in the horizontal direction. In this study it is of spatial interest to correctly simulate the vertical transition probabilities in order to subsequently simulate the flow paths in the shallow groundwater system most accurately. This requires a detailed description of the spatial variability of the vertical direction, with indication of thin sand lenses, only provided by borehole data.

However, MPS is broadly applied in 2D and 3D applications: The *snesim* algorithm (Liu, 2006) combines object-based and pixel-based methods in the general MPS framework, to enforce spatial pattern reproduction and local conditioning, respectively. It was successfully applied by He et al. (2013) in a 3D application. Another promising approach is given by Chugunova and Hu (2008), where MPS is tested on non-stationary 2D structures, by continuous soft conditioning to a secondary variable. Here two training images from the geological structure and from the secondary variable are joint in the simulation.

Many promising geostatistical methods have advanced to incorporate auxiliary information to constrain the simulated target variable: Truncated plurigaussian simulation (Mariethoz et al., 2009a), collocated simulation with probability aggregation (Mariethoz et al., 2009b). Most of them are only tested on 2D applications partly with synthetic data. This present study uses TProGS as the geostatistical tool, because of its reliable model of spatial variability and further it is well established in 3D applications with sparse conditioning data. The application of vast soft conditioning data to a TProGS simulation gives valuable information on how such data can influence the stochastic simulation results.

6.2. TProGS setup

Direct transformation of geophysical data, such as SkyTEM, into a deterministic subsurface model is risky, because too much reliance on geophysical mapping can lead to seriously wrong hydrogeological models (Andersen et al., 2013). Uncertainties are expected in both, geophysical and lithological data and the shape of the fitted histogram curve reflects those. High uncertainty is associated with the transition zone; around 50% sand probability. Although the cut off value that divides the SkyTEM dataset into sand and clay is calibrated, there is a large quantity of high uncertain cells included which make the measured TPs directly dependent on the cut off value. Therefore the facies proportion and mean length are very sensitive to the selection of the cut-off value. As a result, the MCM in the lateral direction, as part of the TProGS setup, is highly dependent on the way the SkyTEM data is treated. Difficulties in the integration of the two data types are indicated in Figure 2. Small scale heterogeneities indicated by the borehole descriptions are not represented by the coarser SkyTEM dataset. This supports computing the horizontal and vertical TPs individually using SkyTEM and borehole data, respectively.

The SkyTEM dataset used in the present study is a 3D grid of 20m x 20m x 2m which was spatially interpolated from soundings with distances of about 17 m and 50-100 m along and between the flight lines, respectively. To reduce the overconditioning problem it might have been preferable to use the direct sounding data instead of the interpolated dataset. A similar effect is achieved by resampling, but here interpolated data with a higher uncertainty than the direct soundings are used.

Simulating a binary system is a crude simplification of the broad range of sediments in the glacial sequence. However, classifying the SkyTEM data into discrete facies or deriving the soft information on facies membership are peculiar in a multi facies environment. Additionally less abundant facies (e.g. gravel) will show extremely uncertain correlations in the histogram probability matching method. Last the less abundant facies might be represented on a 20m domain, but it will often not be visible on the 100m domain chosen for the subsequent hydrological flow simulations. Dell'Arciprete et al. (2012) present a study where geostatistics are implemented to simulate small scale heterogeneities in a multi facies environment.

6.3. Data footprint

Borehole and SkyTEM data are integrated by the histogram probability matching method (He et al., 2014) where differences in support scale are partly neglected. The support scales of the two data types are expected to vary. The lithological data from the boreholes are aggregated to 2m to be in better vertical agreement with the geophysical dataset. The agreement in the lateral direction is more questionable, because the footprint increases with depth for the geophysical data. The footprint is approximately 15-20m on the surface and in the range of 50m at 30m penetration depth. Further the footprint will depend on the material; with a larger energized volume for high conductance materials (high clay content). The two steps of processing the sounding data, namely inversion and kriging are both expected to inflate the footprint by smoothing values. However one can assume that the chosen grid size of 20m x 20m x 2m is suitable for near surface resistivity values, because the footprint of the geophysical data is constantly smaller than the correlation length, which is approximately 500m in vertical direction and 5m in lateral direction.

6.4. Split sample test

Both datasources have advantages and disadvantages: Borehole data have a higher data certainty and a finer spatial resolution in the vertical extent to better represent smaller sand features, but are essentially undersampled in the lateral extend. On the other hand, SkyTEM data have a good spatial coverage and

represent the bigger sand features well, but at the same time the data are associated with a higher data uncertainty. At this point, four major sources of uncertainty can be defined: (1) The inversion that transforms the SkyTEM measurement into resistivity, (2) the borehole data, (3) the relationship between lithology and resistivity and (4) the footprint mismatch between small scale borehole data and large scale SkyTEM data. So it is precarious to assume the SkyTEM data as true geology, but it can serve as a reference/benchmark when validating the simulation results. The *onlyBH* scenario does not capture all of the main sand features, which are revealed by the SkyTEM survey: Only 20% of the high resistivity cells, where the resistivity is greater than 70Ωm are simulated correctly. For the *onlySky20* scenario only 44% of the sand descriptions in the boreholes are simulated correctly, which underlines that the SkyTEM data does not measure the finer sand features correctly. The conducted split sample test does not allow to draw firm conclusions on simulation performance, it rather analyses the agreement between the two dataset propagated through the model.

6.5. Overconditioning

Correlated data, both temporally and spatially are a common problem in hydrogeological investigations. It has not been previously reported how TProGS is able to handle such a conditioning dataset. TProGS stochastically simulates the subsurface facies system by utilizing the two mutually dependent steps SIS and simulated quenching. It is not assured if the soft information is considered accordingly for the cokriging of the local probability estimate in the SIS step nor if it is accounted for in the objective function used for the simulated quenching in the latest TProGS version. However Deutsch and Wen (2000) successfully integrate exhaustive soft data in simulated quenching.

Work around methods have to be developed to overcome the problems associated with overconditioning. Decimating the soft conditioning dataset may seem as an overly simplistic and very crude approach, but the study aims at finding the balance between too few data and too many data. The risk to miss important features is high when conditioning to too few data. This study mainly deals with the latter case, where too many data lead to an underestimation of the simulation uncertainty. Including a moving sampling strategy

ensures that the spatial variation in the original dataset is best represented. A drawback of this approach is that valuable information might be lost, which again underlines the need for model validation, where the entire geophysical dataset is used for the evaluation. The decimation approach works as a very pragmatic solution for a study-specific problem and its generalization might be limited. Decimating the SkyTEM dataset and only considering data on a 200m spaced moving sampling grid gives the most satisfying results. A 200m sampling distance is expected to be sufficient to adequately capture all relevant geological features proxied by the entire dataset; this can be argued by the fraction between the observed mean length and the conditioning spacing. The mean length of a sand lens is found to be 500m and can proxy the correlation length. With a horizontal length scale of 500m and sampling at 200m we still condition the simulation with two to three soft data points in each horizontal direction for each mean sized sand feature.

Concluding it cannot be directly concluded that overconditioning is a general problem in stochastic simulations where a vast conditioning dataset is applied. However it can be presume that heavily spatially correlated data will affect also other stochastic simulation algorithms. TProGS was clearly not developed to run with such comprehensive conditioning. To our knowledge, the problem of overconditioning has not yet been reported nor discussed and with our study we would like to create awareness. In regard to the technique of geophysical prospecting it can be concluded that the problem of overconditioning is clearly not limited to airborne based TEM data.

6.6. Non-stationarity

Non-stationarity can be identified by subdividing the SkyTEM dataset (Figure 3 and 5). It is successfully tested if abundant conditioning data alone is capable of reproducing the observed non-stationary patterns. In a situation of sparse data, e.g. only borehole data for conditioning, these non-stationary trends cannot be reproduced correctly. Seifert and Jensen (1999) present an approach to model non-stationarity, which might be more suitable for sparse conditioning data. They suggested dividing the model domain into several stationary sub-domains, and each subdomain is then characterized using independent MCMs.

When subdiving the model domain, care must be taken, that no major features are cut, because it is then difficult to model them correctly. This approach was tested in the present study, but results revealed that this method is not easily applicable in situations of abundant conditioning data, because large coherent sand features are cut by the sub-division and their connectivity could not be simulated adequately.

6.7. Performance criteria

We identified and tested five performance criteria for validating the model.

(1) *Sand proportion*. Artificial conditioning data outside the target area honoring the defined proportion and MCM may help to make the simulation more homogeneous. In that context, exhaustive hard conditioning outside the simulation target can be tested.

(2) *Mean length.* The simulated and measured TPs are compared by Carle (1997) and Carle et al. (1998). Carle et al. (1998) simulate a four category system and the simulated quenching yields a perfect match between the modeled TPs and the defined MCM. On the other hand, Carle (1997) underlines that small deviations are to be expected and shows this by various examples where different SIS and simulated quenching parameters are tested.

