

Sensitivity and uncertainty analysis of a regional-scale groundwater flow model featuring coal seam gas extraction

Chris Turnadge, Dirk Mallants and Luk Peeters

This report was commissioned by the Department of the Environment and Energy and was prepared by CSIRO.

2018



Copyright

© Commonwealth Scientific and Industrial Research Organisation 2018. To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO.

Citation

Turnadge C, Mallants D and Peeters L (2018) Sensitivity and uncertainty analysis of a regional-scale groundwater flow system stressed by coal seam gas extraction. CSIRO, Australia.

Important disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

CSIRO is committed to providing web accessible content wherever possible. If you are having difficulties with accessing this document please contact csiroenquiries@csiro.au.

Contents

Acknowledgmentsxvi			
Executive summaryxvii			
1	Introduction		1
	1.1	Terms of reference	1
	1.2	Rationale	2
2	CDM Sr	nith (2016) Gunnedah and Surat basins groundwater flow model	. 15
	2.1	Geographic context	. 15
	2.2	Geology	. 16
	2.3	Hydrostratigraphy	. 16
	2.4	Inflows and outflows	. 21
	2.5	Representation of coal seam gas extraction	. 21
	2.6	Numerical solution scheme	. 22
	2.7	Model modifications	. 22
3	Initial p	rediction sensitivity and uncertainty analyses	. 23
	3.1	Model predictions	. 23
	3.2	Model parameters	. 23
	3.3	Model stress testing	. 23
	3.4	Prediction sensitivity and uncertainty analysis - Methods	. 24
	3.5	Prediction sensitivity analysis - Results and Discussion	. 30
	3.6	Prediction uncertainty analysis - Results and Discussion	. 37
4	Improv	ed characterisation of aquitard vertical hydraulic conductivity	. 42
	4.1	Core testing	. 42
	4.2	Parameter upscaling	. 42
5	Predicti	on sensitivity and uncertainty analyses using revised models	. 48
	5.1	Prediction sensitivity and uncertainty analysis - Methods	. 48
	5.2	Prediction sensitivity analysis – Results and Discussion	. 50
	5.3	Prediction uncertainty analysis – Results and Discussion	. 56
6	Analyse	es of the spatial variability of aquitard hydraulic properties	. 63
	6.1	Spatial distributions of aquitard vertical hydraulic conductivity	. 63
	6.2	Variogram analyses	. 65

	6.3	Stochastic generation of spatially distributed parameter fields
	6.4	Prediction uncertainty analysis using spatially distributed parameter fields 71
	6.5	Prediction uncertainty and groundwater management78
7	Data wo	orth analysis82
	7.1	Literature review
	7.2	Bootstrapping resampling
8	Summa	ry and conclusions
References		
9	Append	lix 1
	9.1	Initial model: Magnitude of maximum drawdown prediction (MXD) 100
	9.2	Revised model: Magnitude of maximum drawdown prediction (MXD)
	9.3	Initial model: Timing of maximum drawdown prediction (tMXD) 102
	9.4	Revised model: Timing of maximum drawdown prediction (tMXD) 103
	9.5	Initial model: Number of model cells with drawdown > 2 m prediction (NDD) 104
	9.6 (NDD)	Revised model: Number of model cells with drawdown > 2 m prediction
	9.7	Initial model: Maximum vertical flux prediction (MXQ)106
	9.8	Revised model: Maximum vertical flux prediction (MXQ) 107

Figures

Figure 1-1 Sensitivity and uncertainty analysis flowchart adopted in this study
Figure 1-2. Flowchart of the workflow employed in this study to assess the effects of improved aquitard characterisation (red triangular distributions) on parameter sensitivity (not shown) and prediction uncertainty (one hypothetical flow metric shown)
Figure 2-1 Spatial extent of Gunnedah geological basin (NSW) and numerical grid of groundwater flow model
Figure 2-2 Water production for Maules Creek Formation and Hoskissons Coal seam target formations as predicted by the CDM Smith (2016) groundwater flow model
Figure 3-1. Graphical representation of the Delta Moment Independent Measure (DMIM; Plischke et al., 2013) of global sensitivity (Borgonovo, 2007). The exponential distribution (red solid line) represents the probability density function (PDF) of modelled predictions when all model parameters are varied simultaneously. The skewed Gaussian distribution (dashed blue line) represents the PDF of modelled predictions when all model parameters are varied except a given parameter of interest. The difference in mass density between these two PDFs (solid black shading) is used as the basis for the calculation of the DMIM sensitivity metric
Figure 3-2. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which predictions relating to the Pilliga Sandstone aquifer were most sensitive: (a) maximum drawdown (MXD); (b) time elapsed at which maximum drawdown occurred (tMXD); (c) number of model cells at which drawdown exceeded two metres; and (d) maximum change in vertical flux (MXQ). Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs. Parameters identified are discussed in Section 3.5
Figure 3-3. Statistical distributions of four predictions relating to the Pilliga Sandstone aquifer and simulated using the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model:

Figure 3-4. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the magnitude of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.

Figure 3-5 Global sensitivity analysis metrics of all 30 model parameters in relation to maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation was 300 model runs. 32

Figure 3-6. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the timing of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.

 Figure 3-13. (a) Cumulative density function and (b) frequency histogram of initial modelled predictions of the timing of maximum drawdown, based on a sample size of 300 model runs. 39

 Figure 5-4. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the magnitude of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.

Figure 5-5 Global sensitivity analysis metrics of 20 model parameters in relation to maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation was 300 model runs. 52

Figure 5-8. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of drawdown spatial extent in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001).

Figure 5-9 Global sensitivity analysis metrics of 20 model parameters in relation to spatial extent of maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation is 300 model runs. 54

Figure 5-10. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the maximum change in vertical flux in the Pilliga Sandstone aquifer was most sensitive. Parameter

rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment- Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001)
Figure 5-11 Global sensitivity analysis metrics of 20 model parameters in relation to maximum vertical flux. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation is 300 model runs. 56
Figure 5-12. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the magnitude of maximum drawdown, based on a sample size of 300 model runs
Figure 5-13. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the timing of maximum drawdown, based on a sample size of 300 model runs
Figure 5-14. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the spatial extent of drawdown in excess of 2 m, based on a sample size of 300 model runs
Figure 5-15. Percentiles of spatial distributions of drawdown (solid black lines), observed at the median time of maximum drawdown (i.e., 155 years after the cessation of coal seam gas extraction); purple = Pilliga Sandstone aquifer cells; blue = upper aquitard sequence cells. Percentiles shown are (a) 90 th , (b) 95 th and (c) 99 th . Distributions are based on a set of 300 model runs using the 'revised' model
Figure 5-16. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of changes in vertical groundwater fluxes, based on a sample size of 300 model runs
Figure 6-1. Spatial distributions of upscaled (a) harmonic and (b) numerical mean values for the upper aquitards (Jurassic age). Grey shaded area is extent of aquitard. Arrow indicates focus study area with largest data density
Figure 6-2. Spatial distributions of upscaled (a) harmonic and (b) numerical mean values for the lower aquitards (Permian age). Grey shaded area is extent of aquitard. Arrow indicates focus study area with largest data density
Figure 6-3. Experimental variograms (red, based on semi-variance values) and correlograms (blue, based on autocorrelation values) for the (a) upper aquitard sequence and (b) lower aquitard sequence, using vertical hydraulic conductivity values that were upscaled using numerical model averaging
Figure 6-4 Theoretical variogram models (spherical, exponential, Gaussian) with indication of the practical range (Bohling, 2005)
Figure 6-5. Geostatistical analysis of (a) upper aquitard sequence and (b) lower aquitard sequence K_V values upscaled using a 1-D numerical groundwater flow model. Grey closed circles = data points $[z(x_i+h)-z(x_i)]^2$, red closed circles and lines = experimental variograms, solid green line = spherical variogram model, and dashed blue line = exponential variogram model

Figure 6-6 Multiple scales of heterogeneity. (top) Identification of scales: regional scale (\mathcal{L}), scale of the flow domain (L), and local-scale correlation scale (λ). (bottom) Hypothetical multi-scale semi-variogram corresponding with multiple scales of heterogeneity (modified from Ababou et al. [1989])
Figure 6-7. Heterogeneous spatial distributions of log_{10} vertical hydraulic conductivity ($log_{10} K_V$) for the (a) upper and (b) lower aquitard sequences generated using Sequential Gaussian Simulation based on spherical variogram models and conditioned to values upscaled using numerical model averaging (red = low values; blue = high values; white = non-Pilliga Sandstone aquifer cells)
Figure 6-8. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the magnitude of maximum drawdown, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous')
Figure 6-9. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the timing of maximum drawdown, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous')
Figure 6-10 Maximum drawdown as a function of time of maximum drawdown for homogenous (300 runs) and heterogeneous (50 runs) model. Pink ellipse indicates decreasing trend between maximum drawdown and time of maximum drawdown. Note: clustering of filled red circles at $t = 160$ y was due to the limited temporal extent specified for model runs (see text for further details)
Figure 6-11. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the spatial extent of drawdown in excess of 2 m, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous')
Figure 6-12. Percentiles of spatial distributions of drawdown (solid black lines), observed at the median time of maximum drawdown (i.e., 155 years after the cessation of coal seam gas extraction); purple = Pilliga Sandstone aquifer cells; blue = upper aquitard sequence cells. Percentiles shown are (a) 90 th , (b) 95 th and (c) 99 th . Distributions are based on a set of 50 model runs that featured spatially variable parameterisation of aquitard vertical hydraulic conductivity values (i.e., the 'heterogeneous' model)
Figure 6-13. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of changes in vertical groundwater fluxes, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous')
Figure 6-14 Relationship between impact thresholds and acceptable probability of exceeding a given threshold
Figure 6-15 Cumulative distribution function (CDF) and complementary cumulative distribution function (CCDF) for variable v with a triangular distribution on [1, 10] and a mode at 7 (based on Helton et al. 2004)
Figure 6-16 Boundary line approach to specification of acceptable risk (modified from Helton and Breeding, 1993)

Figure 6-17 Comparison of CCDF for maximum drawdown and single requirement boundary lines for 30% (A), 50% (B), and 80% (C) acceptable probability of exceeding 2 m drawdown.	
Multiple requirement boundary line considers 50% acceptable probability of exceeding for 0.2 m, 10% for 2 m and 1% for 10 m drawdown (D)	
Figure 7-1. Illustration of the bootstrapping approach to estimate parameter (a) and model prediction uncertainty (b) based on a sample size of 50	
Figure 7-2. Illustration of the bootstrapping approach to estimate parameter (a) and model prediction uncertainty (b) based on a sample size of 300	
Figure 7-3 Effect of sample size on robustness (90 % confidence interval) of estimated percentiles of maximum drawdown	,
Figure 7-4 Illustration of the bootstrapping approach to estimate model prediction uncertainty (maximum drawdown) based on a sample size of 50 for the homogeneous (a) and heterogeneous model (b)	,
Figure 7-5 Illustration of the bootstrapping approach to estimate model prediction uncertainty across different sample sizes (heterogeneous model). Box plots show full data range,	
interquartile range (box) and median (red line)	

Tables

Table 2-1. Stratigraphy of the Gunnedah Basin (CDM Smith, 2016; Geoscience Australia,2016).18
Table 2-2. Hydrostratigraphy of the Gunnedah-Surat basins as represented in the CDM Smith(2016) groundwater flow model.20
Table 3-1. Results of parameter stress testing of the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model. Parameter ranges (for each of K_H , K_V and S_S) were specified as the CDM Smith (2016) parameter value ± two orders of magnitude
Table 3-2. Specified and sampled minimum and maximum parameter values used in preliminary global sensitivity analysis of the preliminary CDM Smith (2016) groundwater flow model. All parameters were log_{10} transformed and random sampling of parameters was undertaken from log-uniform distributions. $K_{\rm H}$ and $K_{\rm V}$ in m/day, SS in m ⁻¹
Table 4-1. Statistical summary of \log_{10} aquitard vertical hydraulic conductivity ($\log_{10} K_V$) values upscaled from core-scale observations using harmonic averaging
Table 4-2. Statistical summary of \log_{10} transformed aquitard vertical hydraulic conductivity ($\log_{10} K_V$) values upscaled from core-scale observations using numerical averaging. K_V in m/day
Table 5-1. Median values of four prediction metrics based on sets of 300 model runs for each ofthe 'initial' and 'revised' models.61
Table 5-2. 95th percentiles of four prediction metrics based on sets of 300 model runs for eachof the 'initial' and 'revised' models
Table 6-1. Variogram model parameter (i.e., range, sill and nugget) values and associated coefficient of determination (R^2) values for power and logarithmic variogram models used to characterise the spatial correlation between upscaled K_V values for the upper and lower aquitard sequences. *Note that "Practical range" refers to the distance at which semi-variance values generated using an exponential variogram model reach 95 % of the sill value
Table 6-2. Summary statistics of 50 spatial distributions of equivalent log ₁₀ aquitard vertical hydraulic conductivity (m/d) generated for the upper and lower aquitard sequences using Sequential Gaussian Simulation
Table 6-3. 50 th percentile (i.e., median) values of four prediction metrics based on sets of 300 model runs for each of the 'initial' and 'revised' models and 50 model runs with the 'heterogeneous' models
Table 6-4. 95 th percentiles of four prediction metrics based on sets of 300 model runs for each of the 'initial' and 'revised' models and 50 model runs with the 'heterogeneous' models 77
Table 7-1 Examples of data worth or value of information applications in groundwaterinvestigations and management
Table 7.2.00% confidence intervals derived for summary statistics nEQ. nZE and nQQ.

Abbreviations

Abbreviation	Description
CSG	Coal Seam Gas
CSIRO	Commonwealth Scientific and Industrial Research Organisation (Australia)
HSU	HydroStratigraphic Unit
NSW	New South Wales
Qld	Queensland

Glossary

Term	Description
Anisotropy	A term used to describe the directional dependence of given properties; for example, the hydraulic properties of an aquifer (as opposed to isotropy, which denotes identical properties in all directions)
Aquifer	Rock or sediment in a formation, group of formations or part of a formation, which is saturated and sufficiently permeable to transmit quantities of water to wells and springs
Aquitard	A saturated geological unit that is less permeable than an aquifer and incapable of transmitting useful quantities of water. Aquitards often form a confining layer over aquifers
Coal measure	Geological strata of the Carboniferous or Permian periods usually containing sequences of coal seams
Coal seam	Individual layers containing mostly coal. Coal seams store both water and gas. Coal seams generally contain groundwater that is saltier than that in aquifers that are used for drinking water or agriculture
Coal seam gas	A form of natural gas (generally 95 to 97% pure methane, CH ₄) typically extracted from permeable coal seams at depths of 300 to 1000 m. Also called coal seam methane (CSM) or coalbed methane (CBM)
Confined aquifer	An aquifer that is isolated from the atmosphere by an impermeable layer. Pressure in confined aquifers is generally greater than atmospheric pressure
Covariance	Covariance is a measure of how much two given variables vary together, as a function of either space or time
Darcy flow	Liquid flow that conforms to Darcy's law
Darcy's law	A constitutive equation that describes the flow of a fluid through a porous medium such as rock or soil
Depressurisation	The lowering of static groundwater levels through the partial extraction of available groundwater, usually by means of pumping from one or several groundwater bores or gas wells
Dewatering	The lowering of static groundwater levels through complete extraction of all readily available groundwater, usually by means of pumping from one or several groundwater bores or gas wells
Dirichlet boundary condition	Also known as a first type boundary condition, involves specification of the value that the solution of a differential equation needs to produce along the boundary of a model domain. Applicable to both numerical and analytical models
Gaussian (probability distribution)	A continuous function that approximates the exact binomial distribution and which represents the statistical distribution of many random variables. This can be described using only two parameters: mean (i.e. central tendency) and variance (i.e. spread). Typically visualised as a symmetrical bell-shaped graph
Groundwater	Water occurring naturally below ground level (whether in an aquifer or other low- permeability material), or water occurring at a place below ground that has been pumped,

Term	Description
	diverted or released to that place for storage. This does not include water held in underground tanks, pipes or other works
Groundwater (single phase) flow model	A numerical solution to a partial differential equation used to describe the flow of water in the subsurface. Groundwater flow models involve the flow simulation of a single fluid phase (i.e. water). Common parameters used in groundwater flow models are hydraulic conductivity, specific yield and specific storage
Hydraulic conductivity	A coefficient of proportionality describing the rate at which a fluid can move through a permeable medium
Hydraulic gradient	The difference in hydraulic head between different locations within or between hydrostratigraphic units, as indicated by water levels observed in wells constructed in those units
Hydraulic head	The potential energy contained within groundwater as a result of elevation and pressure. It is indicated by the level to which water will rise within a bore constructed at a particular location and depth. For an unconfined aquifer, it will be largely subject to the elevation of the water table at that location. For a confined aquifer, it is a reflection of the pressure that the groundwater is subject to and will typically manifest in a bore as a water level above the top of the confined aquifer, and in some cases above ground level
Hydraulic pressure	The total pressure that water exerts on the materials comprising the aquifer. Also known as pore pressure
Hydrostratigraphic unit	A formation, part of a formation, or group of formations of significant lateral extent that compose a unit of reasonably distinct (similar) hydrogeologic parameters and responses
Interburden	Material of any nature that lies between two or more bedded ore zones or coal seams
Intrinsic permeability	The permeability of a given medium independent of the type of fluid present
Isotropy	The condition in which the hydraulic properties of a hydrostratigraphic unit are equal in all directions
Kriging	A geostatistical method of spatial interpolation (i.e. prediction) using the weighted mean value of surrounding data points. The data are a set of observations with some spatial correlation present
Monte Carlo sampling	The sampling of uncertain data for use in Monte Carlo risk analysis or simulation
Monte Carlo simulation	The use of Monte Carlo analysis techniques to estimate the most probable outcomes from a model with uncertain input data
Numerical realisation	A numerically generated sample (usually of model parameters) drawn from a probability distribution, used to run a model simulation
Permeability	The measure of the ability of a rock, soil or sediment to yield or transmit a fluid. The magnitude of permeability depends largely on the porosity and the interconnectivity of pores and spaces in the ground
Porosity	The proportion of the volume of rock consisting of pores, usually expressed as a percentage of the total rock or soil mass
Probability density function	A function that describes the relative likelihood for a random variable to take on a given value