(3) *Geobody connectivity.* The connectivity is partly dependent on the proportion. The sand connectivity for the simulation based on the *BH-Sky200moving* scenario is simulated lower and the sand proportion higher in comparison to the results from the *BH-Sky20static* scenario. This shows that the geobody connectivity is not fully depending on the proportion in this study. However it is a more feasible performance criterion for proportions far below the percolation threshold.

(4) *Facies probability distribution.* A good agreement between the simulated facies probability distribution and the original soft dataset doesn't ensure that the allocation pattern of the simulated probability is correct. This becomes evident when validating the results of the *BH-Sky500static* scenario.

(5) *Facies probability – resistivity bias.* The simulated facies probability should be in agreement with a corresponding resistivity observation to ensure that the spatial allocation pattern is simulated correctly. All bins are weighted the same, neglecting the inequality of data in each bin.

We used 25, 10 and 10 realizations to compute the first three performance criteria, respectively. Computing a moving average shows than the mean converges to +/-2% deviation to the final mean after ca. 15 realizations for the first criterion and after ca. 5 realizations for the second and third criteria, which justifies the selected number of realizations. The two latter criteria incorporate the computed probability map based on 25 realizations. Probability maps proved to be a useful tool to investigate the inter variability among realizations (Alabert, 1987; Carle, 2003; Mariethoz et al., 2009b). The results of the *onlyBH* scenario show the highest inter variability and a moving average tested at 10 random locations in the grid shows that after 20 realizations the mean converges to less than +/-20% from the final mean and to less than +/-10% after 23 realizations. These numbers are supposed to decrease as the conditioning data increase and therefore are 25 realizations in the analysis of the two latter criteria justifiable.

Table 4 compiles the five performance criteria for two different TProGS simulations: The *BH-Sky20static*- and the *BH-Sky200moving* scenario. The advantage of using multiple performance criteria is that concentrating on a single criterion may reveal an alleged good result, while another criterion attests a poor performance to the same simulation. Therefore a weighted and balanced analysis of the performance criteria helps to identify the best result. In this study, where abundant data are available, a good performance of the two latter criteria is as important as simulating accurate mean length/proportion. For example, if only considering sand proportion and mean length, it can be argued that the validation favors the *BH-Sky20static* scenario. However both, the facies probability distribution as well as the facies probability - resistivity bias attest poor performance. On the other hand, if interpreting the probability distribution only, it seems that the validation favors the *BH-Sky500static* scenario. Collectively, the conclusion is that the *BH-Sky200moving* scenario generates the overall most balanced results.

7. Acknowledgments

This work has been a part of the NiCA (Nitrate Reduction in a Geologically Heterogeneous Catchment) project (www.nitrat.dk), which is funded by the Danish Council for Strategic Research.

686 References

687 Alabert, F., Stochastic imaging of spatial distributions using hard and soft information, *M. S. thesis*,
688 1987.

689 Andersen, T. R., S. E. Poulsen, S. Christensen and F. Joergensen, A synthetic study of geophysics-based
690 modelling of groundwater flow in catchments with a buried valley, *Hydrogeology Journal*, *21*,
691 491-503, 2013.

692 Auken, E., A. V. Christiansen, J. H. Westergaard, C. Kirkegaard, N. Foged and A. Viezzoli, An
693 integrated processing scheme for high-resolution airborne electromagnetic surveys, the SkyTEM
694 system, *Exploration Geophysics*, *40*(2), 184-192, 2009.

695 Babak, O. and C. V. Deutsch, An intrinsic model of coregionalization that solves variance inflation in
696 collocated cokriging, *Computers & geosciences*, *35*(3), 603-614, 2009.

697 Bowling, J. C., D. L. Harry, A. B. Rodriguez and C. M. Zheng, Integrated geophysical and geological
698 investigation of a heterogeneous fluvial aquifer in Columbus Mississippi, *Journal of Applied*
699 *Geophysics*, *62*(1), 58-73, 2007.

700 Bredehoeft, J., The conceptualization model problem--surprise, *Hydrogeology Journal*, *13*(1), 37-46,
701 2005.

702 Caers, J., History matching under training-image-based geological model constraints, *Spe journal*, *8*(3),
703 218-226, 2003.

704 Caers, J. and T. Zhang, Multiple-point geostatistics: a quantitative vehicle for integrating geologic
705 analogs into multiple reservoir models, *Stanford University, Stanford Center for Reservoir*
706 *Forcasting. California, USA*, 2002.

707 Carle, S. F., T-PROGS:Transition Probability Geostatistical Software, University of California, Davis,
708 1996.

709 Carle, S. F., Implementation schemes for avoiding artifact discontinuities in simulated annealing,
710 *Mathematical Geology*, *29*(2), 231-244, 1997.

711 Carle, S. F., Integration of Soft Data into Categorical Geostatistical Simulation. Not published
712 manuscript, *Water Resources Research*, 2003.

713 Carle, S. F. and G. E. Fogg, Transition probability-based indicator geostatistics, *Mathematical Geology*,
714 *28*(4), 453-476, 1996.

715 Carle, S. F. and G. E. Fogg, Modeling spatial variability with one and multidimensional continuous-lag
716 Markov chains, *Mathematical Geology*, *29*(7), 891-918, 1997.

717 Carle, S. F., G. S. Weissmann and G. E. Fogg, Conditional simulation of hydrofacies architecture: A
718 transition probability approach, *SEPM Special Publication*, *1*(1), 147-170, 1998.

719 Chugunova, T. L. and L. Y. Hu, Multiple-point simulations constrained by continuous auxiliary data,
720 *Mathematical geosciences*, *40*(2), 133-146, 2008.

dell'Arciprete, D., R. Bersezio, F. Felletti, M. Giudici, A. Comunian and P. Renard, Comparison of three geostatistical methods for hydrofacies simulation: a test on alluvial sediments, *Hydrogeology Journal, 20*(2), 299-311, 2012.

Deutsch, C. V. and P. W. Cockerham, Practical Considerations in the Application of Simulated Annealing to Stochastic Simulation, *Mathematical Geology, 26*(1), 67-82, 1994.

Deutsch, C. V. and A. Journel, GSLIB: Geostatistical Software Libary and Users Guide, Oxford University Press, New York, 1992.

Deutsch, C. V. and X. H. Wen, Integrating large-scale soft data by simulated annealing and probability constraints, *Mathematical Geology, 32*(1), 49-67, 2000.

Emery, X. and J. Parra, Integration of crosswell seismic data for simulating porosity in a heterogeneous carbonate aquifer, *Journal of Applied Geophysics, 98*, 254-264, 2013.

Falivene, O., L. Cabrera, J. A. Munoz, P. Arbues, O. Fernandez and A. Saez, Statistical grid-based facies reconstruction and modelling for sedimentary bodies. Alluvial-palustrine and turbiditic examples, *Geologica Acta, 5*(3), 199-230, 2007.

Fleckenstein, J. H., R. G. Niswonger and G. E. Fogg, River-aquifer interactions, geologic heterogeneity, and low-flow management, *Ground Water, 44*(6), 837-852, 2006.

Giudici M., Baratelli F., Bersezio R., dell'Arciprete D., Cattaneo L., Cavalli E., Felletti F., Mele M., Pessina L, Vassena C. (2011). A multidisciplinary study of sediments' connectivity and transport parameters for aquifer analogues. ModelCare 2011. Models – Repositories of Knowledge. The 8th International Conference on Calibration and Reliability in Groundwater Modeling 18-22 September 2011, p.75

Gringarten, E. and C. V. Deutsch, Teacher's Aide Variogram Interpretationand Modeling, *Mathematical Geology, 33*(4), 507-534, 2001.

Gunnink, J. and B. Siemon, Combining airborne electromagnetics and drillings to construct a stochastic 3D lithological model, *15th European Meeting of Environmental and Engineerig Geophysics, Dublin, Ireland*, 2009.

He, X., T. O. Sonnenborg, F. Jorgensen, A. S. Hoyer, R. R. Moller and K. H. Jensen, Analyzing the effects of geological and parameter uncertainty on prediction of groundwater head and travel time, *Hydrology and earth system sciences, 17*(8), 3245-3260, 2013.

He, X., J. Koch, T. O. Sonnenborg, F. Jorgensen, C. Schamper and J. C. Refsgaard, Uncertainties in constructing stochastic geological models using transition probability geostatistics and transient AEM data., *Water Resources Research*, 2014.

Hojberg, A. L. and J. C. Refsgaard, Model uncertainty - parameter uncertainty versus conceptual models, *Water Science and Technology, 52*(6), 177-186, 2005.

Honarkhah, M. and J. Caers, Direct Pattern-Based Simulation of Non-stationary Geostatistical Models, *Mathematical geosciences, 44*(6), 651-672, 2012.

Hovadik, J. M. and D. K. Larue, Static characterizations of reservoirs: refining the concepts of connectivity and continuity, *Petroleum Geoscience, 13*(3), 195-211, 2007.

759 Hubbard, S. S. and Y. Rubin, Hydrogeological parameter estimation using geophysical data: a review of
760 selected techniques, *Journal of contaminant hydrology, 45*(1-2), 3-34, 2000.

761 Jorgensen, F., H. Lykke-Andersen, P. B. E. Sandersen, E. Auken and E. Normark, Geophysical
762 investigations of buried Quaternary valleys in Denmark: an integrated application of transient
763 electromagnetic soundings, reflection seismic surveys and exploratory drillings, *Journal of*
764 *Applied Geophysics, 53*(4), 215-228, 2003a.

765 Jorgensen, F., P. B. E. Sandersen and E. Auken, Imaging buried Quaternary valleys using the transient
766 electromagnetic method, *Journal of Applied Geophysics, 53*(4), 199-213, 2003b.

767 Jorgensen, F., P. B. E. Sandersen, E. Auken, H. Lykke-Andersen and K. Sorensen, Contributions to the
768 geological mapping of Mors, Denmark - A study based on a large-scale TEM survey, *Bulletin of*
769 *the Geological Society of Denmark, 52*, 53-75, 2005.

770 Journel, A., Beyond covariance: the advent of multiple-point geostatistics., *Geostatistics Banff. Springer,*
771 *1*, 225-223, 2004.

772 Krumbein, W. C. and M. F. Dacey, Markov Chains and Embedded Markov Chains in Geology,
773 *Mathematical Geology, 1*(1), 79-96, 1969.

774 Lee, S. Y., S. F. Carle and G. E. Fogg, Geologic heterogeneity and a comparison of two geostatistical
775 models: Sequential Gaussian and transition probability-based geostatistical simulation, *Advances*
776 *in Water Resources, 30*(9), 1914-1932, 2007.

777 Liu, Y. H., Using the Snesim program for multiple-point statistical simulation, *Computers &*
778 *geosciences, 32*(10), 1544-1563, 2006.

779 Mariethoz, G., P. Renard, F. Cornaton and O. Jaquet, Truncated Plurigaussian Simulations to
780 Characterize Aquifer Heterogeneity, *Ground Water, 47*(1), 13-24, 2009a.

781 Mariethoz, G., P. Renard and R. Froidevaux, Integrating collocated auxiliary parameters in geostatistical
782 simulations using joint probability distributions and probability aggregation, *Water Resources*
783 *Research, 45*, W08421, 2009b.

784 McKenna, S. A. and E. P. Poeter, Field example of data fusion in site characterization, *Water Resources*
785 *Research, 31*(12), 3229-3240, 1995.

786 Neuman, S. P., Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic*
787 *Environmental Research and Risk Assessment, 17*(5), 291-305, 2003.

788 Refsgaard, J. C., S. Christensen, T. O. Sonnenborg, D. Seifert, A. L. Hojberg and L. Troldborg, Review
789 of strategies for handling geological uncertainty in groundwater flow and transport modeling,
790 *Advances in Water Resources, 36*, 36-50, 2012.

791 Refsgaard, J. C., J. P. van der Sluijs, J. Brown and P. van der Keur, A framework for dealing with
792 uncertainty due to model structure error, *Advances in Water Resources, 29*(11), 1586-1597, 2006.

793 Refsgaard, J. C., E. Auken, C. A. Bamberg, B. S. Christensen, T. Clausen, E. Dalgaard, F. Effersoe¸, V.
794 Ernstsen, F. Gertz, A. L. Hansen, X. He, B. H. Jacobsen, K. H. Jensen, F. Joegensen, L. F.
795 Joergensen, J. Koch, B. Nilsson, C. Petersen, G. De Schepper, C. Schamper, K. I. Soerensen, R.
796 Therrien, C. Thirup and A. Viezzoli, Nitrate reduction in geologically heterogeneous catchments

- a framework for assessing the scale of predictive capability of hydrological models., *Science of the Total Environment*, -(- 0), 2014.

Renard, P., Stochastic hydrogeology: What professionals really need?, *Ground Water, 45*(5), 531-541, 2007.

Renard, P. and D. Allard, Connectivity metrics for subsurface flow and transport, *Advances in Water Resources, 51*, 168-196, 2013.

Ritzi, R. W., Behavior of indicator variograms and transition probabilities in relation to the variance in lengths of hydrofacies, *Water Resources Research, 36*(11), 3375-3381, 2000.

Rubin, Y., X. Y. Chen, H. Murakami and M. Hahn, A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields, *Water Resources Research, 46*, W10523, 2010.

Schamper, C. and E. Auken, SkyTEM Survey Norsminde and Lillebaek. NiCA pproject 2011, *HydroGeophysics Group. Aarhus University, 2011-06-16*, 2012.

Seifert, D. and J. L. Jensen, Using sequential indicator simulation as a tool in reservoir description: Issues and uncertainties, *Mathematical Geology, 31*(5), 527-550, 1999.

Seifert, D., T. O. Sonnenborg, P. Scharling and K. Hinsby, Use of alternative conceptual models to assess the impact of a buried valley on groundwater vulnerability, *Hydrogeology Journal, 16*(4), 659-674, 2008.

Sorensen, K. I. and E. Auken, SkyTEM - a new high-resolution helicopter transient electromagnetic system, *Exploration Geophysics, 35*(3), 194-202, 2004.

Strebelle, S., Conditional simulation of complex geological structures using multiple-point statistics, *Mathematical Geology, 34*(1), 1-21, 2002.

Troldborg, L., J. C. Refsgaard, K. H. Jensen and P. Engesgaard, The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system, *Hydrogeology Journal, 15*(5), 843-860, 2007.

Weissmann, G. S., S. F. Carle and G. E. Fogg, Three dimensional hydrofacies modeling based on soil surveys and transition probability geostatistics, *Water Resources Research, 35*(6), 1761-1770, 1999.

Weissmann, G. S. and G. E. Fogg, Multi-scale alluvial fan heterogeneity modeled with transition probability geostatistics in a sequence stratigraphic framework, *Journal of Hydrology, 226*(1-2), 48-65, 1999.

Ye, M. and R. Khaleel, A Markov chain model for characterizing medium heterogeneity and sediment layering structure, *Water Resources Research, 44*(9), 2008.

Tables

Table 1. Split sample test showing how many of the high probable sand cells (resistivity > 60 Ωm) are simulated with corresponding sand probabilities (> 70%) or fully deterministic (probability = 1.0) among 25 realizations. Conditioned to onlyBH and onlySky20. The last column shows how many of the areas that are shown as sand in the boreholes are simulated with sand probabilities > 85%.

Conditioning Scenario	**Prob. of sand > 0.7 AND resistivity > 60 Ωm**	**Prob. of sand = 1.0 AND resistivity > 60 Ωm**	**Prob. of sand > 0.85 AND borehole = sand**
onlyBH	20.1 %	1.34 %	74.3 %
onlySky20	99.0 %	59.1%	44.0 %

Table 2. Proportion of high probable sand cells (resistivity > 60 Ωm) that are simulated with corresponding sand probabilities (> 70%) or fully deterministic (probability = 1.0) for six conditioning datasets based on 25 realizations.

Conditioning Dataset	**Prob. of sand > 0.7 AND resistivity > 60 Ωm**	**Prob. of sand = 1.0 AND resistivity > 60 Ωm**
BH-Sky20static	97.9 %	63.8 %
BH-Sky100static / BH-Sky100moving	84.1 % / 87,3	10.4 % / 10.1 %
BH-Sky200static / BH-Sky200moving	75.8 % / 71.0 %	5.4% / 3.6 %
BH-Sky500static	32.7 %	1.5 %

Table 3. Simulated and defined sand proportions for the total domain and two sub-domains based on two simulations with different soft conditioning datasets (BH-Sky20static and BH-Sky200moving), based on 25 realizations.

Mean sand proportion (%) based on 25 realizations	**BH-Sky20static**		
	Total	**South**	**North**
Defined	23	30	13
Simulated	25.0	30.7	13.8
Deviation	+2.0	+0.7	+1.8
	BH-Sky200moving		
Defined	23	30	13
Simulated	29.3	33.7	21.5
Deviation	+6.3	+3.7	+8.5

Table 4. The five performance criteria and categorized SkyTEM data as benchmark that are applied to the two simulations with different soft conditioning datasets: Cell by cell soft conditioning and 200m moving sampling soft conditioning; both including borehole data. The first three criteria are expressed as deviation to the benchmark.