Term	Description
Regional-scale groundwater models	Models that encompass an entire groundwater system, geological basin or other significant area of interest that extends well beyond the measurable influence of individual bores or borefields
Reservoir (hydrocarbon)	Porous or fractured rock formations that contain significant reserves of hydrocarbons. Naturally-occurring hydrocarbons such as crude oil or natural gas are typically trapped in source or host rocks by overlying low permeability formations
Robustness (of model predictions)	Insensitivity of model predictions to data outliers or other small departures from assumptions required by a predictive model, including the types of parametric distributions assumed
Saturated flow	Flow through a porous medium (such as soil or rock) in which the void space within the porous medium is entirely occupied by water (as opposed to water and gas)
Single phase flow	The flow of a single phase, e.g. liquid or gas
Spatial correlation	Spatial dependency (or correlation) between samples
Spatial interpolation	The procedure of estimating the value of properties at unsampled sites within the area covered by existing observations
Stratigraphy	An arrangement of sedimentary, metamorphic and/or igneous rocks
Transmissivity	The rate at which a fluid is transmitted through a unit width of a hydrostratigraphic unit under a hydraulic gradient
Unconfined aquifer	An aquifer in which there are no confining beds between the zone of saturation and land surface
Unconventional gas	Natural gas found in a very low permeability rock, such as coal seam gas, shale gas, and tight gas. Unconventional gas such as coal seam gas is trapped in coal beds by adsorption of the gas molecules to the internal surfaces of coal. It cannot migrate to a trap and form a conventional gas deposit. This distinguishes it from conventional gas resources, which occur as discrete accumulations in traps formed by folds and other structures in sedimentary layers
Upscaling	Upscaling is the process of transforming the detailed description of hydraulic parameters in a grid constructed at measurement scale to a coarser grid with less detailed description. It replaces a heterogeneous domain with a homogeneous one in such a way that both domains produce the same response under some upscaled boundary conditions
Variogram (also semi- variogram)	A function describing the spatial dependency (similarity) between observations of a variable. The shape of the variogram is typically a function of the distance and direction separating observations at two locations; at short distances the semi-variance is small, and typically increases with increasing separation distance. The semi-variance is defined as the variance of the difference between two variables at two locations. At zero separation distance the semi-variance is called nugget (-effect). The sill is the maximum semi-variance or the plateau of the semi variogram; the correlation length or spatial range is the distance over which variables are spatially correlated
Well	Borehole in which a casing (e.g. steel piping) has been placed to restrict connection to specific ground horizons/depths

Symbols

Symbol	Brief description and unit of measurement
А	Range (variogram model parameter)
С	Sill (variogram model parameter)
δ	Delta Moment-Independent Measure global sensitivity metric
γ	Semi-variance
h	Variogram lag distance
k	Permeability [mD or m ²]
k _V	Vertical permeability [mD or m ²]
К	Hydraulic conductivity [m.day ⁻¹]
K _H	Horizontal hydraulic conductivity [m.day ⁻¹]
K _V	Vertical hydraulic conductivity [m.day ⁻¹]
n	Nugget (variogram model parameter)
Q	Volumetric fluid flux [m³.day ⁻¹]
S	Storativity [dimensionless]
Ss	Specific storage [m ⁻¹]
S _Y	Specific yield [dimensionless]
<i>S</i> ₁	First order Sobol' global sensitivity metric
Z	Sample depth [m]

Acknowledgments

The authors would like to acknowledge the Commonwealth Department of the Environment and Energy for funding the project "Research to improve treatment of faults and aquitards in Australian regional groundwater models to improve assessment of impacts of CSG extraction".

The report was subject to internal peer review processes during its development and benefitted from reviews undertaken by Dr. Elmar Plischke (Technische Universität Clausthal, Germany), Dr. Sreekanth Janardhanan (CSIRO Land and Water), Dr. Juan Castilla-Rho (CSIRO Land and Water), Prof. Jim Underschultz (University of Queensland), Prof. Craig Simmons (Flinders University of South Australia), and Dr. Rod Dann (Department of the Environment and Energy).

Executive summary

The project "Research to improve treatment of faults and aquitards in Australian regional groundwater models to improve assessment of impacts of coal seam gas (CSG) extraction" focuses on method development to underpin the risk assessments associated with deep groundwater extraction and depressurisation from energy resource development. The project aims to develop methodologies and techniques that will improve the predictive capability of regional groundwater models used in this context, specifically with respect to the representation of faults and aquitards. The project has three components: 1) an examination of aquitards, 2) an examination of faults, and 3) an examination of the upscaling of aquitard and fault properties such that they can be adequately represented in regional groundwater flow models.

The present report addresses certain aspects of components 1 and 3, that is, to improve understanding of the vertical hydraulic conductivity of aquitards by better conceptualisation, parameterisation and representation of aquitards in regional groundwater models. Specifically this involves the inclusion of upscaled aquitard hydrogeologic parameter values in a numerical model of groundwater flow in the Gunnedah basin and overlying parts of the Surat basin in the New England region of New South Wales, Australia. The model considers the most up-to-date hydrogeological conceptualisation of the study area, including the Jurassic to Late Permian upper aquitard sequence (comprising amongst others the Purlawaugh and Napperby formations) and the lower Mid to Late Permian aquitard sequence (with the Porcupine Formation at its base overlain by the Watermark Formation). Each of these aquitards isolates a coal seam gas target formation from overlying water bearing formations: the Maules Creek coal layers are isolated by the lower aquitard sequence. The key overlying water-bearing formations are the Pilliga Sandstone confined aquifer and the Namoi Alluvium unconfined aquifer.

Although there are a number of geological and parameter uncertainties early in the resource development cycle that could significantly affect the impact of CSG development on adjacent aquifers separated by aquitards, this project chose to only examine a particular subset of these. Choosing to accept a single geological static model realization and the available petrophysical logs and core data from an existing number of well penetrations, this project examined how three different upscaling and parameterization approaches would affect four different groundwater impact predictions. With the available budget and time resources, the project could only examine a limited number of approaches and the results should be considered an example of how parameterisation uncertainty is propagated in predictive outcome uncertainty. The magnitude and associated uncertainties of four key predictions generated by the model (i.e., groundwater impact metrics) were assessed using three parameterisation approaches that were applied sequentially to the same groundwater flow model. Specifically, the four predictions of interest were: the magnitude, timing and spatial extent of maximum drawdown in the Pilliga Sandstone aquifer and the maximum vertical flux across the base of the same aquifer.

The first parameterisation approach used existing model parameters values (i.e., horizontal and vertical hydraulic conductivity [K_H , K_V] and specific storage coefficient [S_S]) derived from prior

modelling studies, with spatially homogeneous parameter values for every model layer. In the second approach vertical hydraulic conductivity data for the two aquitard sequences were generated by combining core-scale K_V values with wireline log data. Those K_V values were subsequently upscaled to the regional scale commensurate with the cellular grid size of the model using analytical and numerical upscaling methods, and incorporated in the flow model. Every hydrostratigraphic unit was assigned a spatially homogeneous average K_V value. The third parameterisation approach featured an alternative, more detailed conceptual model of the aquitards' hydraulic conductivity whereby K_V values were spatially distributed according to an observed spatial correlation model derived from data generated using the second approach. This approach generated heterogeneous K_V domains, while honouring the data at all the observation wells where upscaled K_V values had been determined. The upscaled aquitard K_V values were used to update the prior parameter distributions used in the first parameterisation approach to generate updated model prediction (posterior) distributions using either homogeneous or heterogeneous K_V values.

Global sensitivity methods based on Monte Carlo sampling were adopted to generate and evaluate 300 parameter sets (once for the first and once for the second parameterisation approach). The resulting sets of 300 model predictions were used to identify those hydrogeological parameters to which the four groundwater impact metrics were most sensitive and to quantify the uncertainty of predictions that arises from parameter uncertainty. Parameter sensitivity rankings were based on the Delta Moment-Independent Measure and the first order Sobol' metric. Prediction uncertainties were quantified prior to and following the inclusion of updated aquitard K_V data. This allowed testing of the impact of improved aquitard characterisation on prediction uncertainty. The stochastic approach adhered to in this project to quantify prediction uncertainty is consistent with the Bayesian paradigm, in which prior beliefs (i.e., estimates of parameters and conceptualisation) are iteratively updated when new data and information becomes available.

Results from both sensitivity analyses are summarised for each of the four groundwater impact metrics as follows (based on sets of 300 model runs for each parameterisation approach):

- **The magnitude of maximum drawdown** in the Pilliga Sandstone aquifer was found to be most sensitive to the horizontal hydraulic conductivity (*K*_H) of the Namoi Alluvium aquifer (both parameterisation approaches);
- The timing of maximum drawdown using the first parameterisation was found to be insensitive to all parameters. For the second parameterisation, the timing of maximum drawdown was found to be most sensitive to K_V of the upper aquitard sequence and to the K_H of the Namoi Alluvium;
- **The spatial extent of drawdown propagation** was found to be most sensitive to *K*_H of the Namoi Alluvium aquifer (both parameterisations);
- The maximum vertical flux across the base of the Pilliga Sandstone aquifer using the first parameterisation approach was found to be most sensitive to K_V of the upper aquitard sequence. This was also found true for the second approach.

Two of the four impact metrics were found to be sensitive to the K_V of the upper aquitard sequence; this is one of the two aquitard sequences that was part of the site characterisation. On

the basis of the subsequent prediction uncertainty analysis, the uncertainty about the four groundwater impact metrics was quantified in terms of cumulative distribution functions and their statistical attributes (i.e. percentiles). For both aquitard sequences, the revised prior distributions of upscaled K_V were found to be two orders of magnitude larger (i.e., from $\log_{10} K_V$ -6.7 to -3.2 m/d for the upper aquitard sequence and from $\log_{10} K_V$ -7.5 to -3.1 m/d for the lower aquitard sequence) than the expert opinion-based distributions assumed for the sensitivity analysis with the initial parameter distributions (i.e., from $\log_{10} K_V$ -6.0 to -4.0 m/d for both aquitard sequences).

Furthermore, in the first (initial) parameterisation approach the parameters were described using log-uniform prior distributions due to a paucity of data. Conversely, in the second (revised) parameterisation approach, parameters were characterised using unimodal log-triangular distributions. This represented a change from "uninformative" prior distributions, in which all values with a specified range were considered equally likely, to prior distributions in which a particular value (e.g., mode) was considered most likely.

Redefining the first parameter distribution was based on a combination of field and lab-based data. The spatial density of data was relatively higher in the area anticipated to feature greatest impact upon the four specified groundwater impact metrics. The aquitard characterisation presented here highlighted that initial prior estimates of K_V parameter ranges, as based on expert opinion and international literature, may often be too narrow (i.e., too small). This underscored the need, more broadly, to improve the characterisation of hydrogeologic parameters in order to reduce predictive uncertainty to levels that both the modelling community and water resource regulators are comfortable with. Building robust groundwater models is an iterative process in which uncertainties associated with parameters and model structure can be identified and quantified through sensitivity and uncertainty analyses, and reduced progressively through data collection.

Despite the wider ranges of aquitard K_V values used in the second parameterisation approach, comparisons between prediction uncertainty analyses based on the first (initial) data set and second (revised) parameterisations revealed that the groundwater impact metrics assessed here were only minimally affected by improving the characterisation of aquitard vertical hydraulic conductivity. Specifically, the analysis yielded the following results, based on 300 realisations per analysis (values are given for the initial parameterisation first, then for the revised parameterisation):

- **Magnitude of maximum drawdown**: the median value increased from 0.8 m to 1.2 m, the 95th percentile increased from 7.2 m to 7.9 m;
- **Timing of maximum drawdown**: the median value increased from 154 years to 155 years, while the 95th percentiles were identical (i.e., both 160 years);
- **Spatial extent of drawdown propagation**: the median value remained unchanged at zero model cells, and the 95th percentile increased from 4 490 cells to 5 350 cells.
- Maximum vertical flux: the median value decreased from 737 m³/d to 636 m³/d, and the 95th percentile decreased from 1 310 m³/d to 1 260 m³/d.

As demonstrated in the report, the impact of improving the hydraulic conductivity characterisation on the groundwater impact metrics was most evident when the conceptualisation of the groundwater model was updated such that the information contained within the updated data set was maximally exploited. This required honouring the spatial structure captured by the K_V data by first describing mathematically the spatial heterogeneity in K_V , according to an observed spatial correlation structure (or semi-variogram). This was then followed by the generation of multiple equally probable heterogeneous K_V parameter fields for groundwater modelling, thereby honouring the data at all the observation wells where upscaled K_V values had been determined. In doing so, the K_V domains were constrained or conditioned by observations at multiple cells and were considered to be a more realistic representation of spatial heterogeneity.

Based on this third parameterisation approach, a final uncertainty analysis was undertaken to identify the extent to which an updated model parameterisation could constrain calculated fluxes through aquitards induced by CSG extraction stresses, and the extent to which predictive uncertainty of the impacts was impacted. For this purpose, a set of 50 spatially heterogeneous K_V fields was generated as input to flow modelling. A sample size of 50 random fields was shown to be sufficient to obtain statistically robust estimates most of the groundwater impact metrics. Furthermore, the spatial dimension of the flow domain was shown to be sufficiently large in comparison with the correlation scales of the relevant hydrogeological properties of the formations of interest. The latter is a prerequisite to make meaningful inferences about statistical moments of the distribution of relevant variables. The effects of the second (revised) parameterisation using 300 model runs on the four groundwater impact metrics were compared with the effects based on the third (heterogeneous) parameterisation approach using 50 models runs (results for the second approach are given first, followed by results based on the third approach):

- Magnitude of maximum drawdown: the median value increased from 1.2 m to 3.4 m, while the 95th percentile decreased from 7.9 m to 6.1 m;
- **Timing of maximum drawdown**: the median value decreased from 155 years to 35 years, and the 95th percentile decreased from 160 years to 81 years;
- **Spatial extent of drawdown propagation**: the median value increased from zero cells to 63 cells, while the 95th percentile decreased from 5 350 cells to 551 cells;
- **Maximum vertical flux**: the median value increased from 636 m³/d to 751 m³/d, and the 95th percentile decreased from 1 260 m³/d to 1 074 m³/d.

These results were significant as they demonstrated two key points: first of all, median predictions of the magnitude and spatial extent of maximum drawdown increased slightly when the first parameterisation was replaced by the second. The median maximum drawdown prediction increased further (i.e., from 1.2 m to 3.4 m) when the third (heterogeneous) parameterisation was used, while the median timing of maximum drawdown prediction decreased from 155 years to 35 years. The latter was due to the skewed distribution of predictions generated using the second parameterisation, which was attributed to approximately 30% of the model runs not achieving maximum drawdown (i.e., the time to maximum was > 160 years). Unlike models generated for the two homogeneous parameterisation approaches, all model r uns for the third (heterogeneous) parameterisation achieved maximum drawdown within the total model run time, yielding the smaller median prediction value of 35 years.

Overall, results showed that for the model conditions and data sets used here, the median prediction values were not affected considerably (i.e., are fairly robust) by improved model

parameterisations informed by measurements (i.e., first and second parameterisations), unless a more detailed parameterisation is used (i.e., the third heterogeneous parameterisation).

Results further showed that calculated extreme drawdowns (i.e., 95th percentiles of magnitude and spatial extent of drawdown) are much smaller for the third (heterogeneous) parameterisation. The latter is consistent with maps of the spatial distribution of calculated drawdowns, illustrating that the heterogeneous parameterisation is much better constrained resulting in a much smaller range of extreme drawdowns. The extreme values (e.g., 95th percentiles), however, are materially affected (i.e., smaller) by using the heterogeneous K_V model. The better constrained K_V model is a result from using the conditioned Sequential Gaussian Simulation, which honours the observed K_V data. It is unlikely that the smaller number of model runs, 50 for the heterogeneous models versus 300 for the homogeneous models, is the main reason for the smaller range of drawdowns. The latter conclusion is based on robustness tests with the heterogeneous models, and on a demonstration that the scale of the model is sufficiently large compared to the scale of the heterogeneity in aquitard K_V .

The 95th percentiles for all other impact metrics show the same trend as for the drawdown: using the heterogeneous model decreases the timing of maximum drawdown, and decreases the impact due to the smaller spatial extent and smaller maximum flux. As mentioned above, this difference is unlikely due to the use of a smaller number of models runs (50 compared to 300). Instead, it is the result of using constrained heterogeneous parameter fields that generated models in which extreme K_V values would have less significance on groundwater flow given their small spatial footprint compared to homogeneous models where an extreme K_V would have a much larger impact on flow.