Performance Criteria	**Categorized SkyTEM**	**BH-Sky20static**	**BH-Sky200moving**
1. Sand proportion	23%	+2%	+6.3%
2. Mean length (X/Y)	500m	-21% / -20%	-37% / -37%
3. Geobody Connectivity (θ/Γ)	98.7% / 99.3%	-2.1% / -1.1%	-2.8% / -1.4%
4. Facies probability distribution	n.a.	Poor (approx. 70% cells with zero change)	Satisfying (approx. 15% cells with zero change)
5. Facies probability-resistivity bias	n.a.	0.20	0.06

Figure Captions

Figure 1. The study site in eastern Jutland, Dk. The Norsminde catchment with the delineated glacial structure in the western part of the catchment. Additionally the river network and the topography.

Figure 2. The median resistivity values from the SkyTEM data for the 4- and 16- subarea grid. Dark colors indicate a high median (max: 43.2 Ωm and 45.0 Ωm for the 4- and 16- subarea grid, respectively), light colors a low median (min: 32.0 Ωm and 29.5 Ωm for the 4- and 16- subarea grid, respectively) and white colors the absence of data. Additionally the location of the boreholes, the river network and the delineated glacial structure. The extent is 9km in X- and 12km in Y-direction.

Figure 3: Side-by-side comparison of borehole lithological data and SkyTEM vertical sounding data at borehole number 99.625 (He. et. al, 2014).

Figure 4. The bias corrected histogram curve: The calibrated cut off value (46 Ωm) is added to the histogram and the fitted curve is forced to honor it He et al. (2013).

Figure 5. The computed transition probabilities in vertical and horizontal direction and the fitted MCM: Vertical 5m, horizontal 500m mean length of a sand lens and 23% sand proportion. Additionally the fitted MCM for the north- and south-sub-domain are plotted for the vertical and horizontal sand-sand transitions: 2m, 400m, 13% and 5m, 600m, 30%, respectively.

Figure 6. Upper panel: Two individual realizations for two different conditioning scenarios: onlyBH (a) and onlySky20 data (b). Lower panel: Probability maps for the two scenarios c) and e) showing the probability of sand in each cell based on 25 realizations. The derived sand probability which is used for conditioning the simulation is shown in (d). All maps show data at an elevation of 49m.

Figure 7. The simulated versus the conditioned sand probability over the vertical extent at one borehole (98.918), located in the south western part of the glacial structure. The results originate from the two different soft conditioning scenarios: onlyBH and onlySky20 (based on 25 realizations each).

Figure 8. a): 100 m (small dots) and 500 m (big dots) sampling grids for thinning out the conditioning dataset; b-e): probability of sand at an elevation of 49 m for SkyTEM dataset (b), and for static 20m, 200m and 500m conditioning (c-e) Red colors represent high sand probability and blue colors low sand probability (based on 25 realizations).

Figure 9. The simulated transition probabilities for the south-, north-, and total-domain are compared with the SkyTEM data and the fitted MCM. The results for two soft conditioning dataset are shown: BH-Sky20static and BH-Sky200moving. The simulated TP and the MCM at lag 100m are compared to quantify the underestimation of a sand lens. The TP values are mean values based on 10 realizations. The defined length of a sand lens (X) and the mean simulated length for the BH-Sky20static (Y) and BH-Sky200moving scenario (Z) are given in each graph. (Xm – Ym / Zm).

Figure 10. The simulated facies probability distributions based on sets of realizations conditioned to differently sampled soft datasets (based on 25 realizations): (a) static sampling at different sampling distances and (b) stationary and moving sampling at different sampling distances. Also showing the sand probability distribution of the original soft dataset which is desired to be reproduced.

Figure 11. The simulated facies probability – resistivity bias based on sets of realizations conditioned to differently sampled soft datasets (based on 25 realizations): (a) static sampling at different sampling distances and (b) stationary and moving sampling at different sampling distances. The simulated sand probability is paired with the original resistivity value, grouped into 5 Ωm bins and then plotted as median for each bin. Also showing the observed data and the fitted curve from the histogram which is desired to be reproduced.

Figures

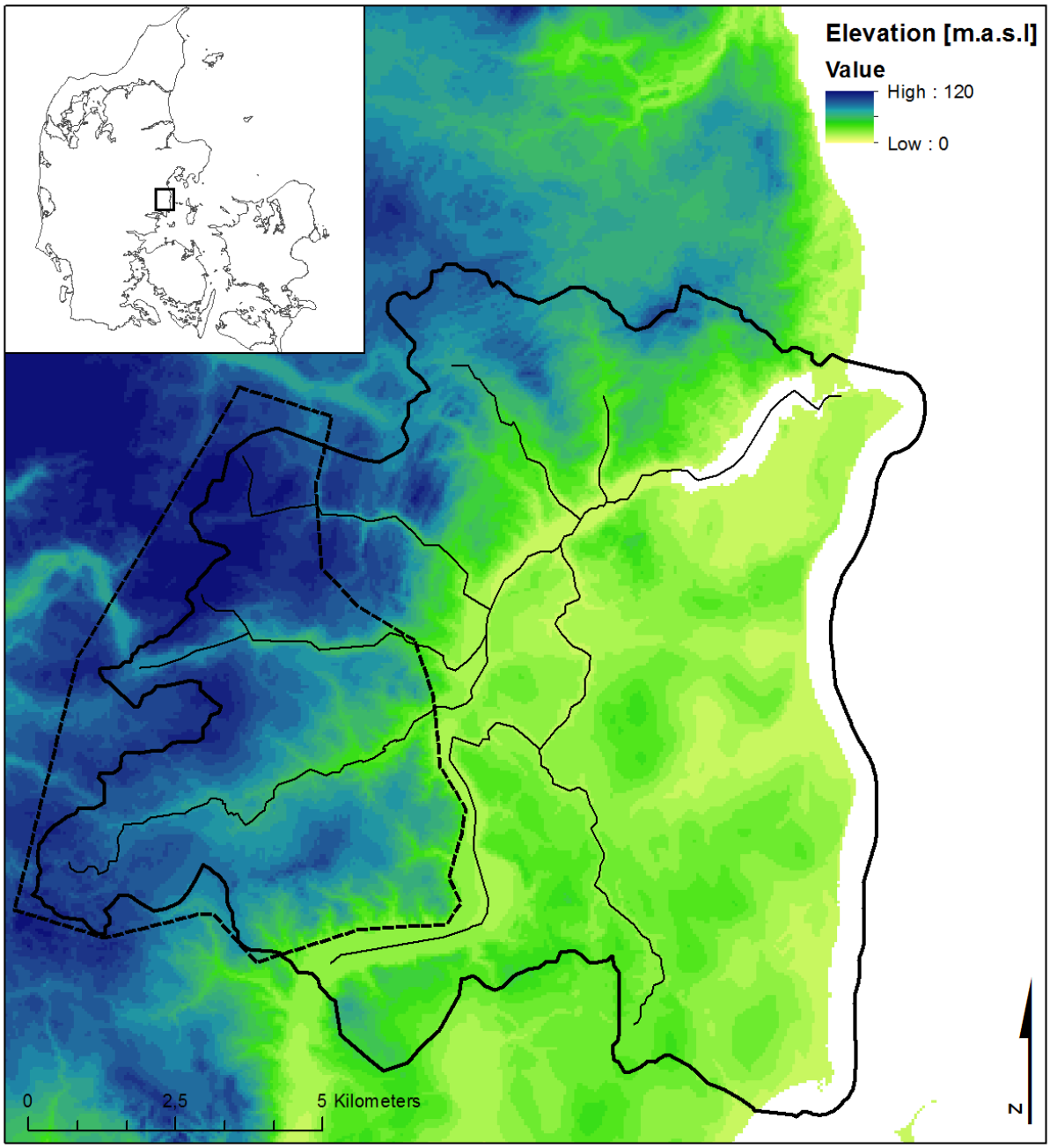

Figure 1. The study site in eastern Jutland, Dk. The Norsminde catchment with the delineated glacial structure in the western part of the catchment. Additionally the river network and the topography.