In a final step, a bootstrapping resampling method was implemented as a means to determine how robust the statistical moments (50^{th} , 75^{th} , and 90^{th} percentile) of groundwater impact metrics were. The analysis highlighted how the size of upscaled aquitard K_V samples affected the robustness of the summary statistics.

1 Introduction

1.1 Terms of reference

The Department of the Environment and Energy, through its Office of Water Science, requires hydrology research to better include faults and aquitards in Australian regional groundwater models to improve assessment of impacts of CSG extraction and coal mining.

This research addresses research priorities that the IESC has identified: "to increase the scientific evidence that underpins decisions about coal seam gas and large coal mining development, enabling decisions to be based on the most rigorous science available."

The research theme is expected to:

- 1. assist better decision-making, regulation, natural resource management and industry practice
- 2. build knowledge about the highest risks to freshwater resources, land and ecosystems
- 3. help provide data and knowledge that can support the Bioregional Assessments in priority areas.

The current focus is on identifying and assessing the risks associated with deep groundwater extraction and depressurisation. Recent research, discussion with the IESC and consultation with industry stakeholders identified the need for a project to specifically address the following three issues:

- 1. <u>Component 1 (Aquitards)</u>: *Improving understanding of vertical hydraulic conductivity in aquitards to examine the risk of depressurisation at a range of scales (this report).*
- 2. <u>Component 2 (Faults)</u>: Understanding the influence of faults on groundwater flow in Australian sedimentary basins, and the risk of faults propagating depressurisation to linked aquifers and surface environments (discussed elsewhere).
- 3. <u>Component 3 (Modelling)</u>: Better conceptualisation, representation and parameterisation of aquitards and faults in regional groundwater models to reduce uncertainty in regional groundwater flow and pressure simulation (aquitards are discussed in this report).

The current report discusses component 1 and 3 of the Terms of Reference, that is, to improve understanding of the vertical hydraulic conductivity of aquitards by better conceptualisation, parameterisation and representation of aquitards in regional groundwater models. This report builds on findings documented in three previous reports that were all part of this project:

 An overview of approaches to simulating the hydrological influence of aquitards and faults in regional groundwater models, and a summary of the literature relating to regional scale groundwater modelling approaches (Turnadge et al., 2018a). The overview provides a framework that can be used to guide research into appropriate methodologies and procedures for aquitard and fault zone representation in regional groundwater models. The report highlights the highly simplified representation of local scale processes such as dual phase flow and geomechanical deformation and regional scale considerations in the majority of groundwater models used in ten Australian CSG impact studies. Simplifications involved adopting spatially uniform values of hydraulic conductivity and storativity for aquitards, thus neglecting the spatial heterogeneity.

- 2. Based on a workflow to combine existing geophysical wireline logs available from coal seam gas exploration wells with laboratory measurements of porosity and permeability, continuous porosity and permeability profiles had been generated for four key aquitards (Purlawaugh Formation, Napperby Formation, Watermark Formation and Porcupine Formation) within the Gunnedah and Surat basins of New South Wales and generated 97 profiles of continuous permeability (or vertical hydraulic conductivity, *K*_v), with depths below surface of aquitards tops ranging from about 250 m to 1300 m. These high-resolution *K*_v profiles were upscaled into equivalent *K*_v values representative for large units commensurate with the typical size (i.e. vertical interval) of the numerical grid of the regional scale groundwater flow model (Turnadge et al., 2018b).
- 3. Developed and implemented a novel approach using the environmental tracer helium (⁴He) to derive the formation-scale hydraulic conductivity of key aquitards in the Gunnedah and Surat basins of New South Wales. By modelling the formation-scale transport, production and partitioning of helium in the aquitard sequence, a very slow formation-scale vertical fluid velocity on the order of 0.002–0.02 mm/year (about $10^{-13} 10^{-12}$ m/s) was derived (Smith et al., 2018).

1.2 Rationale

Historically, hydrogeology has focused on the characterisation of aquifers for water supplies. In comparison, the study of aquitards and their hydraulic properties has been relatively limited (Mazurek et al., 2011; Yu et al., 2013). The recent development of unconventional energy extraction industries (e.g., coal seam gas, shale gas and tight gas) has provided a new impetus for the study of aquitards. In the specific case of coal seam gas extraction, aquitards can restrict the vertical propagation of depressurisation from an extraction target unit (i.e. a coal seam) to a developed aquifer. Coal seam gas extraction development proposals typically involve predictive modelling of potential impacts on a groundwater flow system (e.g., OGIA, 2016). The representation of aquitards and their hydraulic properties in such models is typically limited (Turnadge et al., 2018a). The values assigned to aquitard hydraulic properties are often derived from prior modelling studies or textbooks, rather than from field or laboratory observations of specific aquitard units. The spatial variability of aquitard properties is also often omitted in favour of uniform values, typically due to the paucity of aquitard property data.

Based on a comparison of ten regional groundwater models, Turnadge et al. (2018a) reported that the OGIA model (OGIA, 2016) is the most advanced in many aspects. It is a typical example of a model having gone through several iterations with increased scientific underpinning, credibility, and accuracy. In the current study the starting point was to consider the least developed models, which represent the majority of cases, and then develop a generic workflow that in a stepwise manner increases the scientific underpinning, credibility, and accuracy. Once field data from depressurisation or other state variables becomes available, it can and should be used to further

constrain the models. Until then, the current approach provides a cost-effective, practical, repeatable and transparent approach to materially improve the existing models.

The uncertainty around aquitard hydrogeological properties and groundwater flow modelling based on such properties is typically large. Two broad categories of uncertainty exist in groundwater flow and transport modelling: aleatory and epistemic uncertainty (Helton et al., 2008; Ross et al., 2009; Swiler et al., 2009). Epistemic uncertainty refers to a lack of knowledge about the appropriate value to use for a quantity; this uncertainty can be reduced through increased understanding or collecting more data. Aleatory uncertainty is characterised by inherent randomness of a system that cannot be reduced by further data collection.

Epistemic uncertainty has two components, i.e. parametric and structural (model) uncertainty (Srinivasan et al., 2007). Parametric uncertainty, for example, reflects our partial knowledge about the appropriate value to use for the spatially averaged hydraulic conductivity *K* in groundwater flow analysis; the spatially-averaged *K* has, by definition, a single value but this single "effective" value can never be known with certainty. Neuman and Di Federico (1998) defined effective parameters as those parameters that are used in ensemble-averaged Darcy equations (e.g., effective hydraulic conductivity relating the ensemble average flux to the ensemble mean gradient). More importantly, as discussed by Turnadge et al. (2018a), the use of such effective *K* values may be justified to some extent by the regional scale of model applications and where there are insufficient data to support representations of spatial variability. Unfortunately, the rigorous basis for deriving and using effective *K* values such as using an appropriate averaging method or using Darcy's Law-based approaches is currently missing in many regional scale groundwater flow applications (Turnadge et al., 2018).

Parameter uncertainties are strictly epistemic because the uncertainty in the estimation decreases and may asymptotically vanish with increasing quantity and quality of the available observational data (direct measurements of parameters or measurements of state variables like heads from which parameters can be derived) (Der Kiureghian and Ditlevsen, 2007). Model (structural, conceptual) uncertainty in groundwater modelling shows itself on a multiplicity of scales, from pore scale to regional scale. In groundwater modelling different conceptual models are typically based on different geological interpretations (Højberg and Refsgaard, 2005; Rojas et al., 2010).

Several studies have recognized that geological structural uncertainty often is the most important source of uncertainty (Bredehoeft, 2005; Højberg and Refsgaard, 2005; Refsgaard et al., 2012). The most commonly used approach to assess uncertainty of model predictions due to conceptual geological uncertainty is to run multiple geological models in a scenario modelling or multimodel approach (Neuman and Wierenga, 2003; Rojas et al., 2010; Troldborg et al., 2007). Højberg and Refsgaard (2005) analysed the importance of parameter uncertainty relative to conceptual geological uncertainty by constructing three alternative groundwater models on the basis of three different geological interpretations for their study area in Denmark. Inverse model calibrations against groundwater heads and streamflows revealed a similar performance by the three models. A Monte Carlo based parameter uncertainty analysis showed that the model parameter uncertainty was the dominating source of uncertainty for prediction of groundwater heads throughout the model area. However, the results illustrated that the conceptual model uncertainty became relatively more important for predictions of groundwater recharge and even more important for prediction of chemical concentrations in abstraction wells. Højberg and

Refsgaard (2005) concluded that conceptual geological uncertainty will be more dominating than parameter uncertainty, the more the model predictions are extrapolations from the basis of model calibration (e.g. uncalibrated chemical velocities and concentrations). Because the relative importance of conceptual geological uncertainty will be region specific and will be depending on which system component needs interrogating to answer a specific management question (e.g., drawdown in a shallow bore or degree of depressurisation in a deep coal seam gas reservoir), it is recommended to explore in a step-wise manner its likely contribution to the overall uncertainty. A qualitative assessment commensurate with the generic methodology discussed in Section 1.4 would be a possible first step, followed by a more quantitative analysis where justified.

The uncertainty associated with geologic formations can be summarised into large-scale stratigraphic heterogeneity and smaller-scale heterogeneity within stratigraphic layers (facies distribution). Both types of geologic uncertainty will result in parametric uncertainty about flow and transport parameters. One approach to address geologic uncertainty is to adopt stratigraphic modelling and generate multiple geologic models. Stratigraphic forward modelling (SFM) was used by Ravenstein et al. (2015) to generate and characterise a static reservoir model using limited well data to assess geological storage of CO₂ in the Surat Basin, Queensland. SFM involves numerical simulation of the depositional processes to predict reservoir properties at appropriate scales (i.e., where the course and fine grained sediments are likely to be deposited), away from wells and below seismic resolution. This information can be used to get an independent view, based on sedimentary processes rather than stratigraphic correlation, on the risk of lateral continuity of reservoir or sealing geobodies. Ravenstein et al. (2015) developed a basin-scale model of formation thickness, sediment type and sediment heterogeneity with a 10 km grid spacing. Within this basin-scale model, a nested tenement-scale simulation was run with a 1 km grid spacing. Finally, a 3D permeability volume was calculated from the porosity model using a porositypermeability transform derived from an observed porosity/permeability relationship. Their regional-scale application still uses a relatively coarse spatial discretisation of 10 km, with the higher resolution model limited to relatively small areas (tenement scale) due to computational limitations.

A high-resolution (200 m horizontal cell size and 614 layers) tenement scale (17.7×19.5 km) geologic facies model was developed by Moore et al. (2015) in a comparison between a dualphase reservoir simulator and a single-phase groundwater model. Both models were used to examine depressurization and water desaturation processes in the vicinity of an extractive wellfield. While the focus of the Moore et al. (2015) study was mainly on testing upscaling methods for a highly layered coal – interburden system, the current study focuses on aquitard heterogeneity and uses existing water production curves as boundary condition.

A widely used alternative to explicit representation of geologic heterogeneity followed by coupling with a porosity-hydraulic conductivity relationship is to directly work with the heterogeneity in hydraulic conductivity *K*. Indeed, uncertainty about hydraulic properties (i.e., parametric uncertainty about hydraulic conductivity *K*) is often represented through multiple realisations of a 3D hydraulic conductivity model. Such realisations of aquitard properties represent realistic and statistically meaningful reconstructions of subsurface heterogeneity (Kolterman and Gorelick, 1996); they account for uncertainties about hydrostratigraphy, facies variability, etc., all of which affect hydraulic property variability. Realisations can be constrained by hard data including field measurements of hydraulic properties (i.e. *K*). For low-permeable aquitards formed in low-energy

depositional environments, the geologic heterogeneity is typically smaller compared to more permeable formations such as aquifers formed under high-energy environments. Relationships between the depositional environment and patterns of heterogeneity of clastic sedimentary rocks have been reported by Weber (1982). Deep marine shales, for instance, may be continuous for a hundred km (Richardson et al., 1978). Depositional flow regime features cause spatial variations in the average permeability (Kolterman and Gorelick, 1996). It is expected that for low-energy environments the hydraulic conductivity heterogeneity will be relatively small with a spatial structure dominated by large-scale heterogeneity (with typical hydraulic conductivity images which appear continuous). In relatively homogeneous formations with limited facies heterogeneities, fractures or large-scale conduits such as geological faults, the scale-dependency of K will also be limited. Examples of formations that are practically scale-invariant in K include marine clays (Yu et al., 2013), argillites (Distinguin and Lavanchy, 2007), and several types of sandstones (Schulze-Makuch et al., 1999). Furthermore, Moore et al. (2015) reported a two orders of magnitude range in K_V for the lower Sprinbok aquitard (sandstone/siltstone/shale) compared to three orders of magnitude for the Walloon Coal Measure interburden K_H and up to five orders of magnitude for the Walloon Coal Measure coal layers $K_{\rm H}$. For environments with large lateral continuity, layer-cake geometries are appropriate, while for more heterogeneous environments a labyrinth geometry is more appropriate (Kolterman and Gorelick, 1996).

1.3 Description

In the present study we sought to improve the representation of two aquitard sequences in an existing groundwater flow model that had previously been used to provide predictions of coal seam gas production impacts in the Gunnedah and Surat basins in NSW. The values assigned to these aquitard sequences had been derived from prior modelling studies and textbooks (CDM Smith, 2016). Turnadge et al. (2018a) noted that based on ten impact assessment studies involving coal seam gas production in Australian coal basins, groundwater flow models used to estimate potential impacts are often highly simplified. Specifically for aquitards, simplification typically involves neglecting the spatial heterogeneity of hydraulic conductivity and storativity by adopting spatially uniform values, often based on literature values, inverse modelling or limited field investigations. The existing groundwater flow model used in this project therefore represents a typical situation of many groundwater models currently in use by industry. The case study developed here is aimed at:

- 1. demonstrating how a materially better parameterisation of aquitard K_V can be achieved by using mainly existing data augmented with targeted additional measurements,
- 2. developing a workflow that can be readily implemented by industry on a sufficiently large spatial scale (i.e., regional scale) without putting an undue burden on computational resources,
- 3. applying methodologies that have been validated in similar geological settings and that are fit-for-purpose, and
- 4. developing a workflow that uses existing simulation software that is either free domain or can be obtained at a reasonable cost.

In selecting a fit-for-purpose methodology, care must be taken that the level of complexity is commensurate with the complexity of the system and the objective of the study (Neuman and

Wierenga, 2003); the general principle "as simple as possible, but not simpler" should be followed (Hill, 2006; Simmons and Hunt, 2012). Furthermore, increasing model complexity is only justified if the data base improves in quantity and quality. In other words, if the data base is limited and/or of poor quality, there seems little justification for selecting an elaborate model with numerous parameters. In this case, a simpler model with fewer parameters is then preferred, which still reflects adequately the underlying hydrogeologic structure of the system and the corresponding flow behaviour (Neuman and Wierenga, 2003). Unfortunately, for most modelling applications involving scenario analyses using simple models (not limited to this study), the validation data to test "whether the simpler model still reflects adequately the underlying hydrogeologic structure of the system and the corresponding flow behaviour" simply will never be available. What is key then is for the modeller to demonstrate that the modelling approach is fit for purpose; e.g., by demonstrating the impacts are conservative (i.e., they do not underestimate the impact) and statistically robust (Zuidema, 1994). The former can be shown by a comparison with a more complicated model (e.g., a model that has more processes, heterogeneous parameters rather than homogeneous, finer spatial discretisation, etc.). Finally, conservatism is not a static concept: conservative assumptions – imposed because of lack of data – can be replaced by more realistic ones when more data become available.

When relatively simple models are embedded in a stochastic framework, they may be offering a way to deal with complex heterogeneous systems (Hunt and Zheng, 1999). In a discussion on the practical use of simplicity in developing groundwater models, Hill (2006) argues that features or processes that often can most advantageously be represented as deterministic include the thicknesses and extents of hydrogeologic units, while features that often can most advantageously be represented as stochastic include heterogeneity in hydraulic conductivity, storage, or recharge.

1.4 Methodology

The methodology presented here aimed to (i) quantify the uncertainty in four key predictions that arises from uncertainty in aquitard *K* parameterisation; (ii) to identify the parameters contributing most to the predictive uncertainty; and (iii) to evaluate how to include measurements of aquitard hydraulic parameters in the uncertainty quantification.

Prior to the quantitative uncertainty analysis, a qualitative uncertainty analysis was undertaken to list key sources of uncertainty and discuss which source of uncertainty is relevance for being incorporated in the subsequent quantitative uncertainty analysis.

1.4.1 Qualitative uncertainty analysis

The qualitative uncertainty analysis is the first step out a series of steps that are part of the sensitivity and uncertainty analysis workflow adopted in this study (Figure 1-1). Following an initial qualitative analysis of uncertainty factors, existing hydrogeological data was compiled and prior parameter distributions derived followed by model stress testing. The latter was done to ascertain the groundwater flow model would run satisfactorily (i.e., test the robustness of numerical model convergence) for the imposed parameter range, which would typically include parameters that are considerably different from the calibration set. The next step then involved running a global sensitivity analysis with a large number of uncertain factors or parameters (a total of 30

hydrogeologic parameters for ten hydrostratigraphic units was considered here). Only the most sensitive parameters were then included in a stochastic quantitative uncertainty analysis, considering four groundwater impact metrics. After the first iteration of the sensitivity and uncertainty analysis, an improved characterisation was undertaken for one of the most sensitive and uncertain model parameters. Based on an updated prior parameter distribution, a second iteration of the uncertainty analysis was undertaken which produced the final statistics for the four groundwater impact metrics.