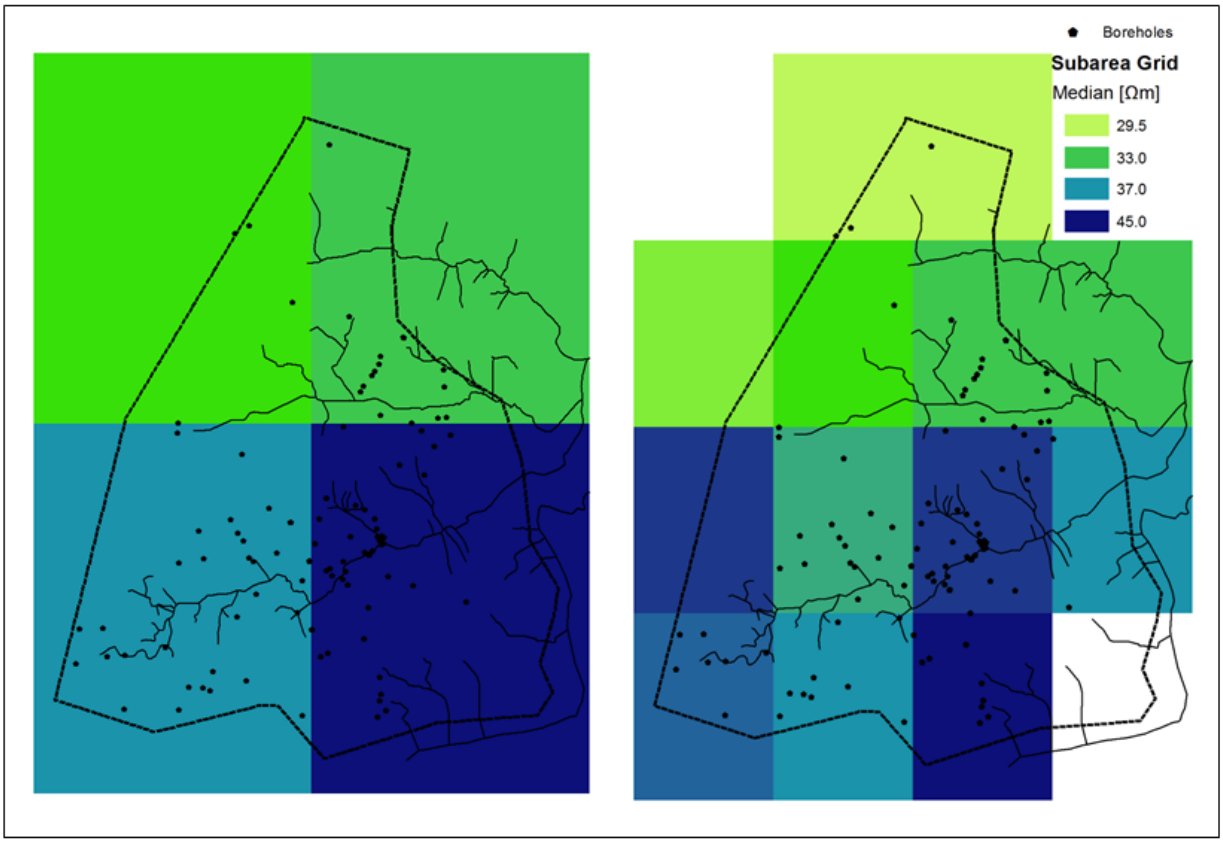

Figure 2. The median resistivity values from the SkyTEM data for the 4- and 16- subarea grid. Dark colors indicate a high median (max: 43.2 Ωm and 45.0 Ωm for the 4- and 16- subarea grid, respectively), light colors a low median (min: 32.0 Ωm and 29.5 Ωm for the 4- and 16- subarea grid, respectively) and white colors the absence of data. Additionally the location of the boreholes, the river network and the delineated glacial structure. The extent is

9km in X- and 12km in Y-direction.

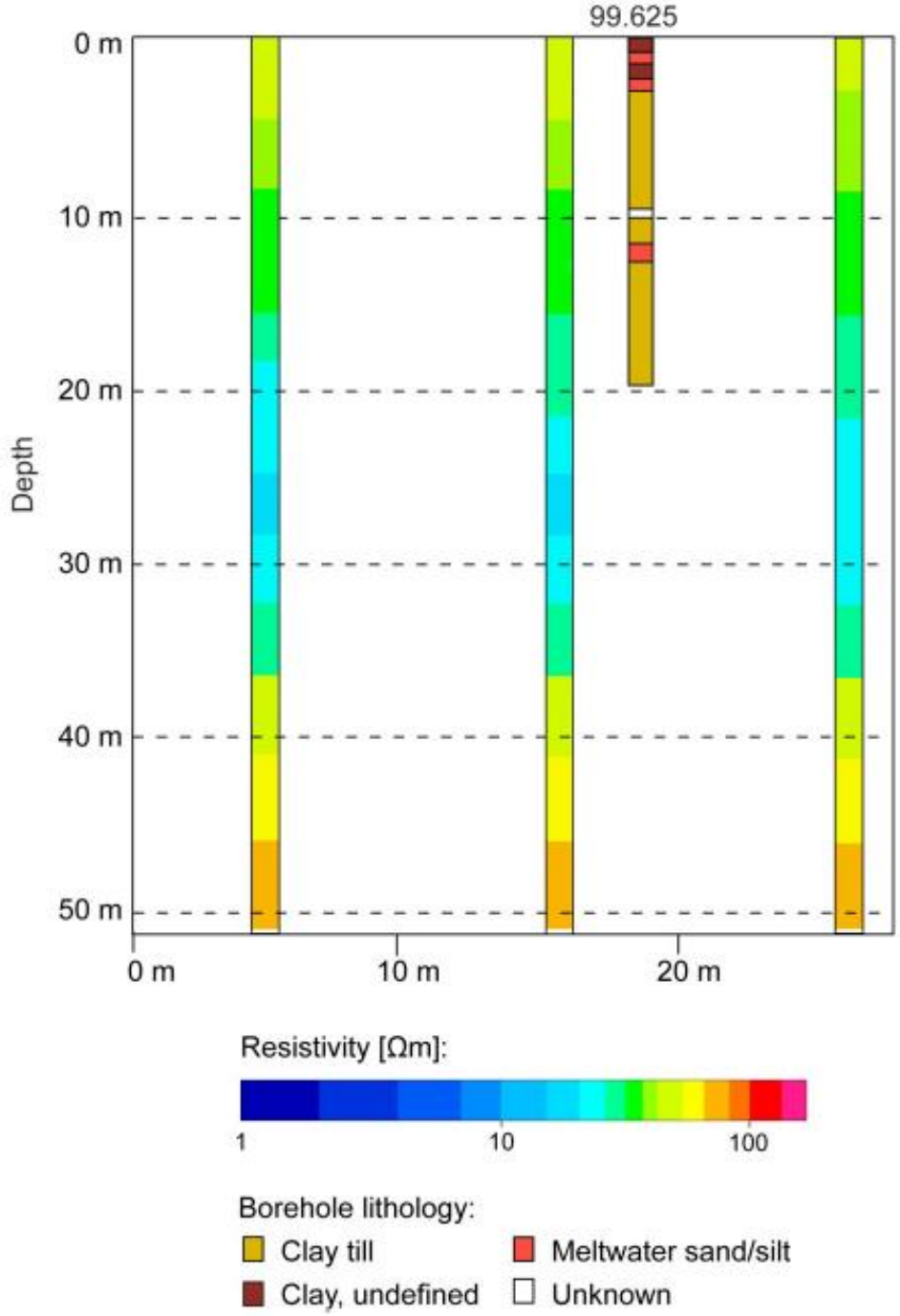

Figure 3 Side-by-side comparison of borehole lithological data and SkyTEM vertical sounding data at borehole number 99.625 (He. et. al, 2014).