Figure 1-1 Sensitivity and uncertainty analysis flowchart adopted in this study.

The qualitative uncertainty analysis considered a broad range of factors (both conceptual and parametric) that could contribute to the uncertainty of predictions produced by numerical models of CSG extraction impacts. Potential sources of uncertainty include, but are not limited to (a more detailed discussion follows from Section 1.4.1.1 to Section 1.4.1.6):

- Aquifer hydraulic conductivity (K) and storativity (S),
- Aquitard hydraulic conductivity and storativity (S),
- Changes in rates of non-CSG groundwater extraction, recharge (e.g., due to climate and land use change over time, particularly over decadal timescales),
- Variations in the geometry (i.e., extent and/or thickness) of geological units,
- The translation between geological and hydrostratigraphic units; for example, the aggregation of units in a simple model, or the separation of units in a complex model;
- Changes in hydraulic properties of coal layers over time due to geomechanical deformation resulting from CSG extraction;
- The accuracy of predictions of CSG water extraction rates, which have historically been fraught with difficulty (Moore et al., 2015).

Some high-level generic insights in complex coupled systems such as coal seam-aquitard-aquifer systems may be obtained by considering simplified models. For example, Cook et al. (2016) show that the flux q(t) through the bottom of an aquifer due to depressurisation of a deeper gas-

bearing formation, separated by an aquitard from the aquifer, can be approximated analytically as (eq. 14 in Cook et al., 2016):

$$q(t) = 2\Delta H \sqrt{\frac{KS}{\pi t}} \exp\left(\frac{-L^2 S}{4Kt}\right)$$
(1)

where ΔH [L] is the change in groundwater head in the gas-bearing formation due to coal seam gas depressurisation, K [L/T] is the vertical hydraulic conductivity of the aquitard, S [–] is the storage of the aquitard, L [L] is the thickness of the aquitard and t [T] is the time since coal seam gas depressurisation commenced.

From equation (1) it is clear that the flux is proportional to the change in groundwater head in the gas-bearing formation, i.e. a doubling of ΔH will result in a doubling of the flux through the bottom of the aquifer. As equation (1) also shows, the effect of the aquitard properties on the flux (conductivity, storage and thickness) is non-linear, and therefore detailed analyses are required to evaluate for a specific geology and specific flow conditions and for specific management questions, how sensitive such a coupled system is to the aquitard properties.

In the subsequent sections the different sources of uncertainty are briefly discussed. Several of the uncertainty factors will not be taken into consideration in the current study; therefore, the current study does not present a comprehensive uncertainty analysis. A fully comprehensive assessment of model predictive uncertainty would simultaneously assess the sensitivity of predictions to a range of model parameters, boundary conditions and initial conditions. Such analysis is carried out as part of the Bioregional Assessments Programme in this region (Janardhanan et al., 2018).

1.4.1.1 Uncertainty due to aquifer hydraulic conductivity and storativity

As the groundwater impact metrics are calculated for the key aquifer of this study, the parameters that determine propagation of drawdown in this aquifer will need to be considered. Therefore, all key aquifers and their hydrogeologic parameters *K* and *S* were included in the initial sensitivity analysis (see discussion in Section 3.2).

1.4.1.2 Uncertainty due to aquitard hydraulic conductivity and storativity

As shown through Equation (1), the hydraulic properties of an aquitard constitute important parameters in determining the water flux *q*. The non-linear nature of the analytic solution warrants exploring its effect through global sensitivity analysis techniques based on stochastic analysis of a regional groundwater model. All aquitards considered in this study had not been the subject of detailed investigations, making their hydrogeologic parameters highly uncertain (see discussion in Section 4.1). The uncertainties associated with aquitard hydraulic properties were represented through multiple realisations of a 3-D hydraulic conductivity model. Realisations were constrained by hard data including field measurements of hydraulic properties (i.e. *K*).

1.4.1.3 Uncertainty due to groundwater Recharge

Groundwater recharge is expected to affect the groundwater balance of aquifers. The calibrated groundwater that was used in this study was not suitable to have recharge as part of a sensitivity and uncertainty analysis because of the way the recharge was calibrated. The standard calibration

approach is to match the simulated hydraulic head in a groundwater model to the observed water table elevation by iteratively updating the hydrogeological properties. For the Namoi Alluvium, CDM Smith (2016) used an alternative approach which does not require the hydrogeological properties of the model to be changed; recharge outside the Namoi Alluvium was based on average rates of regional rainfall and outcrop/subcrop geology (CDM Smith 2016). The approach for the Namoi Alluvium involved optimising the recharge fluxes at the water table until a good match between observed and calculated heads was obtained. The approach also combined rainfall and flood recharge, irrigation returns, groundwater pumping and evapotranspiration into a single estimate of net flux, and therefore these processes are not individually represented in the modelling. Note that this approach was only applied within the Namoi Alluvium. For these reasons, groundwater recharge was not included in the uncertainty analysis.

1.4.1.4 Geologic uncertainty

For the purposes of the present study, it was assumed that the conceptualisation and parameterisation of the Gunnedah and Surat basins groundwater flow system was adequate, with the exception of the parameterisation of aquitard hydraulic properties. The geological model was constructed using Leapfrog Hydro (v. 1.7)¹ and was based on a combination of geological datasets, including drilling logs, stratigraphic surfaces, outcrop geology, and ground surface. The twentynine stratigraphic units present within the assessment area were represented by 13 model layers (CDM Smith, 2016). Because the focus of this study was on aquitards, conceptual geological uncertainty was deemed less critical compared to parametric uncertainty. There are several reasons that support this approach for aquitards: (1) aquitards are geologically more homogeneous than aquifers or coal layers, especially when formed under low-energy depositional environments, (2) the larger-scale heterogeneity can be captured by means of sufficient hydraulic conductivity measurements, and (3) low-permeable aquitards are naturally less sensitive to geological or hydraulic heterogeneity (provided the hydraulic conductivity is low enough). As mentioned above, the relative importance of conceptual geological uncertainty is site-specific and depending on the specific management question being addressed. It is therefore recommended to explore its likely contribution to the overall uncertainty first in a qualitative manner, before undertaking a quantitative study. Indeed, any uncertainty analysis should try to be as comprehensive as possible, but this does not mean that all possible parameters, boundary conditions, etc. need to be included in a quantitative analysis.

1.4.1.5 Uncertainty due to dynamic behaviour of coal layer hydraulic properties

Permeability in coal is controlled by the magnitude of net stress in the hydrocarbon reservoir. This can vary across the field, in different coal seams, and also change over time with production. Production influences permeability in two opposing ways. First, a decrease in permeability may occur due to cleat compaction. Secondly, an increase in permeability may happen due to coal matrix shrinkage as gas desorbs. If the matrix shrinkage is large enough, then this could counteract

¹ http://www.leapfroghydro.com/hydro/

any decrease in permeability from the loss of pore pressure and cleat closure (Moore, 2012). The dynamic behaviour of coal layer permeability is not normally accounted for in reservoir models. If deemed important, its uncertainty would need to be part of the calculation of the water production rates.

1.4.1.6 Uncertainty due to water production rates

Potentially one of the largest sources of uncertainty to estimate impacts of coal seam gas depressurisation in shallow aquifers are the water production rates. Detailed reservoir simulation by coal seam gas proponents produces such drawdown estimates as part of their assessment of the water production rates required to sufficiently depressurise for methane to desorb. Most proponents recognize that there is considerable uncertainty in the water production rates (and hence drawdowns), not only due to uncertainty in the geological characterisation of the coal formations but also due to uncertainty in the planned production schedule, and typically provide an ensemble of water production rates based on stochastic analysis (CDM Smith, 2016).

The water production rates from reservoir simulation modelling are often used in regional groundwater models as either a direct boundary condition (prescribed flux) or as a constraint on the water production rate if depressurisation is implemented as a specified drawdown boundary condition. The work of Herckenrath et al. (2015) and Moore and Doherty (2015) point out some of the major issues related to simulating coal seam gas extraction in regional scale groundwater models, such as the single phase versus dual phase flow and the upscaling of the hydraulic properties of the gas bearing formations. They show that many of the simplifying assumptions in regional groundwater flow modelling, such as uniform properties and single phase flow, lead to overestimates of drawdown in the gas bearing formations.

Available water production rates for the current case study area demonstrate a relatively small uncertainty, based on the available model simulations for the deepest of two coal formations (CDM Smith, 2016). Indeed, the water production volumes for the Low Case (P90 exceedance probability), Base Case (P50 exceedance probability) and High Case (P10 exceedance probability) realisations are 72.3 GL, 83.8 GL and 98.3 GL, respectively. The difference between the Low and High Cases is 26 GL (31% variation of the Base Case). The High Case is 14.5 GL higher than the Base Case, or 17%. The Base Case was used in the current study. There were no stochastic variations available of the estimates of water production from the shallower coal formations. Although not generally recommended, this study did not include water production as an uncertain factor due to lack of data for the shallowest coal layers and due to a relatively small uncertainty for the deepest coal layers. As data becomes available, the initially calculated water production rates can be updated and the uncertainty analysis repeated if observed data shows a much larger variability than the initial estimate.

1.4.2 Quantitative uncertainty analysis

Global sensitivity analysis methods were used to characterise the uncertainty of four predictions (i.e., groundwater impact metrics) of interest. The initial prior statistical distributions of 30 model parameters were defined using log-uniform distributions (Figure 1-2). The use of uniform distributions (or log-uniform distributions if values span several orders of magnitude) is

recommended when very little is known about the true parameter distribution due to a lack of data (i.e., when one is unable to decide which values within a given range are more likely than others) (Mallants et al., 2003).



Figure 1-2. Flowchart of the workflow employed in this study to assess the effects of improved aquitard characterisation (red triangular distributions) on parameter sensitivity (not shown) and prediction uncertainty (one hypothetical flow metric shown).

In order to provide a more rigorous basis for the representation of these aquitard sequences, laboratory porosity-permeability measurements were undertaken and relationships between

these two parameters were established for key hydrogeological formations. Borehole wireline log data were then used to generate continuous porosity profiles, which were subsequently transformed into permeability profiles based on the derived porosity-permeability relationships. High-resolution permeability profiles were upscaled to aquitard formation scale prior to inclusion in the groundwater flow model (Turnadge et al., 2018a).

The ranges of these distributions were specified arbitrarily using CDM Smith (2016) model parameter values ± one order of magnitude. These ranges are consistent with the variability ascribed to aquitard hydraulic properties by modelling studies prior to that of CDM Smith (2016). The combinations of parameter values sampled from these distributions represented groundwater flow models that were assumed to be equally likely (i.e., with likelihood functions equal to unity). Furthermore, the models generated from these prior distributions were not constrained by observations of system states, such as hydraulic heads or groundwater discharge fluxes; e.g., as often undertaken in a traditional calibration-based approach.

Following an initial estimation of prediction uncertainties based on initial model parameter values (Figure 1-2), the prior parameter distributions for two aquitard sequences were updated on the basis of upscaled aquitard vertical hydraulic conductivity values using a combination of core measurements and geophysical wireline logging data from around 100 deep exploration wells (Turnadge et al., 2018b). A second global sensitivity analysis was then performed using these updated prior parameter distributions (Figure 1-2). The uncertainties of the four predictions of interest were again estimated and subsequently compared to initial results.

An important source of uncertainty about predicted impacts originates from imperfect conceptual models (Gupta et al., 2012; Rojas et al., 2010). To address conceptual model uncertainty we include, as part of the broader uncertainty analysis, a more complex conceptual model that differs in its representation of hydraulic conductivity heterogeneity within aquitard formations.

It is of further importance to note that the stochastic approach adhered to in this project is consistent with the Bayesian paradigm, in which prior beliefs and estimates of parameters and conceptualisation are iteratively updated when new data and information becomes available. In this study no effort is invested in constraining parameters by observations of state variables, a process referred to variously as calibration, history matching or inverse modelling (Carrera et al., 2005). The calibration process assumes that a parameter set that minimises the mismatch between observed and simulated heads will result in more accurate predictions. White et al. (2014) demonstrated that, in many instances, this assumption is not valid and that calibration can even lead to bias in predictions. Another example where a well-calibrated model provided inaccurate predictions was reported by Moore and Doherty (2006). They demonstrated that transport predictions made by a model that calibrated perfectly to ground water elevation data can be 100% wrong as a consequence of the simplifications required to achieve a unique calibration. Furthermore, the stochastic uncertainty quantification undertaken in this study involves a future scenario of how a groundwater system may evolve under a given stress (i.e., depressurisation). As the stress will only eventuate in the future, there are currently no head or flux data available for model calibration. It should also be noted that in unconstrained Bayesian approaches, such as presented here, the ensemble of models generated will typically include a model featuring parameters that would otherwise be estimated through calibration to
observations. Therefore, the range of prediction uncertainty explored through such Bayesian approaches can be considered comparatively more robust.

Furthermore, calibration-based approaches also assume that the information content of available observations is suitable and relevant to the parameters estimated through model inversion. In the context of propagation of depressurisation due to CSG extraction, the theoretical analysis of Cook et al. (2016) indicated that the current set of predevelopment groundwater level and flux observations in the Gunnedah and Surat basins are highly unlikely to yield information on the salient hydrogeological parameters as they are mostly controlled by external driving forces such as pumping, groundwater surface water interactions and recharge (Giambastani et al., 2009).

1.5 Workflow

CDM Smith (2016) developed a transient model of the Gunnedah-Surat basins groundwater flow system as part of Environmental Impact Assessment requirements for the Santos Narrabri Gas Project, to be located near Narrabri, NSW. The model was developed using the MODFLOW-SURFACT simulator (HydroGeoLogic, 1996), which is a proprietary version of the widely used MODFLOW groundwater flow simulator (Harbaugh, 2005). This deterministic model featured spatially uniform values for each hydrostratigraphic unit represented. Two aquitard sequences were represented in the model, both of which occur between the primary CSG target unit (the Maules Creek Formation coal seams) and the key confined aquifer in the region (the Pilliga Sandstone aquifer).

For the purposes of the present study, four simulation metrics were defined for the assessment of prediction uncertainty, based on groundwater flow predictions relating specifically to the Pilliga Sandstone aquifer:

- (1) the maximum hydraulic head reduction in the Pilliga Sandstone aquifer with respect to preproduction conditions (i.e., drawdown);
- (2) the total time elapsed after the cessation of CSG extraction at which maximum drawdown conditions occur in the Pilliga Sandstone aquifer;
- (3) the number of model cells in the Pilliga Sandstone aquifer with drawdown greater than or equal to two metres; and,
- (4) the maximum vertical flux across the low hydraulic conductivity formations that separate the Pilliga Sandstone aquifer from underlying CSG target formations.

With respect to prediction (3), a nominal drawdown value of two metres was selected, which is consistent with trigger-level thresholds specified by the NSW Aquifer Interference Policy (NSW DPI OW, 2012). In many cases, when drawdowns induced by groundwater extraction exceed two metres, "make-good" provisions are enacted. The sensitivities of these four predictions to 30 model parameters were tested. These parameters included the horizontal and vertical hydraulic conductivities (K_H , K_V) and specific storage coefficients (S_s) of the ten hydrostratigraphic units represented in the model (Table 2-1). For each parameter, the value specified by CDM Smith (2016) was assumed to be the mean of a log-uniform distribution with upper and lower bounds specified as the mean value \pm one order of magnitude. Parameter values were sampled 300 times from these prior distributions and used as model inputs. The resulting set of 300 model outputs were

used to characterise the initial uncertainty (i.e., unconstrained by data) in the four metrics of interest. Statistical tests were undertaken to confirm that this sample size was sufficiently large to capture the full range of parameter variability (see Figure 3-2 and Figure 5-2). As mentioned before, it should be noted that the approach presented differs from calibration-based approaches. The approach involved quantification of the uncertainty of each of the four predictions was assessed by sampling from all possible parameter combinations, each of which were considered equally likely.

As described by Turnadge et al. (2018b), the laboratory testing of aquitard core samples, in conjunction with the processing of wireline log-derived porosity profiles from exploration drilling, were used to generate revised prior K_V distributions for the two aquitard sequences represented in the CDM Smith (2016) groundwater flow model. These revised prior distributions were used as a basis for a second set of 300 model runs that provided revised estimates of the uncertainty of the four predictions of interest (Section 5). As will be shown later (Section 4), at least the K_V of the first aquitard sequence was identified in the sensitivity analysis to be a key parameter to which three out of four groundwater impact metrics were sensitive.

In a subsequent step of the uncertainty analysis, upscaled values were used to estimate the spatial structure of aquitard K_V for both aquitard sequence. This was achieved by estimating semi-variogram models which were subsequently used to generate 50 stochastic random fields. An upper limit of 50 realisations was selected due to time limitations, as each model run typically required between 20 and 60 minutes to complete. All realisations were considered to be equally-likely spatial distributions of heterogeneous aquitard K_V values. Predictions generated using this alternative conceptualisation were compared to those generated using the spatially uniform parameterisation used in steps 1 and 2.

The adopted workflow involves high-resolution computational Monte Carlo simulations that produce a large number (i.e., 300 for the spatially uniform model and 50 for the heterogeneous model) of equally likely results. These non-unique results are summarized in terms of statistically averaged quantities and sample probability distributions. A further advantage of the 50 stochastic random fields is that their results honour measured values of aquitard properties (K_V), i.e. they are said to be conditioned on observed data. An important improvement of conditioned over unconditioned simulations is that one obtains (among others) conditional mean flow variables that constitute optimum unbiased predictors of these unknown random quantities (Neuman and Wierenga, 2003).

In a final step, the robustness of the summary statistics (the percentiles) of the predictions are evaluated in a simplified data worth analysis using a bootstrap resampling methodology. The bootstrap resampling approach is a non-parametric method of calculating prediction confidence intervals and parameter uncertainty. If the uncertainty is high, the summary statistic is prone to change due to omission or inclusion of only a few data points. The summary statistic can then not be considered robust and the data density is insufficient to characterise the parameter distribution according to an acceptable degree of uncertainty.

2 CDM Smith (2016) Gunnedah and Surat basins groundwater flow model

2.1 Geographic context

The Gunnedah geological basin is located approximately 230 km north west of Sydney and encompasses an area of approximately 15 000 km². The Gunnedah Basin underlies much of the Liverpool Plains area which includes the municipalities of Dubbo, Narrabri and Gunnedah. The study area is located west of the line connecting the towns of Narrabri and Gunnedah (Figure 2-1).



Figure 2-1 Spatial extent of Gunnedah geological basin (NSW) and numerical grid of groundwater flow model. The highlighted row (A-A') and column (B-B') was used to display the hydrostratigraphic cross-section.

2.2 Geology

The Permian-age Gunnedah geological basin represents the central portion of the Sydney-Gunnedah-Bowen Basin. A narrow foreland basin, it is located between the Thomson Orogen in the west and the New England Fold Belt in the east. The Gunnedah Basin is overlain by the Jurassic-Cretaceous age Surat Basin (a sub-unit of the larger Great Artesian Basin). Collectively, these two geological basins contain a layered sequence of marine, non-marine and volcanolithic sedimentary rocks up to 1 200 m thick (Table 2-1). Coarse-grained units include conglomerates (e.g., Napperby and Porcupine Formations) and sandstones (e.g., Mooga, Pilliga and Clare Sandstones). Fine-grained units include claystones (e.g., Benelabri and Leard Formations) and shales (e.g., Porcupine and Watermark Formations). In parts of the Gunnedah Basin area, sedimentary rocks are intruded by igneous units such as the Garrawilla, Liverpool Range and Warrumbungle Volcanics. The Gunnedah Basin also includes significant coal-bearing units such as the Maules Creek Formation and the Hoskissons Coal.

2.3 Hydrostratigraphy

The following description of the hydrostratigraphy of the Gunnedah and -Surat basins follows that proposed by CDM Smith (2016; Table 2-2). Basement units include the Werrie Basalt and Boggabri Volcanics and the Goonbri and Leard formations. For the purposes of groundwater flow simulation, these units are assumed to be mostly impermeable to flow and do not interact with overlying units. The Maules Creek Formation overlies basement units and contains coal beds which were the primary target for coal seam gas development, with the estimated water production shown in Figure 2-4. The model cross-sections in Figure 2-2 and Figure 2-3 are located along the highlighted row of the model grid (Figure 1-2).

The hydraulic conductivity of the coal beds (e.g., $K_H = 0.1 \text{ m/d}$) was assumed to be two orders of magnitude higher than the remainder of the Maules Creek Formation (e.g., $K_H = 0.001 \text{ m/d}$) and thereby consistent with that of an aquifer. The Maules Creek Formation is overlain by a thick sequence of middle to late Permian sedimentary rocks which collectively represent an aquitard unit. This is overlain by the Hoskissons Coal, which was the secondary target for coal seam gas development with a much smaller estimated water production (Figure 2-4). As with the Maules Creek coal beds, the hydraulic conductivity of the Hoskissons Coal unit (e.g., $K_H = 0.1 \text{ m/d}$) was assumed to be consistent with that of an aquifer. The Hoskissons Coal is overlain by a second thick sequence of late Permian to Jurassic sedimentary rocks which collectively represent an aquitard unit. This is overlain by a key confined aquifer, the Pilliga Sandstone, which is the main Great Artesian Basin aquifer in the Namoi subregion. Its outcrop in the central part of the Namoi subregion corresponds to the boundary of the Great Artesian Basin (CSIRO, 2012).

The Pilliga Sandstone aquifer is overlain by a number of Cretaceous age sedimentary rocks including the Mooga Sandstone and the Orallo and Bungil formations. The Liverpool Range Volcanics are also included in this hydrostratigraphic unit, which is considered to be an aquitard. At ground surface, the Namoi Alluvium and Gunnedah-Oxley Basin Formations are present in various parts of the study area. The predominant water supply development has occurred within the Namoi Alluvium, with smaller developments in the Gunnedah-Oxley Basin.



Figure 2-2 Approximately east-west cross-section (transect A-A' in Figure 2-1) with hydrostratigraphic units of the groundwater model for the case study area, Gunnedah and Surat basins (NSW).



Figure 2-3 Approximately north-south cross-section (transect B-B' in Figure 2-1) with hydrostratigraphic units of the groundwater model for the case study area, Gunnedah and Surat basins (NSW).

Table 2-1. Stratigraphy of the Gunnedah Basin (CDM Smith, 2016; Geoscience Australia, 2016).

Group	Stratigraphic unit	Lithology
N/A	Narrabri Formation	Clay, silt, sand
	Gunnedah Formation	Gravel, sand, clay
Liverpool Range Volcanics	Liverpool Range Volcanics	Basalt, dolerite, conglomerate, sandstone, shale, gravel, siltstone
Rolling Downs	Wallumbilla Formation	Mudstone, siltstone, sandstone, limestone
Blythesdale	Bungil Formation	Sandstone, siltstone, mudstone, coal
	Mooga Sandstone	Sandstone, siltstone, shale, mudstone, coal
,	Orallo Formation	Sandstone, siltstone, mudstone, coal
Injune Creek	Pilliga Sandstone	Sandstone, conglomerate, mudstone, siltstone, coal
	Purlawaugh Formation	Sandstone, siltstone, mudstone, coal
	Garrawilla Volcanics	Dolerite, basalt, trachyte, tuff, breccia
,	Deriah Formation	Sandstone, mudstone
	Napperby Formation	Sandstone, siltstone, conglomerate, shale
,	Digby Formation	Conglomerate, sandstone
Black Jack	Trinkey Formation	Claystone, siltstone, sandstone, tuff, claystone, coal
,	Wallala Formation	Conglomerate, sandstone, siltstone, claystone, coal
	Clare Sandstone	Sandstone, conglomerate, coal, claystone
,	Benelabri Formation	Claystone, siltstone, sandstone, coal
	Hoskissons Coal	Coal, sandstone, siltstone, claystone, tuff
,	Brigalow Formation	Sandstone, siltstone
	Arkarula Formation	Sandstone, siltstone
,	Pamboola Formation	Sandstone, siltstone, claystone, conglomerate, coal
Millie	Watermark Formation	Siltstone, claystone, sandstone
, 	Porcupine Formation	Conglomerate, sandstone, siltstone
Bellata	Maules Creek Formation	Claystone, sandstone, coal, siltstone, conglomerate
	Goonbri Formation	Siltstone, coal, claystone, sandstone
	Leard Formation	Sandstone, conglomerate, coal
N/A	Werrie Basalt	Basalt, tuff, coal
	Boggabri Volcanics	Rhyolite, dacite, tuff, shale, trachyte, andesite



Figure 2-4 Water production for Maules Creek Formation and Hoskissons Coal seam target formations as predicted by the CDM Smith (2016) groundwater flow model.

In summary, CDM Smith (2016) conceptualised the Gunnedah-Surat basins groundwater flow system as being composed of one unconfined aquifer, one confined aquifer, six distinct aquitards (including two coal interburden units), and two coal measures (Table 2-2). When representing these hydrostratigraphic units in a numerical groundwater flow model, CDM Smith (2016) used a single model layer to represent the unconfined Namoi Alluvium aquifer (i.e., layer 1). Each of the Pilliga Sandstone aquifer, Wallumbilla Formation aquitard, and Bungil Formation–Orallo Formation aquitard were distributed across six model layers (i.e., layers 1-6). The Pilliga Sandstone aquifer was confined in some parts of the model (i.e., was overlain by the Wallumbilla Formation or Bungil–Mooga–Orallo Formation aquitards). In other parts of the model it was unconfined (i.e., was overlain by the Namoi Alluvium, or was outcropping). Six model layers were used to represent the upper aquitard sequence (i.e., Purlawaugh Formation to Benelabri Formation; layers 7-12). Seven model layers were used to represent the lower aquitard sequence (i.e., Brigalow Formation to Porcupine Formation; layers 13-19). Both coal measures (Hoskissons Coal and Maules Creek Formation coal member) were each represented using a single model layer (i.e., layers 13 and 22). Maules Creek Formation interburden members, which overlie and underlie the Maules Creek Formation coal members, were represented using five model layers (i.e., layers 20-24). Consequently, it should be noted that many model layers contain multiple hydrostratigraphic units; for example, model layer 13 contains a discrete area of Hoskissons Coal within the lower aquitard sequence. Similarly, model layer 22 contains a discrete area of the Maules Creek Formation coal member within the Maules Creek Formation interburden member.

Table 2-2. Hydrostratigraphy of the Gunnedah-Surat basins as represented in the CDM Smith (2016) groundwater flow model.

Geological	Geological	Hydrostrati-	Groundwater	Groundwater
unit	model layer	graphic unit	model parameter ID	model layer(s)
Namoi Formation	1	aquifar	01	1
Gunnedah Formation		aquiter	01	1
Liverpool Range Volcanics	2	aquitard	02	1.6
Wallumbilla Formation		aquitaru	02	1-0
Bungil Formation				
Mooga Sandstone	3	aquitard	03	1–6
Orallo Formation	-			
Pilliga Sandstone	4	aquifer	04	1–6
Purlawaugh Formation	5			
Garrawilla Volcanics	6			
Deriah Formation	7	-		
Napperby Formation				
Digby Formation	8			
Trinkey Formation		aquitard	05	7–12
Wallala Formation				
Breeza Coal Member				
Clare Sandstone	9			
Howes Hill Coal Member				
Benelabri Formation	,			
Hoskissons Coal	10	coal	06	13
Brigalow Formation				
Arkarula Formation				
Melvilles Coal Member	, 11	aquitard	07	12 10
Pamboola Formation		aquitard	07	13-19
Watermark Formation	12	-		
Porcupine Formation				
Maules Creek Formation (upper)	·	interburden	08	20, 21
Maules Creek Formation (coal)	13	coal	09	22
Maules Creek Formation (lower)	,	interburden	10	22–24

In summary, the main aquitards of interest were the upper aquitard sequence (i.e., groundwater model layers 7-12) and the lower aquitard sequence (i.e., groundwater model layers 13-19). These aquitards were the focus of improved characterisation as well as the estimation of the effects of improved representation of vertical hydraulic conductivity.

2.4 Inflows and outflows

The majority of boundary condition fluxes applied to the model occurred on the top model boundary. These were specified separately for the Namoi Alluvium and for the remainder of the top model layer and are described as follows.

Within the Namoi Alluvium, the Namoi River was represented using Cauchy boundary conditions (i.e., RIV package), which were parameterised using river stage, base and conductance parameters. Within the remainder of the Namoi Alluvium, Neumann boundary conditions (i.e., RCH package) were applied in order to represent net recharge; i.e., gross recharge minus evapotranspiration (ET) minus extraction. For cells in which extraction and/or ET exceeded gross recharge, negative net fluxes of zero to -0.5 mm/d were specified. Elsewhere, positive net fluxes of zero to +0.5 mm/d were specified (CDM Smith, 2016; Figure 6-13).

For cells located outside the extent of the Namoi Alluvium in the top model layer, Neumann and Cauchy boundary conditions were specified simultaneously. Spatially variable gross recharge rates ranging from 0.5 mm/y to 16 mm/y were applied using Neumann boundary conditions (i.e., RCH package). Evapotranspiration fluxes were represented using Cauchy boundary conditions (i.e., the EVT package). This package calculates ET fluxes as a linear function of depth below ground surface (i.e., model top). Calculated ET rates varied from a maximum of 600 mm/y for a watertable at ground surface to zero for a watertable ≥ 5 m depth (CDM Smith, 2016; Figure 6-14).

Other model inflows and outflows occurred via lateral connections to other groundwater flow systems. This included connections between the Pilliga Sandstone aquifer and the Great Artesian Basin, as well as connections between the Namoi Alluvium aquifer represented in the model and the Lower Namoi Alluvium. All lateral inflows and outflows were represented using Dirichlet boundary conditions.

2.5 Representation of coal seam gas extraction

The extraction of fluid (i.e., water and methane gas) for both the primary coal target (Maules Creek Formation), and the secondary coal target (Hoskissons Coal) was represented in the model using Neumann boundary conditions (i.e., WEL Package). Time-varying extraction rates were assigned to relevant model cells in accordance with typical production dynamics (for details, see Figure 2-4 and CDM Smith, 2016). The water extraction rates used here represent the CDM Smith 'base case', which did not include groundwater extraction for mine dewatering. In practice, CSG water extraction rates are typically higher upon the commencement of production (in order to extract sufficient water to lower the pressure in coal seams, thereby allowing gas desorption and flow of gas to the well) and are reduced more or less exponentially over time (e.g., Moore 2005). It should be noted that past studies have highlighted that the prediction of extraction rates and durations required for CSG extraction is often fraught with difficulty (KCB, 2012; Moore et al., 2015), and that this can represent one of the most significant uncertainties in estimating the cumulative impacts of CSG production on adjacent aquifers. For reasons discussed in Section 1.4.1, this uncertainty is not the focus of this study.

2.6 Numerical solution scheme

The proprietary finite-difference code MODFLOW-SURFACT (HydroGeoLogic, 1996) was used to solve the groundwater flow equation. While this version of MODFLOW includes a number of additions and alterations (including adaptive time stepping for transient models and improved treatment of dewatered model cells), none of these features were used in the development of the CDM Smith (2016) model. Instead, the sole reason for the use of MODFLOW-SURFACT rather than public domain versions was the use of an alternative preconditioned conjugate gradient numerical solver (PCG4). Preliminary testing of the model found that numerical convergence could not be achieved when using either the PCG2 (Hill, 2003) or NWT (Niswonger et al., 2011) standard solvers.

2.7 Model modifications

Prior to undertaking sensitivity analyses, a number of modifications were made to the CDM Smith (2016) groundwater flow model. First of all, numerical model convergence criteria were adjusted. These modifications are described as follows. The hydraulic head closure criterion for the preconjugate gradient solver was increased in size from 0.001 m to 0.1 m, in order to minimise model run times. The number of outer solver iterations was also reduced from 2 000 to 100, also to reduce model run times. Model mass balance errors were checked to ensure that relaxation of the convergence criterion did not affect the accuracy of the numerical solution (in terms of the total groundwater balance). Mass balance errors for steady state and transient solutions were < 5 % and < 0.1 %, respectively.

The input files for the constant head, recharge and evapotranspiration packages were rewritten to reduce their size and thereby reduce model run times. The writing of hydraulic head outputs was deactivated, also in order to reduce model run times.

Finally, the temporal extent of the model was modified by including a single final stress period of 160 years in length, which was subdivided into 160 time steps. This period represented the response of the groundwater flow system upon the cessation of groundwater extraction for CSG production. The length of the stress period was consistent with similar simulations undertaken by the Bioregional Assessments Programme. Groundwater impact calculations for the Pilliga Sandstone aquifer for periods in excess of 100 – 150 years are considered uncertain due to likely changes in climatic boundary conditions that have not been accounted for here.

3 Initial prediction sensitivity and uncertainty analyses

3.1 Model predictions

The model simulated groundwater flow in ten hydrostratigraphic units of the Gunnedah-Surat basins. The sensitivities of four modelled predictions to three flow parameters per unit (i.e., 30 parameters in total) were examined. All predictions related to potential hydraulic impacts to the Pilliga Sandstone aquifer and were compared to steady-state model outputs, which represented pre-development conditions. Specifically, the four predictions were:

- (1) the maximum hydraulic head change (i.e. drawdown) across all model cells;
- (2) the time elapsed when maximum drawdown conditions occurred;
- (3) the spatial extent of drawdown greater than or equal to two metres (at any given time); and,
- (4) the maximum vertical flux across the base of the Pilliga Sandstone aquifer (at any active model cell).

As the model used to generate predictions was of numerical type, the spatial extent of the model was discretised using a number of cells. For this reason, prediction (3) was calculated discretely rather than by integrating a single continuous area. This prediction was calculated as the total number of cells with drawdown in excess of two metres, hereafter referred to as the number of drawdown cells ('NDD'). It should be noted that the model featured cells of various sizes (i.e., 1 km², 5 km² and 25 km²); therefore areas affected could not be calculated directly from the number of cells affected.

3.2 Model parameters

The model parameters to which prediction sensitivity was tested were the horizontal (K_H) and vertical (K_V) hydraulic conductivity and specific storage (S_S) of each of the ten hydrostratigraphic units, resulting in a total of 30 parameters. As the CDM Smith (2016) model was of deterministic type, it used single values for each parameter, which were specified based upon reviews of existing literature and models (and referred-to as "order-of-magnitude estimates", CDM Smith, 2016). In comparison, the sensitivity analysis performed for the present study required the definition of prior distributions for each of the model parameters tested. These were specified using log-uniform distributions with arbitrary ranges defined by the CDM Smith (2016) parameter value (for each of K_H , K_V and S_S) ± one order of magnitude.