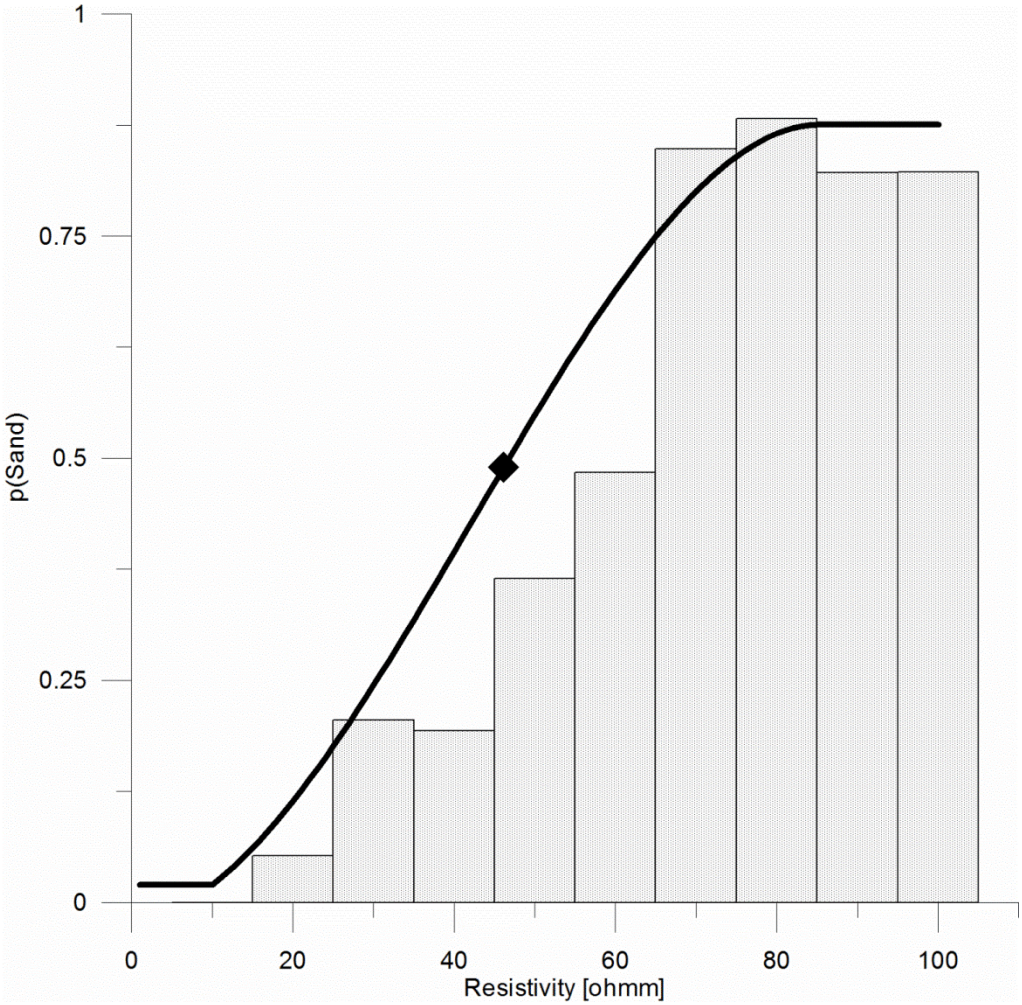

Figure 4. The bias corrected histogram curve: The calibrated cut off value (46 Ωm) is added to the histogram and the fitted curve is forced to honor it He et al. (2014).

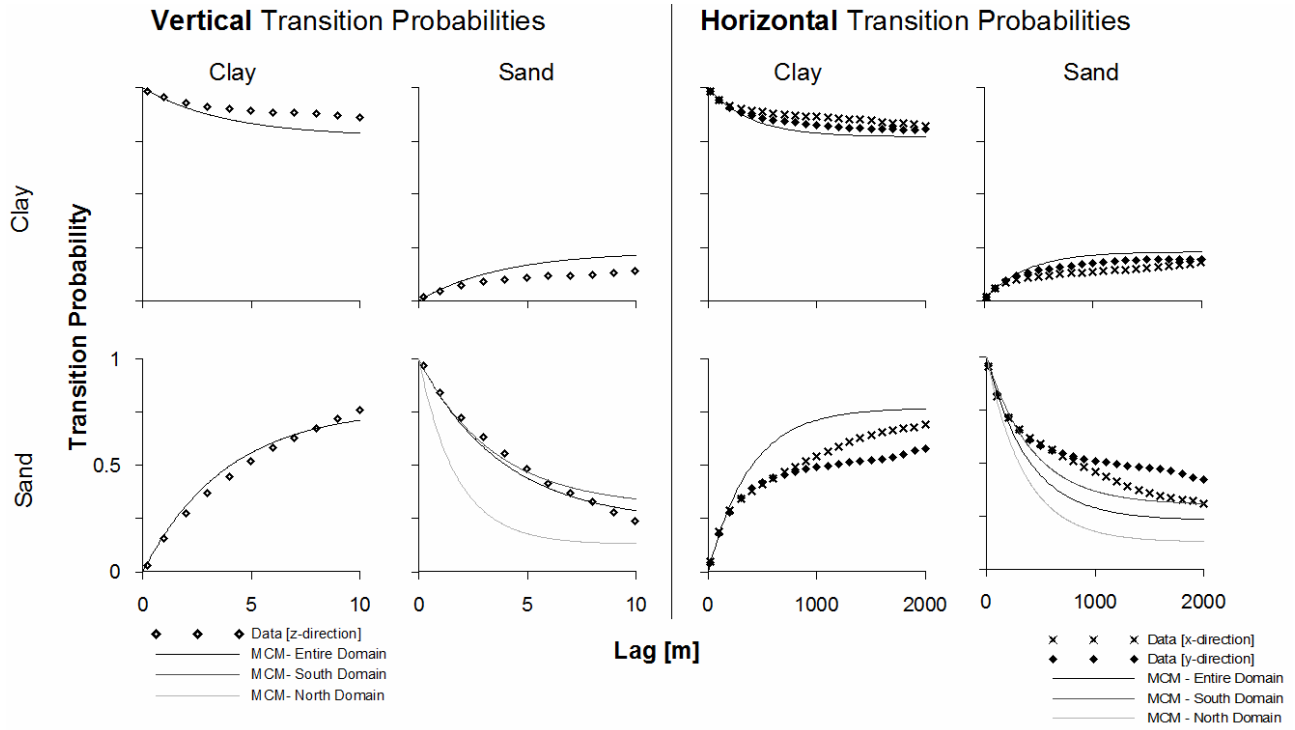

Figure 5. The computed transition probabilities in vertical and horizontal direction and the fitted MCM: Vertical 5m, horizontal 500m mean length of a sand lens and 23% sand proportion. Additionally the fitted MCM for the north- and south-sub-domain are plotted for the vertical and horizontal sand-sand transitions: 2m, 400m, 13% and 5m, 600m, 30%, respectively.

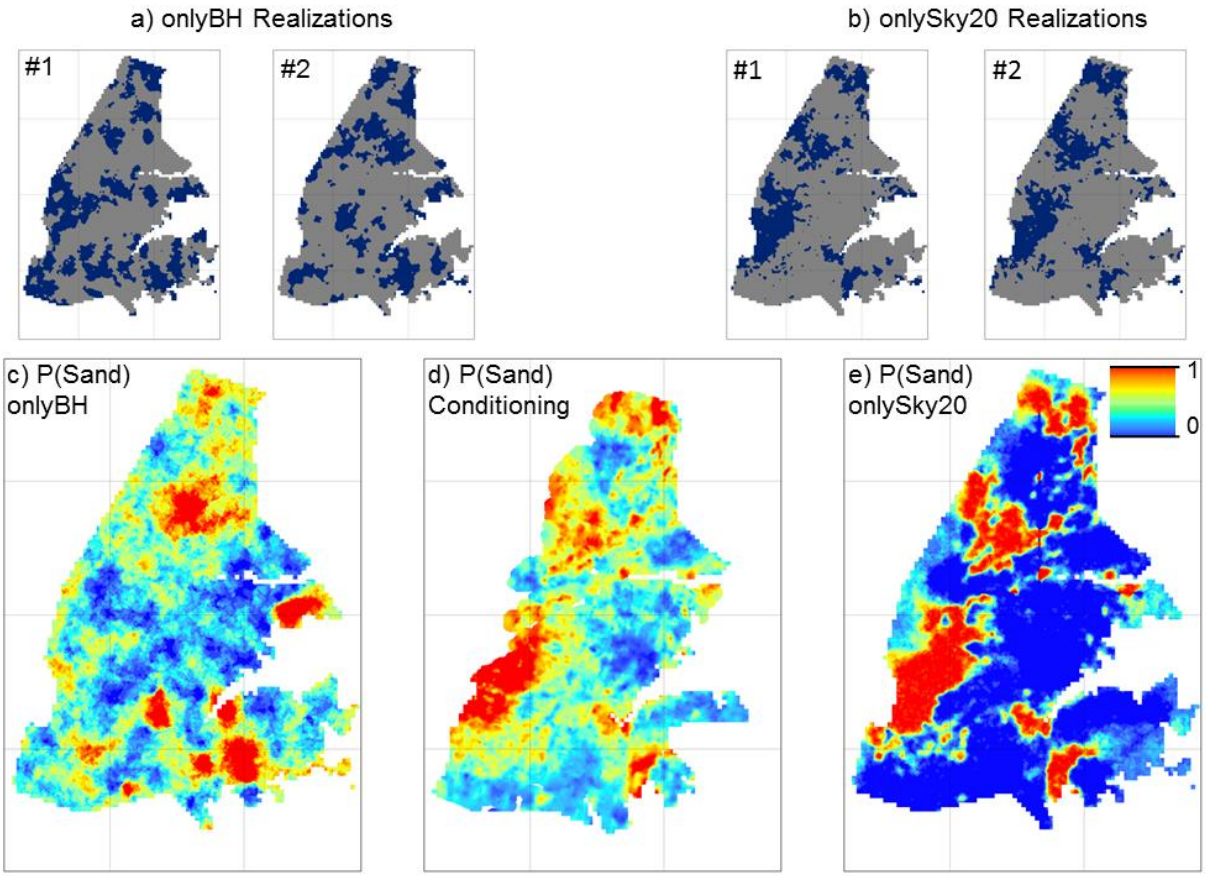

Figure 6. Upper panel: Two individual realizations for two different conditioning scenarios: *onlyBH* (a) and *onlySky20* data (b). Lower panel: Probability maps for the two scenarios c) and e) showing the probability of sand in each cell based on 25 realizations. The derived sand probability which is used for conditioning the simulation is shown in (d). All maps show data at an elevation of 49m.