3.3 Model stress testing

Prior to undertaking global sensitivity analyses, the CDM Smith (2016) model was subjected to a number of "stress tests" in order to test the robustness of numerical model convergence. Based

upon the prior distributions specified for model parameters, 11 parameter combinations were devised (Table 3-1). These combinations primarily explored minimum and maximum possible values, either for all parameters (i.e., n=2, 3) or for groups of parameters (i.e., n=4-9). The effects of increasing and decreasing hydraulic diffusivity (i.e., K_H/S_S) were also examined (i.e., n=10, 11).

Table 3-1. Results of parameter stress testing of the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model. Parameter ranges (for each of K_{H} , K_{V} and S_{S}) were specified as the CDM Smith (2016) parameter value ± two orders of magnitude.

n	Parameter set	Converged?
1	Base parameter set	yes
2	All parameters at minimum	no
3	All parameters at maximum	no
4	K _H parameters at minimum	no
5	K _H parameters at maximum	no
6	K_V parameters at minimum	no
7	K_V parameters at maximum	yes
8	S ₅ parameters at minimum	yes
9	S ₅ parameters at maximum	yes
10	K_H and K_V parameters at maximum, S_S parameters at minimum (i.e. high diffusivity)	no
11	K_H and K_V parameters at minimum, S_S parameters at maximum (i.e. low diffusivity)	yes

Seven of the 13 parameter combinations tested did not achieve convergence. These included: when all parameters were set to their minimum or maximum values; when all K_H parameters were set to their minimum or maximum values; when all K_V parameters were set to their minimum value; and when all hydraulic diffusivity values were set to their maximum value. Based on these results, it was decided to reduce the initial parameter range from four to two orders of magnitude. For the latter conditions the model did converge.

3.4 Prediction sensitivity and uncertainty analysis - Methods

A global sensitivity analysis of the transient CDM Smith (2016) coal seam gas production impact model was undertaken using two global approaches: (1) the delta moment-independent measure (DMIM; Borgonovo, 2007; Plischke et al., 2013), and; (2) a variance-based measure (Sobol', 2001). Global sensitivity analyses, which provide a comprehensive assessment of the sensitivity of modelled outputs to model parameter variation, is currently under-utilised (Saltelli et al., 2008; Pianosi et al., 2016). Instead, local, one-at-a-time (OAT) sensitivity analyses are employed in the vast majority (i.e., 96 %) of all published model sensitivity analyses (Ferretti et al., 2016). Model outputs generated through global sensitivity analyses were also used as bases for prediction uncertainty quantification. In order to explain the advantages of global approaches over local sensitivity analysis, local OAT sensitivity analyses are first described.

3.4.1 One-at-a-time sensitivity analysis

Local sensitivity analysis is typically performed using a one-at-a-time procedure. Here, a base set of parameter values is assumed, which is often derived from deterministic model calibration. For a given modelled prediction, each model parameter value is adjusted by an infinitesimal amount (typically ± 1 %) while all other parameters retain "base" values; e.g., for a forward finite difference (Saltelli et al., 2000):

$$OAT(X_i) = \frac{Y|X_{i=x+1.01} - Y|X_{i=x}}{X_{i=x+1.01} - X_{i=x}}$$
(2)

where X_i = parameter *i* from a set of parameters $X_{i=1...N}$, x = the "base" value of parameter *i*, and Y|X = the value of prediction Y when using parameter set X. Although trivial to apply, the inability to vary parameter values simultaneously in this approach typically results in gross undersampling of the range of possible models and, therefore, predictions. For example, OAT sensitivity analysis of a twelve parameter model will explore less than one-thousandth of the total number of possible parameter combinations (Saltelli and Annoni, 2010). Furthermore, OAT approaches assume that relationships between a prediction between parameters are independent. For models featuring considerable correlation between parameters (such as many groundwater flow models), this assumption may be invalidated. Consequently, for such models, estimates of prediction uncertainty based on OAT sensitivity analyses are not robust. The ability to explore the effects of parameter correlation on prediction uncertainty was a key motivation behind the development of global sensitivity metrics.

3.4.2 Delta Moment-Independent Measure

The first global method used is known as the Delta Moment-Independent Measure (DMIM; Borgonovo et al., 2007; Plischke et al., 2013). The DMIM approach is based upon computing the differences in mass density between probability density functions (PDFs) of prediction values computed (A) when all parameter values are varied simultaneously and (B) when one parameter of interest is fixed at a constant value. An example of this approach is provided in Figure 3-1. The PDF of (A) is represented by a solid red line while the PDF of (B) is represented by a dashed blue line. The difference between these two PDFs is represented by the solid black shading. Mathematically, this difference in mass density [$s_i(x)$] is expressed as (Plischke et al., 2013):

$$s_i(x) = \int_{V} |f_Y(y) - f_{Y|X_i = x}(y)| \, \mathrm{d}y$$
(3)

where $f_Y(y)$ represents the PDF of predictions y and $|\cdot|$ represents the L¹ norm; i.e., the sum of absolute values. Essentially, this equation is used to compute the integral with respect to y of the absolute difference between the PDF of (A), expressed as $f_Y(y)$, and the PDF of (B), expressed as $f_{Y|X_i=x}(y)$. The DMIM sensitivity of a given model prediction to a parameter of interest $[\delta_i]$, is then calculated as one half of the expected value of $s_i(x)$; i.e.: (Plischke et al., 2013)

$$\delta_i = \delta(Y, X_i) = \frac{1}{2} E[s_i(x)] = \frac{1}{2} E\left[\int_{y} |f_Y(y) \ f_{Y|X=x}(y)| \, \mathrm{d}y\right]$$
(4)

where *E* is the expected value. A large δ_i value indicates that the prediction of interest is highly sensitive to parameter X_i . In practice, the integrals in equations (3) and (4) are evaluated numerically using a kernel density estimator.



Figure 3-1. Graphical representation of the Delta Moment Independent Measure (DMIM; Plischke et al., 2013) of global sensitivity (Borgonovo, 2007). The exponential distribution (red solid line) represents the probability density function (PDF) of modelled predictions when all model parameters are varied simultaneously. The skewed Gaussian distribution (dashed blue line) represents the PDF of modelled predictions when all model parameters are varied except a given parameter of interest. The difference in mass density between these two PDFs (solid black shading) is used as the basis for the calculation of the DMIM sensitivity metric.

3.4.3 Sobol' variance-based metric

The second global method used was a relatively simple approach to global sensitivity analysis derived by Sobol' (2001). This approach quantifies the uncertainty of a given prediction by the variance of the range of values. As such, it is implicitly assumed that the statistical distributions of predicted values can be characterised by their first two moments (i.e., mean and variance) alone. The primary advantage of this approach lies in its simplicity, whereby prediction uncertainty is characterised using scalar values while avoiding the need for numerical integration or quadrature.

The Sobol' approach assumes that both prior and posterior distributions can be characterised by their first two moments and that their uncertainty can be quantified by the second moment (i.e., variance) of these distributions. The first-order sensitivity of model predictions to model parameters is quantified as the ratio (i.e., S_1) between two values (Saltelli et al., 2008):

$$S_1(i) = \frac{V[E(Y|X_i)]}{V(Y)}$$
(5)

The numerator $V[E(Y|X_i)]$ represents the variance of the expected values of prediction Y when parameter value X_i is held constant (Saltelli et al., 2008). The denominator V(Y) represents the sum of (A) the variance of all expected values of prediction Y when X_i is held constant and (B) the expected value of the variance of prediction Y when X_i is held constant; i.e. (Saltelli et al., 2008):

$$V(Y) = V[E(Y|X_i)] + E[V(Y|X_i)]$$
(6)

Consequently, the Sobol' sensitivity metric represents the variance of model outputs when a parameter is fixed at a given value, normalised by the total model output variance. In practice, the sensitivity metric can be calculated for all model parameters; i.e., from $X_{i=1}$ to $X_{i=N}$.

3.4.4 Implementation of methods

DMIM and Sobol' global sensitivity analyses were implemented using the Python language library SALib (Herman and Usher, 2016). The Monte Carlo sampling employed by these algorithms provided the basis for prediction uncertainty analyses, which were visualised as scatterplots (see Appendix). A sample size of 300 model runs provided the basis for both algorithms. Uniform prior distributions with a range of two orders of magnitude were specified for all parameters tested. Mean values were specified according to the parameter values described by CDM Smith (2016). Parameter sampling was undertaken using a Latin Hypercube approach, implemented using the *SALib* library. Customised Python scripts were used to generate MODFLOW model input files, to undertake parallel model runs, and to post-process model outputs, including the calculation of sensitivity metrics (quantitative sensitivity analysis) and the generation of scatter plots of prediction versus parameter values for each of the four groundwater impact metrics (qualitative sensitivity analysis) (see Pianosi et al. (2016) for a review of helpful visualisation tools for global sensitivity analysis).

3.4.5 Real-time parameter re-sampling

As indicated by model stress testing results (Section 3.3), the numerical convergence of the CDM Smith (2016) model was not unconditionally robust, given the specified parameter prior distributions. Therefore it was not known prior to undertaking global sensitivity analyses whether parameter sets generated by the *SALib* library would be convergent. For this reason, parameter resampling was incorporated into the GSA algorithm. After the completion of each model run, a model convergence check as performed. If the model was found to have not converged then a new parameter set was generated and the model was rerun. This procedure was repeated until a convergent parameter set was achieved, which was then recorded. Inspection of the final set of 300 convergent parameter sets confirmed that comprehensive sampling of the ranges specified for each parameter was achieved (Table 3-2).

Mean values for a uniform or log-uniform distribution function are calculated as mean = (minimum + maximum)/2. Based on a log-uniform distribution, the mean $\log_{10} K_V$ (with K_V in m/day) for both the upper and lower aquitard is calculated to be $\log_{10} K_V = -5.0$. This value will be compared with updated values following the improved aquitard characterisation (Section 4).

Table 3-2. Specified and sampled minimum and maximum parameter values used in preliminary global sensitivity analysis of the preliminary CDM Smith (2016) groundwater flow model. All parameters were log_{10} transformed and random sampling of parameters was undertaken from log-uniform distributions. $K_{\rm H}$ and $K_{\rm V}$ in m/day, SS in m⁻¹.

n	Hydrostratigraphic unit	Parameter	Specified minimum	Sampled minimum	Specified maximum	Sampled maximum
1	Namoi Alluvium aquifer	<i>К_Н</i> 01	-0.300	0.338	1.700	1.613
2	Wallumbilla Formation aquitard	<i>К_н</i> 02	-5.000	-4.993	-3.000	-3.002
3	Blythesdale Group aquitard	<i>К</i> _н 03	-4.000	-3.994	-2.000	-2.002
4	Pilliga Sandstone aquifer	<i>К_н</i> 04	-2.000	-1.941	0.000	-0.016
5	Upper aquitard sequence	<i>К_н</i> 05	-4.000	-3.994	-2.000	-2.006
6	Hoskissons Coal	<i>К_н</i> 06	-2.000	-1.993	0.000	-0.006
7	Lower aquitard sequence	<i>К_н</i> 07	-4.000	-3.990	-2.000	-2.002
8	Maules Creek Formation interburden	<i>К_Н</i> 08	-4.000	-3.984	-2.000	-2.002
9	Maules Creek Formation coal	<i>К_н</i> 09	-2.000	-1.996	0.000	-0.000
10	Maules Creek Formation interburden	<i>K_H</i> 10	-4.000	-3.993	-2.000	-2.005
11	Namoi Alluvium aquifer	<i>K</i> _V 01	-1.300	-1.295	0.700	0.697
12	Wallumbilla Formation aquitard	<i>K</i> _v 02	-6.000	-6.295	-4.000	-4.305
13	Blythesdale Group aquitard	<i>K</i> _v 03	-6.000	-5.990	-4.000	-4.010
14	Pilliga Sandstone aquifer	<i>K</i> _V 04	-3.000	-2.996	-1.000	-1.000
15	Upper aquitard sequence	<i>K</i> _v 05	-6.000	-5.968	-4.000	-4.007
16	Hoskissons Coal	<i>K</i> _v 06	-3.000	-2.981	-1.000	-1.001
17	Lower aquitard sequence	<i>K</i> _v 07	-6.000	-5.989	-4.000	-4.004
18	Maules Creek Formation interburden	<i>K</i> _V 08	-6.000	-5.993	-4.000	-4.012
19	Maules Creek Formation coal	<i>K</i> _v 09	-3.000	-2.993	-1.000	-1.000
20	Maules Creek Formation interburden	<i>K</i> _V 10	-6.000	-5.999	-4.000	-4.001
21	Namoi Alluvium aquifer	<i>S</i> ₅ 01	-6.000	-5.999	-4.000	-4.003
22	Wallumbilla Formation aquitard	S ₅ 02	-6.000	-5.999	-4.000	-4.001
23	Blythesdale Group aquitard	S ₅ 03	-6.000	-5.995	-4.000	-4.005
24	Pilliga Sandstone aquifer	<i>S</i> _s 04	-6.000	-6.000	-4.000	-4.005
25	Upper aquitard sequence	<i>S</i> _s 05	-6.000	-5.978	-4.000	-4.006
26	Hoskissons Coal	<i>S</i> _s 06	-6.000	-5.992	-4.000	-4.019
27	Lower aquitard sequence	S _s 07	-6.000	-5.991	-4.000	-4.001
28	Maules Creek Formation interburden	<i>S</i> ₅ 08	-6.000	-5.977	-4.000	-4.001
29	Maules Creek Formation coal	S ₅ 09	-6.000	-5.991	-4.000	-4.010
30	Maules Creek Formation interburden	S _s 10	-6.000	-5.994	-4.000	-4.050

3.4.6 Statistical convergence testing

A sample size of 300 model runs was initially selected for pragmatic reasons; i.e., to limit both the total model processing time and computer storage required. Each model run required approximately 30 minutes to complete and produced approximately two gigabytes of output files.

Furthermore, because of the proprietary status of the finite-difference code MODFLOW-SURFACT (HydroGeoLogic, 1996), distributed or parallel computing was not an option, although this would have reduced the overall runtime and thus would have allowed for a larger number of model runs. Such practical implications need careful consideration when selecting the modelling software, especially for stochastic simulations. Whatever the initial number of model runs selected, it is important to test if the sensitivity metrics have converged for a given number of Monte Carlo runs (Nossent et al., 2011).

Following the completion of 300 convergent model runs, the statistical robustness of results based on this sample size was then tested by recalculating the relative ranking of each parameter, based on the DMIM δ and Sobol' S_1 quantitative sensitivity metrics. These two metrics were recalculated using sample sizes of 200, 220, 240, 260, 280 and 300 model runs. The relative rankings of the four modelled predictions are compared with respect to increasing sample size in Figure 3-2.



Figure 3-2. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which predictions relating to the Pilliga Sandstone aquifer were most sensitive: (a) maximum drawdown (MXD); (b) time elapsed at which maximum drawdown occurred (tMXD); (c) number of model cells at which drawdown exceeded two metres; and (d) maximum change in vertical flux (MXQ). Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs. Parameters identified are discussed in Section 3.5.

For each prediction, at least one model parameter was consistently ranked first based on sample sizes greater than or equal to 260 model runs. Note that the relative rankings are only significant if at least one of the two metrics occupies systematically the top rank for each and every sample size. For example, for the maximum drawdown (MXD), there is only one parameter for which the Sobol' *S*₁ metric is consistently ranked first. For the extent of maximum drawdown (NDD), there are two parameters consistently ranked first and second, indicating only these two parameters are significantly affect the groundwater impact metric (here NDD). As will be shown later, these diagrams of relative rankings are best supplemented with diagrams displaying the magnitude of

the sensitivity metrics for all predictions tested (i.e., the larger the metric, the more sensitive is the model prediction to this parameter).

The statistical robustness of results based on a sample size of 300 model runs was also tested by recalculating the uncertainty (i.e., statistical spread) of model predictions. The uncertainties of the four modelled predictions are compared with respect to increasing sample size (from 200 to 300, Figure 3-3. The uncertainties of all predictions was estimated consistently based on sample sizes greater than or equal to 260 model runs for prediction MXD, 200 model runs for prediction tMXD, and 280 model runs for predictions NDD and MXQ. Notably, the statistical distributions of three of the four predictions were highly skewed. For the right-skewed magnitude of maximum drawdown prediction (MXD), this was due to a large number of predictions less than one metre in magnitude. For the left-skewed timing of maximum drawdown (tMXD), this was due to the specification of a model temporal extent (i.e., a 160 year-long period of response after the cessation of CSG extraction) that was insufficient to capture peak drawdown conditions for a subset of the model runs. Given that a set of 300 model runs took approximately one week to complete, it was not feasible to repeat the sampling process using an extended model duration within the duration of the present study. For the right-skewed prediction of the spatial extent of drawdown (NDD), the skewed distribution was due to a large number of zero-value predictions (i.e. no grid cells with maximum drawdown exceeding 2 m).



Figure 3-3. Statistical distributions of four predictions relating to the Pilliga Sandstone aquifer and simulated using the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model: (a) maximum drawdown (MXD); (b) time elapsed at which maximum drawdown occurred; (c) number of model cells at which drawdown exceeded two metres (NDD) ; and (d) maximum change in vertical flux (MXQ). The size of samples described by each of these statistical distributions ranged from 200 to 300 model runs.

3.5 Prediction sensitivity analysis - Results and Discussion

A full set of qualitative results and quantitative results (i.e., DMIM δ and Sobol' S_1 metric values, and associated relative rankings) of preliminary prediction sensitivity analyses are presented in the next sections. Key results, in terms of parameters to which predictions were found to be most sensitive, are summarised as follows.

3.5.1 Prediction metric 1: Magnitude of maximum drawdown

The prediction of maximum drawdown was found to be most sensitive to the horizontal K of the Namoi Alluvium aquifer (K_H 01), which was consistently ranked as the most influential parameter

by the Sobol' S_1 metric across all sample sizes (Figure 3-4). This parameter (K_H 01) featured the largest δ and S_1 values (0.096 and 0.118, respectively) across all parameters (Figure 3-5). Using the DMIM δ metric, the horizontal K of the Namoi Alluvium was ranked first in all but one sample size (i.e., a DMIM δ analysis of a subset containing 280 samples, Figure 3-4a). The high sensitivity of this prediction to the horizontal K of the Namoi Alluvium is explained as follows. Significant volumes of water were removed from the model via negative net recharge conditions imposed within the Namoi Alluvium. This water was sourced predominantly from underlying hydrostratigraphic units; specifically, the Pilliga Sandstone aquifer. Changes to the horizontal K of the Namoi Alluvium directly affected the hydraulic gradient between the Namoi Alluvium and the Pilliga Sandstone. Hydraulic head values (and therefore drawdown values) were directly affected by variations in the horizontal K of the Namoi Alluvium.

As discussed in Section 3.4.6, the uncertainty of the magnitude of maximum drawdown did not vary significantly for sample sizes \geq 280 (Figure 3-3a). Predicted values of MXD ranged from 0.0 m to 5.5 m. The 10th percentile value ranged from 0.1 m to 0.2 m; the median value ranged from 0.8 m to 0.9 m; and the 90th percentile value ranged from 2.4 m to 2.5 m.



Figure 3-4. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the magnitude of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.



Figure 3-5 Global sensitivity analysis metrics of all 30 model parameters in relation to maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation was 300 model runs.

Maximum drawdown was also found to be weakly sensitive to the vertical *K* of the upper aquitard sequence (K_V 05; Figure 3-4). The ranking of the latter parameter converged with increasing sample size. Note, however, that inspection of the magnitude of sensitivity metrics (Figure 3-5) shows that K_V 05 sits in the noise and is only weakly significantly more sensitive than other parameters (δ =0.092, S_1 =0.048). Indeed, DMIM δ values ranged from 0.016 to 0.069 for the remainder of the parameters examined.

3.5.2 Prediction metric 2: Timing of maximum drawdown

The prediction of time of maximum drawdown was found to be equally insensitive to most parameters tested (Figure 3-6a, b); i.e., DMIM δ values ranged within a narrow band between 0.041 and 0.078 (Figure 3-7a), while the Sobol' S_1 metric also varied within a relatively narrow range (Figure 3-7b). As a result, for the model conditions tested here, there is no single model parameter to which the timing of maximum drawdown is more sensitive than any other parameter.

As demonstrated in Section 3.4.6, the uncertainty about timing of maximum drawdown did not vary significantly with increasing sample size (Figure 3-3b). Predicted values ranged from 19 to 261 years across all 300 model runs. For sample sizes of 200 to 300, the 10th percentile value ranged from 117 years to 134 years and the median value ranged from 253 years to 255 years. The 90th percentile value was consistently 261 years, regardless of the sample size. When interpreting these results, and as stated previously, for a subset of model runs the model temporal extent specified was insufficient to capture peak drawdown conditions. Due to large model run times it was not feasible to repeat the sampling process using an extended model duration within the duration of the present study.



Figure 3-6. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the timing of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.





3.5.3 Prediction metric 3: Spatial extent of drawdown propagation

The prediction of the spatial extent of drawdown greater than or equal to two metres was found to be most sensitive to the horizontal K of the Namoi Alluvium aquifer (K_H 01), which was

consistently ranked as the most influential parameter with increasing sample size (Figure 3-8a, b). For interpretation of this result with regards to the structure of the groundwater flow model, see Section 3.5.1. Sobol' analyses also indicated consistent sensitivity to the horizontal K of the Maules Creek Formation interburden (K_H 08) (Figure 3-9b).

The K_H 01 parameter had significantly larger δ and S_1 values (0.125 and 0.101, respectively) across all parameters (Figure 3-9a, b). The values of both sensitivity metrics for this model parameter were significantly larger than all other model parameters. DMIM δ values ranged from 0.062 to 0.089 for the remainder of the parameters examined (Figure 3-9a).

As shown previously in Section 3.6.4, the uncertainty of this prediction did not vary significantly with increasing sample size (Figure 3-3). Predicted values ranged from zero to 102 cells displaying drawdowns ≥ 2 m. The 10th percentile and median values were consistently zero (i.e., no model cells featured drawdown ≥ 2 m) and the 90th percentile value was consistently 60 cells with drawdown ≥ 2 m.



Figure 3-8. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of drawdown spatial extent in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.



Figure 3-9 Global sensitivity analysis metrics of all 30 model parameters in relation to spatial extent of maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation was 300 model runs.

3.5.4 Prediction metric 4: Maximum vertical flux

The prediction of maximum flux was found to be most sensitive to the vertical *K* of the upper aquitard sequence (i.e., K_V 05) for both sensitivity metrics (Figure 3-10a, b). The values of both sensitivity metrics for this parameter were significantly larger than all other parameters (δ =0.206, S_1 =0.193) (Figure 3-11a, b). DMIM δ values for the remainder of the parameters examined ranged within a narrow band from 0.028 to 0.087; Sobol' S_1 values ranged with a narrow band from 0.05 to 0.1.

The uncertainty of this prediction did not vary significantly with increasing sample size (Section 3.4.6, Figure 3-3). Predicted values ranged from 17 to 38 m³/d across all 300 model runs. For sample sizes of 200 to 300, the 10th percentile value was consistently 23 m³/d; the median value ranged from 26.7 to 27.0 m³/d; and the 90th percentile value ranged from 33 to 34 m³/d.



Figure 3-10. Relative rankings of ten parameters used by the CDM Smith (2016) Gunnedah Basin groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of maximum change in vertical flux in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.



Figure 3-11 Global sensitivity analysis metrics of all 30 model parameters in relation to maximum vertical flux. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation was 300 model runs.

3.5.5 Summary of prediction metric sensitivities

With the exception of one prediction metric (i.e., the time of maximum drawdown), each of the other three prediction types were found to be most sensitive to one of three parameters: the horizontal K of the Namoi Alluvium aquifer (K_H 01), the horizontal K of the Maules Creek Formation

interburden (K_H 08), and the vertical K of the upper aquitard sequence (K_V 05). The influence of the K_H 01 parameter is interpreted as follows: this parameter affects the movement of water removed from the top layer of the model as negative net recharge. When K_H 01 is relatively small, the hydraulic gradient between the Namoi Alluvium aquifer and the indirectly underlying Pilliga Sandstone aquifer (i.e., the nearest additional water source) is higher. This results in relatively lower hydraulic heads in the Pilliga Sandstone aquifer, which are observed as drawdown. Conversely, when K_H 01 is relatively large, the hydraulic gradient between the two aquifers is lower; consequently, hydraulic heads in the Pilliga Sandstone aquifer are not significantly affected. The influence of the vertical K of the upper aquitard sequence K_V 05 is interpreted as the parameter controlling the upward propagation of hydraulic stresses to the Pilliga Sandstone aquifer.

Based on these findings, further characterisation of the three parameters (K_H 01, K_V 05, K_H 08) is warranted, assuming that the uncertainty associated with each of these parameters is either poorly characterised or known to be large. The question of parameter uncertainty was first addressed in Section 3.4. The initial parameter uncertainty was defined by assuming log-uniform distributions with arbitrary ranges defined by the CDM Smith (2016) parameter value (for each of K_H , K_V and S_S) ± two orders of magnitude. Very few data were available to define a more credible parameter distribution; hence, the initial uncertainty was indeed poorly defined and therefore improved characterisation of the three parameters (K_H 01, K_V 05, K_H 08) was justified. Given the specific focus of the present study on aquitard properties and their representation, the improved characterisation of aquitard (e.g., K_V 05) hydraulic conductivity was pursued in this study, rather than aquifer (e.g., K_H 01) or coal seam (e.g., K_H 08) hydraulic conductivity.

3.6 Prediction uncertainty analysis - Results and Discussion

The uncertainty of predictions generated using the CDM Smith (2016) Gunnedah-Surat basins model is described using summary statistics as follows. Differences between maximum and minimum values do not provide robust estimates of prediction uncertainty, as these are highly sensitive to the presence of outlying values. Instead, prediction uncertainty is quantified here using three statistical measures: (1) standard deviation, which is commonly used in linearised prediction uncertainty analyses (though typically implemented as variance); (2) the interquartile range (i.e., the difference between the 25th and 75th percentiles), which is most robust to outliers; and (3) the 90 % confidence interval (i.e., the difference between the 5th and 95th percentiles).

3.6.1 Prediction 1: Magnitude of maximum drawdown

Using a sample size of 300 model runs, predictions of the magnitude of maximum drawdown at any active Pilliga Sandstone aquifer cell in the model domain ranged from < 1.0 m to 29.7 m, with a median value of 0.7 m (Figure 3-12). The strongly right-skewed frequency histogram indicated that many model runs predicted near-zero maximum drawdown values. The standard deviation of predicted values was 3.0 m. The interquartile range of predicted values was 3.0 m, ranging from 0.1 m to 3.1 m. The 90 % confidence interval was 0.6 m, ranging from 1.8 m to 2.3 m. It should be noted that the numerical convergence criterion with regards to hydraulic head (i.e., 0.1 m) should be considered as an acceptance threshold for this prediction. The magnitude of maximum drawdown was calculated as \leq 0.1 m in 65 of the 300 model runs.



Figure 3-12. (a) Cumulative density function and (b) frequency histogram of initial modelled predictions of the magnitude of maximum drawdown, based on a sample size of 300 model runs.

3.6.2 Prediction 2: Timing of maximum drawdown

Using a sample size of 300 model runs, predictions of the timing of maximum drawdown ranged from 1 years to 160 years, with a median value of 154 years (Figure 3-13). The standard deviation of predicted values was 42 years. The interquartile range of predicted values was 46 years, ranging from 114 years to 160 years. The 90 % confidence interval was 8 years, ranging from 127 years to 135 years. It should be noted that the maximum simulation time specified for the model (i.e., 160 years after the cessation of CSG extraction) significantly affected this prediction type. Nearly one third of all model runs (i.e., 95 runs) resulted in a predicted timing of maximum drawdown value of 160 years. These results indicated that the maximum simulation time of these models was possibly not sufficiently large to capture the maximum drawdown responses. Due to large model run times it was not feasible to repeat the sampling process using an extended model duration within the duration of the present study. This factor is also the cause of the non-asymptotic behaviour of the cumulative density function for this prediction (Figure 3-13a) and the left-skewed frequency histogram (Figure 3-13b). As mentioned earlier, the 160-year long simulation period was chosen to be consistent with the approach taken for the Bioregional Assessments Programme.



Figure 3-13. (a) Cumulative density function and (b) frequency histogram of initial modelled predictions of the timing of maximum drawdown, based on a sample size of 300 model runs.

3.6.3 Prediction 3: Spatial extent of drawdown propagation

Using a sample size of 300 model runs, predictions of the spatial extent of drawdown propagation in the Pilliga Sandstone aquifer (i.e., model cells with drawdown ≥ 2 m) ranged from zero to 10 418 cells, with a median value of zero cells (Figure 3-14a). The strongly right-skewed frequency histogram indicated that a large number of model runs featured zero cells with drawdown ≥ 2 m (Figure 3-14b).



Figure 3-14. (a) Cumulative density function and (b) frequency histogram of initial modelled predictions of the spatial extent of drawdown ≥ 2 m, based on a sample size of 300 model runs.

Note that the cumulative density function presented is a truncated distribution, in which approximately 70 % of the model runs featured 'zero impact' (i.e., for which the maximum reported drawdown for Pilliga Sandstone aquifer cells was less than 2 m). The standard deviation of predicted values was 1 829 cells. The interquartile range of predicted values was 1 496 cells,

ranging from zero to 1 496 cells. The 90 % confidence interval was 347 cells, ranging from 822 cells to 1 169 cells.

Spatial maps with isolines of drawdowns are shown in Figure 3-15 for the 5th, 50th, 90th, 95th, and 99th percentiles. Maps were constructed by calculating at each model cell the cumulative distribution of drawdowns, based on 300 calculated values of drawdown for each model cell. The drawdowns after 155 years following cessation of gas extraction are used here; this time corresponds to the median time to maximum drawdown across all 300 model runs. Isolines shown are for 1 and 2 m drawdown (Figure 3-15); as expected, larger areas of drawdown are observed for the higher percentiles, while the 5th and 50th percentiles have zero cells with drawdowns > 1 m.



Figure 3-15. Percentiles of spatial distributions of 1 and 2 m drawdown (solid black lines), observed at the median time of maximum drawdown (i.e., 155 years after the cessation of coal seam gas extraction); purple = Pilliga Sandstone aquifer cells; blue = upper aquitard sequence cells. Drawdowns < 1 m are not shown. Percentiles shown are (a) 5th, (b) 50th, (c) 90th, (d) 95th and (e) 99th. Distributions are based on a set of 300 model runs using the 'initial' model.

3.6.4 Prediction 4: Maximum vertical flux

Using a sample size of 300 model runs, predictions of the maximum vertical flux in any active Pilliga Sandstone aquifer cell in the model domain ranged from 294 m³/d to 1418 m³/d, with a median value of 737 m³/d (Figure 3-16). Approximately 25% of all model runs featured a maximum vertical flux of around 550 m³/d. The standard deviation of predicted values was 241 m³/d. The interquartile range of predicted values ranged over 363 m³/d, from 548 m³/d to 911 m³/d. The 90 % confidence interval was 46 m³/d, ranging from 757 m³/d to 803 m³/d.



Figure 3-16. (a) Cumulative density function and (b) frequency histogram of initial modelled predictions of changes in vertical groundwater fluxes, based on a sample size of 300 model runs.

4 Improved characterisation of aquitard vertical hydraulic conductivity

4.1 Core testing

CDM Smith (2016) reported there is generally a lack of published information describing the range and spatial distribution of hydrogeological properties of the deep consolidated strata within the Gunnedah and Surat basins. Only limited data were available for strata within the vicinity of the Project area, which relate predominantly to coal seam geology. Drill stem tests had been undertaken almost exclusively on the coal seams. For the Napperby and Purlawaugh aquitard formations, limited falling head data was available from the Narrabri Coal Mine groundwater model. No *K*_V estimates were available for Watermark and Porcupine aquitard formations (CDM Smith, 2016). This paucity of hydraulic conductivity data illustrates that there is very little opportunity to couple existing hydrogeologic parameters to geological, geophysical or seismic data. Because of this paucity of data, Turnadge et al. (2018b) decided to collect core samples of the key aquitard sequences and correlate the measured permeability with the available geophysical borehole data.

Tri-axial core testing was used to estimate porosity-permeability relationships for the two aquitard sequences underlying the Pilliga Sandstone aquifer (Turnadge et al., 2018b); i.e., (a) the Triassic-Jurassic age Purlawaugh to Napperby formation sequence and (b) the Permian age Watermark to Porcupine formation sequence. Parametric analytical models were then derived to describe the porosity-permeability relationship for four aquitard units located within the two aquitard sequences (i.e., upper sequence: Purlawaugh and Napperby formations; lower sequence: Watermark and Porcupine formations). Downhole porosity logs were then obtained for 97 exploration wells located across the Gunnedah Basin. The derived porosity-permeability models were applied to these logs to estimate vertical distributions of permeability (and therefore vertical hydraulic conductivity, or K_V) at various locations across the Gunnedah Basin. Upscaling approaches were subsequently applied to upscale these core scale aquitard K_V values for inclusion in the regional-scale numerical groundwater flow model.

4.2 Parameter upscaling

The upscaling of parameter values from small support scales (e.g., centimetres, for core samples) to large support scales (e.g., kilometres, for regional-scale flow processes) can be undertaken using (1) analytical, (2) numerical or (3) geostatistical methods (Sanchez-Vila et al., 1995, 2006; Wen and Gomez-Hernandez, 1996; Renard and de Marsily, 1997; Li et al., 2011; Moore et al., 2013; Turnadge et al., 2018a). In the present study, the former two approaches were used to upscale core-scale aquitard vertical hydraulic conductivity values to values suitable for inclusion in a regional-scale groundwater flow model. The primary analytical approach to upscaled values were also derived using arithmetic and geometric mean values. The numerical upscaling approach involved

the use of a one-dimensional steady-state confined groundwater flow model which was solved using a finite difference scheme.

The inherent assumption when upscaling parameters from the core-scale to numerical grid cells is that whatever upscaling approach is used does not introduce unacceptable errors and thus uncertainties in the large-scale K_V when extrapolating such parameter beyond the measurement support. Several studies have demonstrated that in low-permeable media, K values are fairly insensitive to the scale of measurement, even beyond the typical core or borehole scale (up to a certain degree). E.g., Yu et al. (2013) demonstrated for plastic Boom Clay of marine origin that K_V was nearly identical across a range of scales from cm to a macro-permeameter with a flow crosssectional area of about 190 m². In other words, the representative elementary volume (REV) for K_V is orders of magnitude larger than the core scale. Using numerical experiments, McKenna and Rautman (1996) demonstrated that upscaling errors were around 5% for mildly heterogeneous Kfields ($\sigma \log K = 0.5$) when specific upscaling methods were used (i.e., geometric mean, numericalinverse method, and renormalization scaling techniques). While there are definitely stratigraphic and other differences between the literature studies and the current study area, the presence of fine-grained units such as claystones (e.g., Benelabri and Leard Formations) and shales (e.g., Porcupine and Watermark Formations) indicate that a similarly large REV may exist.