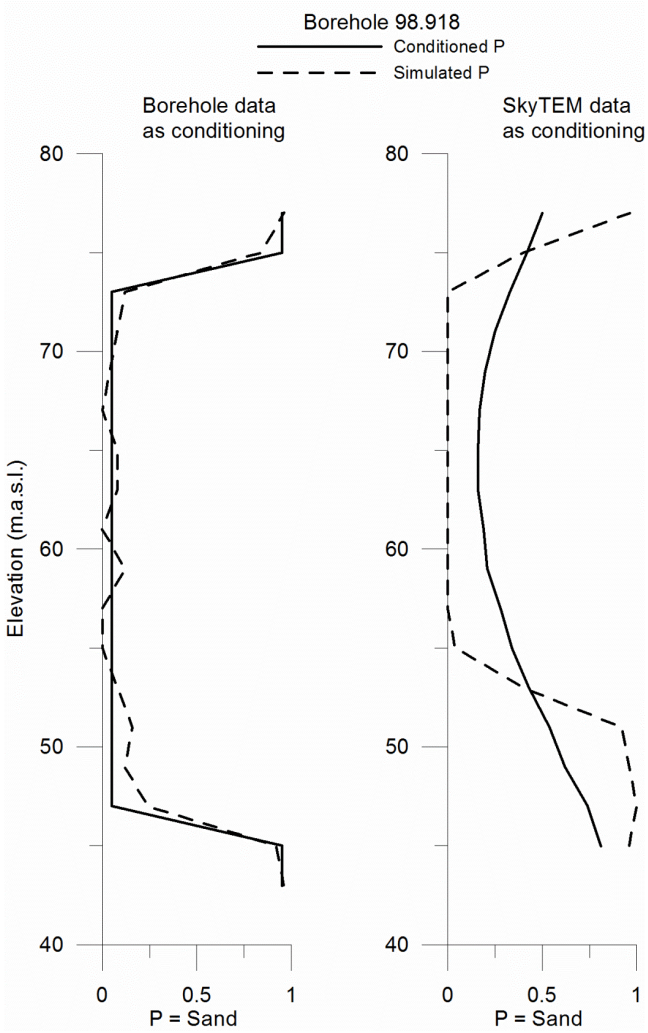

Figure 7. The simulated versus the conditioned sand probability over the vertical extent at one borehole (98.918), located in the south western part of the glacial structure. The results originate from the two different soft conditioning scenarios: *onlyBH* and *onlySky20* (based on 25 realizations each).

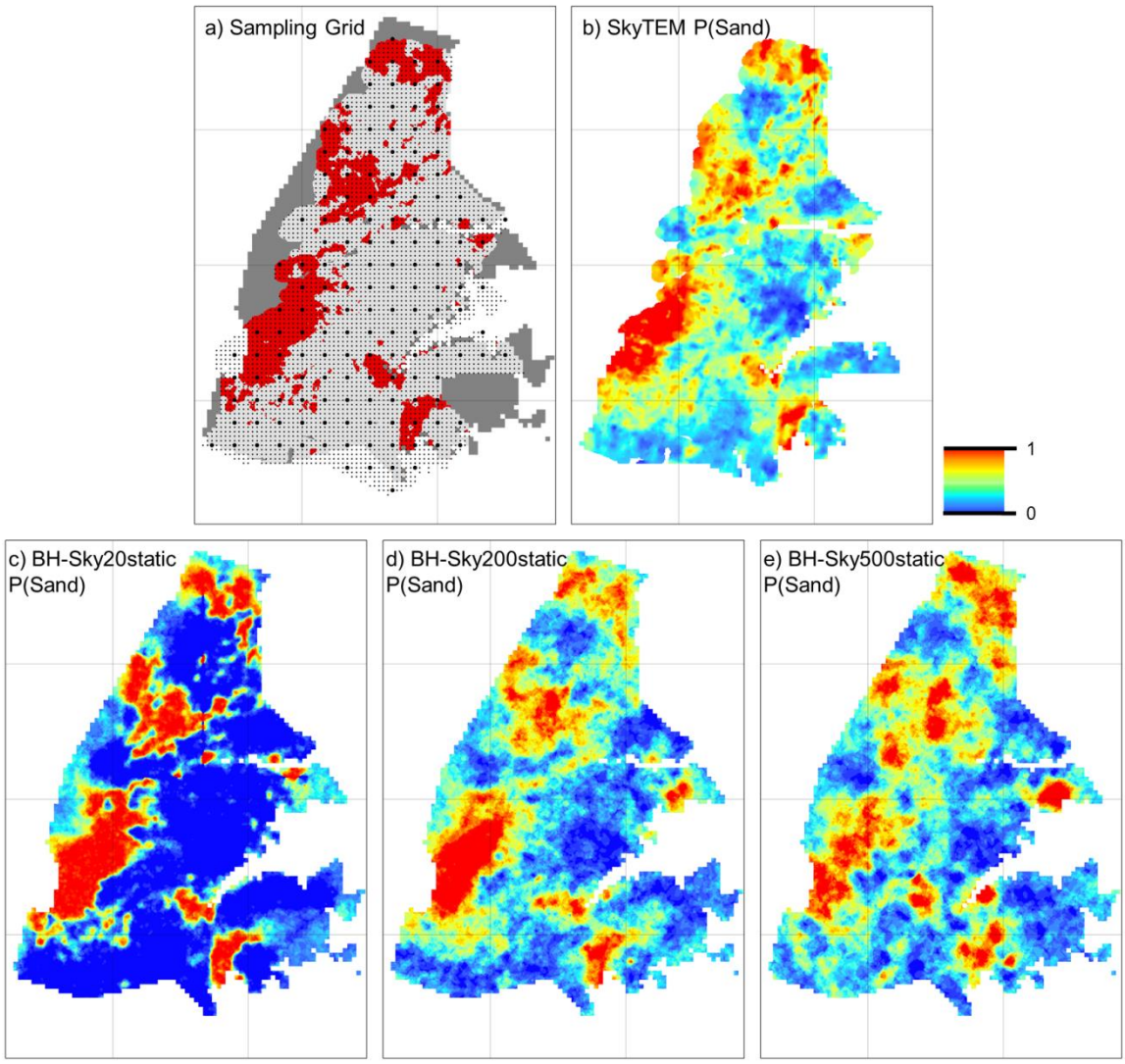

Figure 8. a): 100 m (small dots) and 500 m (big dots) sampling grids for thinning out the conditioning dataset; b-e): probability of sand at an elevation of 49 m for SkyTEM dataset (b), and for static 20m, 200m and 500m conditioning (c-e) Red colors represent high sand probability and blue colors low sand probability (based on 25 realizations).

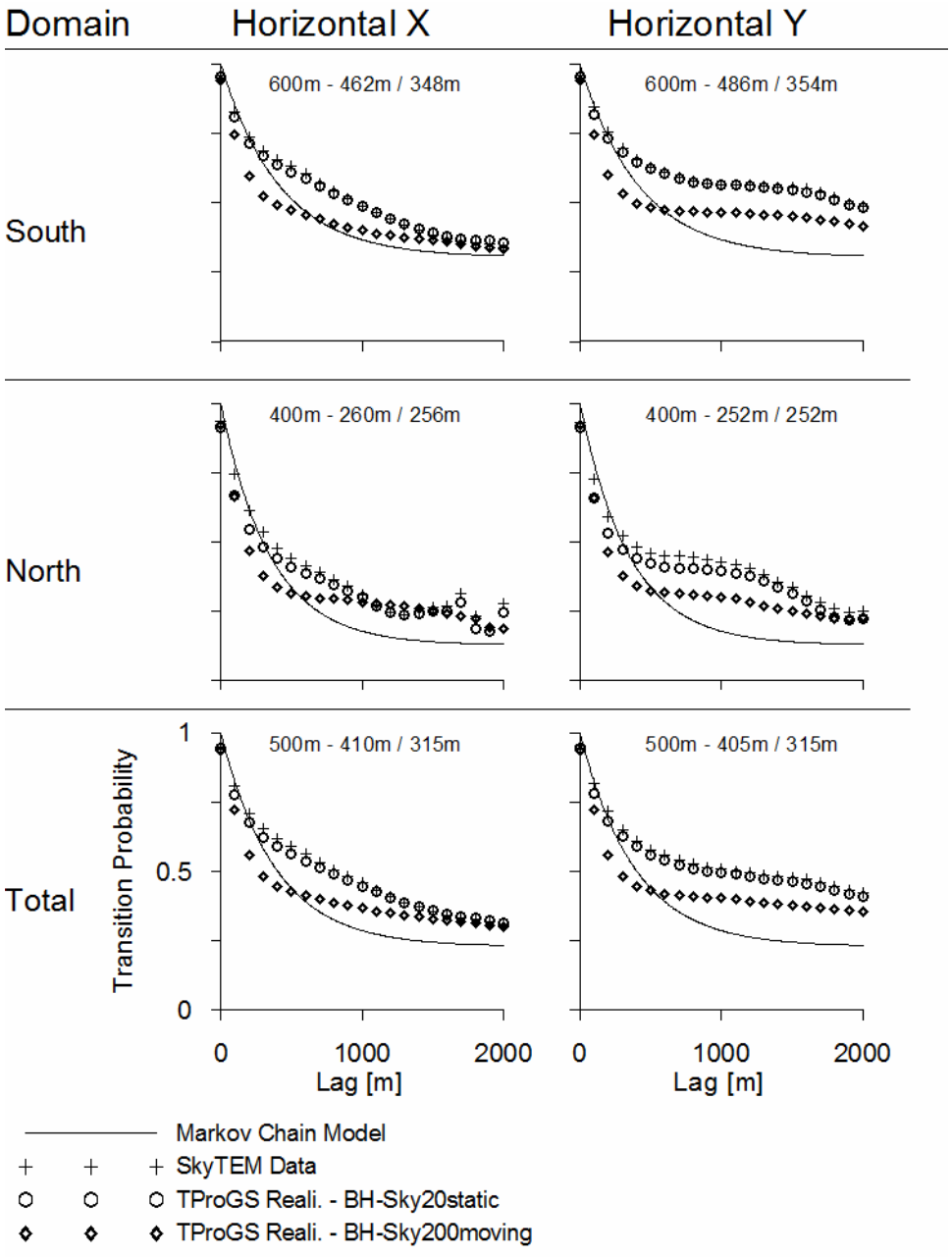

Figure 9. The simulated transition probabilities for the south-, north-, and total-domain are compared with the SkyTEM data and the fitted MCM. The results for two soft conditioning dataset are shown: *BH-Sky20static* and *BH-Sky200moving*. The simulated TP and the MCM at lag 100m are compared to quantify the underestimation of a sand lens. The TP values are mean values based on 10 realizations. The defined length of a sand lens (X) and the mean simulated length for the *BH-Sky20static* (Y) and *BH-Sky200moving* scenario (Z) are given in each graph. (Xm – Ym / Zm).

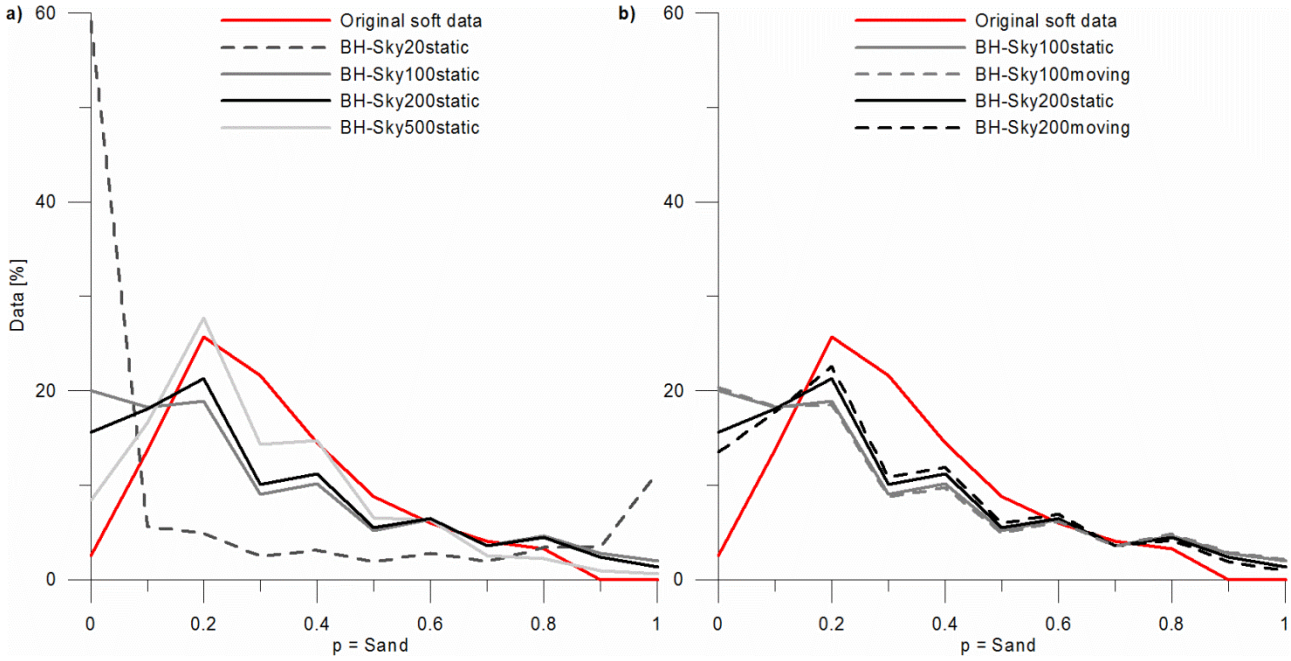

Figure 10. The simulated facies probability distributions based on sets of realizations conditioned to differently sampled soft datasets (based on 25 realizations): (a) static sampling at different sampling distances and (b) stationary and moving sampling at different sampling distances. Also showing the sand probability distribution of the original soft dataset which is desired to be reproduced.

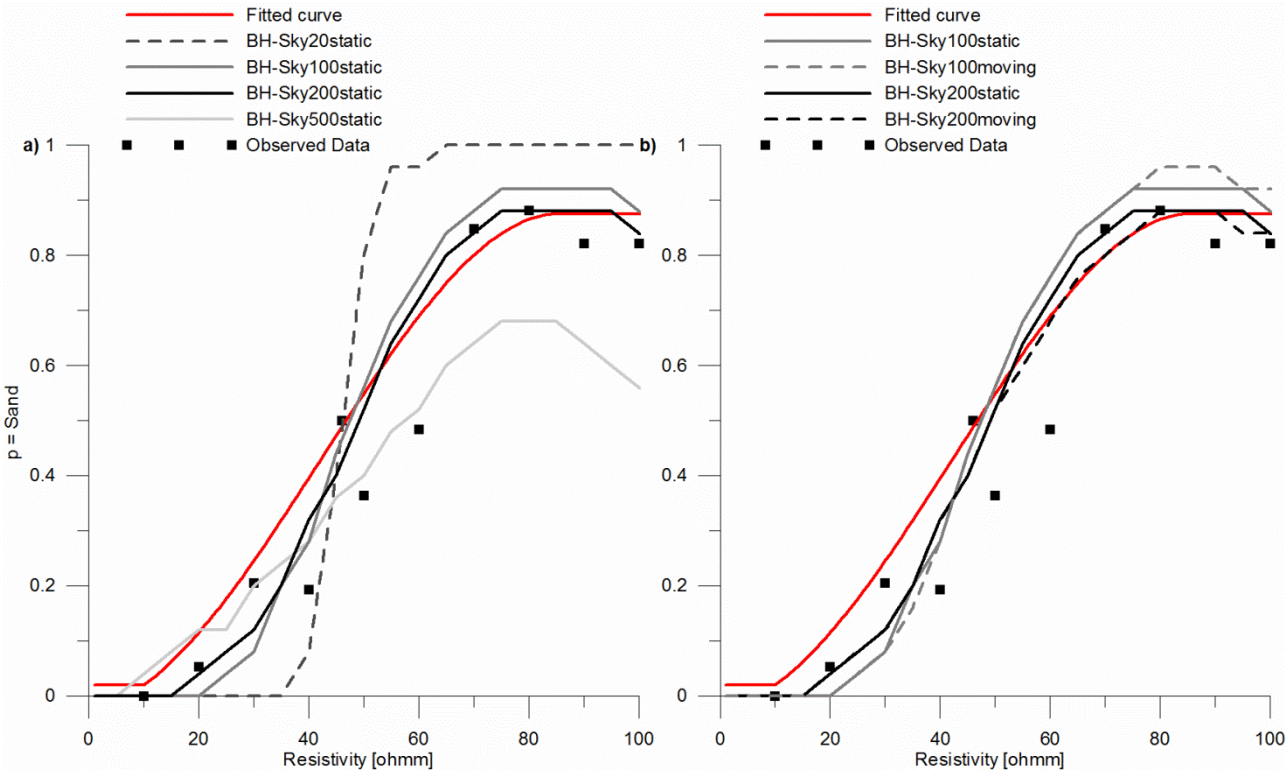

Figure 11. The simulated facies probability – resistivity bias based on sets of realizations conditioned to differently sampled soft datasets (based on 25 realizations): (a) static sampling at different sampling distances and (b) stationary and moving sampling at different sampling distances. The simulated sand probability is paired with the original resistivity value, grouped into 5 Ωm bins and then plotted as median for each bin. Also showing the observed data and the fitted curve from the histogram which is desired to be reproduced.