An alternative approach to upscaling properties from the pore to the core scale was discussed by Arena et al. (2016). These authors used upscaling techniques based on micro-CT imaging data at scales ranging from micro-scale to core-scale combined with independent core-plug measurements of permeability and wireline log data. Main impediments to developing a mature whole core to log scale workflow were associated with the ability to robustly identify rock types and upscale data from the plug scale to the whole core scale and the need for improved direct calibration of the log response to the whole core scale properties.

4.2.1 Analytical upscaling methods

Analytical approaches to upscaling parameter values involve calculation of the arithmetic, geometric, harmonic, power or volume-weighted mean of a given sample of values (Li et al., 2011). For one-dimensional flow in heterogeneous porous media, the equivalent hydraulic conductivity can be calculated using the harmonic mean (Freeze and Cherry, 1979). For two-dimensional heterogeneous media, it has been shown that under certain conditions the geometric mean is an appropriate means of parameter upscaling (Gomez-Hernandez and Wen, 1998; Sanchez-Vila et al., 1995).

In the present study harmonic averaging was used to upscale 64 and 78 vertical profiles of K_V to scalar values for the upper and lower aquitard sequences, respectively. Numerous vertical profiles were omitted from further analysis (from the initial sample set of 97 profiles) on the basis of unsuitable location and/or geology. For the upper aquitard sequence, only the wells that featured the presence of either the Purlawaugh or Napperby formations were retained (64 out of 97). Similarly, 78 of the 97 wells featured the presence of either the Watermark or Porcupine formations for the lower aquitard sequence. The resulting $log_{10} K_V$ values (m/d) are summarised in Table 4-1.

Table 4-1. Statistical summary of log₁₀ aquitard vertical hydraulic conductivity (log₁₀ K_V) values upscaled from corescale observations using harmonic averaging.

Statistic	Upper aquitard sequence	Lower aquitard sequence
count	64	78
minimum	-7.0	-7.5
5 th percentile	-6.6	-7.0
median	-5.4	-4.6
mean	-5.3	-4.9
95 th percentile	-4.0	-3.3
maximum	-3.2	-3.1
standard deviation	0.8	1.1



Figure 4-1. Original and revised statistical distributions of vertical permeability values (left y-axis) and vertical hydraulic conductivity values (right y-axis) for (a) upper and (b) lower aquitard sequences represented in the CDM Smith (2016) model of groundwater flow in the Gunnedah-Surat basins. Original values and ranges were described by CDM Smith (2016). Revised values and ranges are based on aquitard core porosity-permeability analyses and upscaled using harmonic averaging, described herein. Whiskers indicate the 90 % confidence interval and fliers indicate outlying data values.

From Figure 4-1 it may be observed that, for both aquitard sequences, the revised prior distributions of K_V are two orders of magnitude larger than the distributions assumed for the performance of the initial sensitivity analyses. Note that the original log-uniform prior distribution was replaced by a uniform log-triangular distribution, which featured finite minimum and maximum values (unlike, for example, a Gaussian distribution). This represented a change from an "uninformed" prior distribution (in which all values were considered equally likely) to a distribution featuring a single, most-likely value (e.g., mode).

4.2.2 Numerical upscaling methods

Numerical flux-based approaches aim to preserve the hydraulic mass balance (and therefore groundwater flow rates) when upscaling parameter values from one scale to another larger scale.

Li et al. (2011) summarised flux-based upscaling methods involving the Laplace equation for groundwater flow. The upscaled hydraulic conductivity in any given direction can be calculated using Darcy's Law as the ratio of the flow rate to the cross-sectional thickness divided by the hydraulic gradient. Numerical methods have been widely used to upscale hydraulic conductivities in petroleum engineering and hydrogeology (e.g., Warren and Price, 1961; Journel et al., 1986; Desbarats, 1987; Deutsch, 1989).

Numerical methods were used to upscale 54 and 69 vertical profiles of K_V obtained from core-log analyses for the upper and lower aquitard sequences, respectively. Numerous vertical profiles of the initial set used for analytical upscaling were omitted from further analysis (from the initial sample set of 64 for the upper aquitard, 54 were retained; from the initial 78 for the lower aquitard 69 were retained here). The reason for excluding additional wells was lack of hydraulic conductivity data for interburden layers connecting the two aquitards within a single sequence.

The vertical resolution of each vertical profile was first upscaled from non-uniform centimetrescale intervals to uniform one metre intervals by use of arithmetic mean values. These values were then incorporated into a steady-state 1-D numerical groundwater flow model. Each value was assigned to a unique cell of one metre thickness. Dirichlet (i.e., fixed head) boundary conditions were applied to the top and bottom cells of the model in order to apply a unit vertical hydraulic gradient (i.e., equal to the vertical extent of the model) across the model. The dimensions of model cells in the *x-y* plane were each specified as $1 \times 1 \text{ m}^2$; therefore the cross-sectional area perpendicular to outflow from the model was equal to unity. The equivalent vertical hydraulic conductivity was thereby calculated as being equal to the volumetric outflux from the model. The numerical approach was implemented using the MODFLOW-2005 code (Harbaugh, 2005) with preand post-processing of model files performed using the *FloPy* Python language library (Bakker et al., 2016).

Statistic	Upper aquitard sequence	Lower aquitard sequence
count	54	69
minimum	-6.7	-7.5
5 th percentile	-6.6	-7.0
median	-5.3	-4.4
mean	-5.2	-4.7
95 th percentile	-3.5	-3.3
maximum	-3.2	-3.1
standard deviation	0.9	1.2

Table 4-2. Statistical summary of \log_{10} transformed aquitard vertical hydraulic conductivity ($\log_{10} K_V$) values upscaled from core-scale observations using numerical averaging. K_V in m/day.

It should be noted that the range (i.e., width) of the revised statistical distribution of aquitard K_V (i.e., four orders of magnitude) was two orders of magnitude larger than the distribution initially assumed. This difference was also observed for aquitard K_V values upscaled using harmonic averaging. (Figure 4-2).



Figure 4-2. Original and revised statistical distributions of vertical hydraulic conductivity values for (a) upper and (b) lower aquitard sequences represented in the CDM Smith (2016) model of groundwater flow in the Gunnedah-Surat basins. Original values and ranges were described by CDM Smith (2016). Revised values and ranges are based on aquitard core porosity-permeability analyses and upscaled using numerical averaging, described herein. Whiskers indicate the 90 % confidence interval and fliers indicate outlying data values.

4.2.3 Comparisons between methods

Histograms of K_V values computed using the analytical and numerical upscaling approaches, for both upper and lower aquitard sequences, are presented in Figure 4-3. Good agreement is observed between the two approaches. This was confirmed through the calculation of Walsh's *t*test for correlation between two independent samples. Specifically, this test is used to test whether the null hypothesis, in this case, that both distributions are identical, can be rejected. The test returned *p* values of 0.532 and 0.559 for the upper and lower aquitard sequences respectively. These large *p* values (i.e., >> 0.01) indicated that the null hypothesis (i.e., that both distributions are identical) could not be rejected.

Upper aquitard sequence values were right-skewed, with mean $\log_{10} K_V$ values of approximately -5.2 m/d and -5.3 m/d for the harmonic and numerical upscaling approaches, respectively. Conversely, lower aquitard sequence values were left-skewed, with mean $\log_{10} K_V$ values of approximately -4.9 m/d and -4.7 m/d for the harmonic and numerical upscaling approaches, respectively. For the numerical upscaling approach applied to the lower aquitard sequence, the minimum and maximum $\log_{10} K_V$ was -7 and -3.1, respectively (Table 4-2). Smith et al. (2018) previously estimated K_V from rates of vertical fluid flow across the lower aquitard sequence through quartz-helium analyses. Smith et al. (2018) estimated rates of fluid flow of 0.02 to 0.002 mm/year, equivalent to K_V values of 0.0018 – 0.018 mm/year or $\log_{10} K_V$ between -8.3 and -7.3. This demonstrates that the range of (numerically) upscaled K_V values for the lower aquitard brackets K_V values independently estimated based on vertical fluid flow in the same aquitard unit. Because the helium in quartz technique is more of an in situ approach (helium concentrations are fairly insensitive to sampling), it does not have the issues typically experienced when permeability is determined on core samples that have been disrupted during sampling (Schulze-Makuch et al., 1999).



Figure 4-3. Histograms of aquitard vertical hydraulic conductivity values upscaled using (a) harmonic mean averaging and (b) numerical model averaging for the upper aquitard sequence (upper row) and lower aquitard sequence (lower row).

5 Prediction sensitivity and uncertainty analyses using revised models

Following the improved characterisation of aquitard vertical hydraulic conductivities, a second global sensitivity and uncertainty analysis was undertaken using updated prior parameter distributions. The uncertainties of the four predictions of interest were again estimated and subsequently compared to the initial results.

5.1 Prediction sensitivity and uncertainty analysis - Methods

As described in Chapter 4, the characterisation of vertical hydraulic conductivity of the upper and lower aguitard sequences of the Gunnedah-Surat basins was revised. As mentioned earlier, in the study area the Gunnedah Basin is overlain by the Jurassic-Cretaceous Surat Basin. The Surat Basin strata present in the vicinity of the study area include the Blythesdale Group, (Keelindi Beds), Pilliga Sandstone, Purlawaugh Formation and basal Garrawilla Volcanics. Core scale observations previously reported by Turnadge et al. (2018a) were upscaled using analytical and numerical methods and the prior distributions of aquitard K_V parameters were adjusted accordingly. The variability of both prior distributions, one for each aquitard sequence, was increased from two orders of magnitude (chosen arbitrarily) to four orders of magnitude (informed by observed data). The revised distributions were both characterised using triangular distributions (Figure 5-1). The distribution used to characterise the upper aquitard sequence was bounded at $log_{10} K_V = -7.0$ and $\log_{10} K_V = -3.0$ with a peak value of ~20 located at $\log_{10} K_V = -5.8$. Note that distribution parameters were nearly identical for both harmonic and numerical upscaling approaches. Similarly, the distribution used to characterise the lower aquitard sequence was bounded at $log_{10} K_V = -7.0$ and $\log_{10} K_V = -3.0$ but featured a peak value of ~20 located at $\log_{10} K_V = -4.2$, with similar parameter values observed for both harmonic and numerical upscaling approaches. Given the insensitivity of the four modelled predictions to specific storage parameters (as demonstrated in Section 3), the ten S_s parameters were omitted from the global sensitivity analysis of the revised CDM Smith (2016) model. This omission also served to considerably reduce the total model run time required for global sensitivity analysis.


Figure 5-1. Histograms (grey) and fitted triangular distributions (red) of aquitard vertical hydraulic conductivity values upscaled using (a) harmonic mean averaging and (b) numerical model averaging for the upper aquitard sequence (upper row) and lower aquitard sequence (lower row).

5.1.1 Statistical convergence testing

Following the completion of 300 convergent model runs, the statistical robustness of results based on this sample size were again tested by recalculating the relative ranking of each parameter based on the DMIM δ and Sobol' S_1 quantitative sensitivity metrics. The metrics were recalculated using sample sizes of 200, 220, 240, 260, 280 and 300 model runs. The relative rankings of the four modelled predictions are compared with respect to increasing sample size in Figure 5-2.

The statistical robustness of results based on a sample size of 300 model runs were also again tested by recalculating the uncertainty (i.e., statistical spread) of model predictions. The uncertainties of the four modelled predictions are compared with respect to increasing sample size in Figure 5-3. The uncertainties of all predictions was estimated consistently based on sample sizes greater than or equal to 200 model runs. As observed in the initial sensitivity analysis (Section 3.4.6), the statistical distributions of three of the four predictions were highly skewed. For the magnitude of maximum drawdown prediction (MXD), this was due to a large number of predictions less than one metre in magnitude. For the timing of maximum drawdown (tMXD), this was due to a total simulation time that was too short to capture peak drawdown conditions. For the prediction of the spatial extent of drawdown (NDD), this was due to a large number of model runs that did not feature drawdown ≥ 2m.



Figure 5-2. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (red, blue, green; others shown in grey) to which predictions relating to the Pilliga Sandstone aquifer were most sensitive: (a) maximum drawdown (MXD); (b) time elapsed at which maximum drawdown occurred (tMXD); (c) number of model cells at which drawdown exceeded two metres (NDD); and (d) maximum change in vertical flux (MXQ). Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (Plischke et al., 2013) and (b) the first order Sobol' metric (Sobol', 2001). Parameters identified are discussed in Section 5.2.



Figure 5-3. Statistical distributions of four predictions relating to the Pilliga Sandstone aquifer and simulated using the revised Gunnedah Basin groundwater flow model: (a) maximum drawdown (MXD); (b) time elapsed at which maximum drawdown occurred (tMXD); (c) number of model cells at which drawdown exceeded two metres (NDD); and (d) maximum vertical flux (MXQ).

5.2 Prediction sensitivity analysis – Results and Discussion

A full set of qualitative results (i.e., scatterplots) and quantitative results (i.e., DMIM δ and Sobol' S_1 metric values and associated relative rankings) of global prediction sensitivity analyses using the revised CDM Smith (2016) model are presented in the accompanying Appendix. Key results, in terms of parameters to which predictions were found to be most sensitive, are summarised as follows.

5.2.1 Prediction metric 1: Maximum drawdown

The prediction of maximum drawdown was again found to be most sensitive to the horizontal *K* of the Namoi Alluvium aquifer (K_H 01), which was consistently ranked as the most influential parameter with increasing sample size (Figure 5-4). The K_H 01 parameter prediction was also showing significantly larger values for the δ and S_1 parameter (δ =0.268 and S_1 =0.504) compared to all other model parameters (Figure 5-5). For interpretation of this result with regards to the structure of the groundwater flow model, see Section 3.5.1. Maximum drawdown was again found to be sensitive to the vertical *K* of the upper aquitard sequence, K_V 05 (Figure 5-4), with magnitudes δ =0.152, S_1 =0.119) (Figure 5-5). This prediction was also found to be sensitive to the horizontal *K* of the Blythesdale Group aquifer (K_H 03).



Figure 5-4. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the magnitude of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001). Sample sizes used as the basis for metric computation ranged from 200 to 300 model runs.

The uncertainty associated with this prediction did not vary significantly with increasing sample size (Figure 5-3). Predicted values ranged from 0.0 m to 4.2 m. The 10th percentile value ranged from 0.1 m to 0.1 m; the median value was consistently 1.1 m; and the 90th percentile value ranged from 2.5 m to 2.6 m.



Figure 5-5 Global sensitivity analysis metrics of 20 model parameters in relation to maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation was 300 model runs.

5.2.2 Prediction metric 2: Time of maximum drawdown

The prediction of time of maximum drawdown was found to be most sensitive to the vertical *K* of the upper aquitard sequence, K_V 05 (Figure 5-6a). Sobol' analyses also indicated that that prediction was sensitive to the horizontal *K* of the Namoi Alluvium aquifer K_H 01 (Figure 5-6b). These two model parameters have sensitivity metrics that are significantly larger than that of all the other parameters, i.e. δ =0.199 and S_1 =0.240 for K_V 05 and δ =0.121 and S_1 =0.090 for K_H 01 (Figure 5-7).



Figure 5-6. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the timing of maximum drawdown in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001).

The uncertainty of this prediction did not vary significantly with increasing sample size (Figure 5-3). Predicted values ranged from 19 years to 261 years. The 10th percentile value ranged from 99 years to 114 years; the median value was consistently 255 years; and the 90th percentile value was consistently 261 years. When interpreting these results, and as stated previously, for a subset of model runs the model temporal extent specified was insufficient to capture peak drawdown conditions. Due to large model run times it was not feasible to repeat the sampling process using an extended model duration within the duration of the present study.



Figure 5-7 Global sensitivity analysis metrics of 20 model parameters in relation to time to maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation is 300 model runs.

5.2.3 Prediction metric 3: Spatial extent of drawdown propagation

The prediction of drawdown propagation extent was found to be most sensitive to the horizontal K of the Namoi Alluvium aquifer, K_H 01, which was consistently ranked as the most influential parameter with increasing sample size (Figure 5-8). For interpretation of this result with regards to the structure of the groundwater flow model, see Section 3.5.1. This prediction was also found to be sensitive to the horizontal K of the Blythesdale Group aquitard K_H 03 (Figure 5-8b). Figure 5-9 confirms that model parameter K_H 01 is the only significant model parameter with δ =0.287 and S_1 =0.644; although Figure 5-8 suggests that parameter K_H 03 is also a sensitive parameter, this is not confirmed when considering the magnitude of the sensitivity metrics (δ =0.100, S_1 =0.049).

The uncertainty of this metric did not vary significantly with increasing sample size (Figure 5-3). Predicted values ranged from zero to 100 cells. The 10th percentile and median values were consistently zeros cells and the 90th percentile value was consistently 64 cells.



Figure 5-8. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of drawdown spatial extent in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001).



Figure 5-9 Global sensitivity analysis metrics of 20 model parameters in relation to spatial extent of maximum drawdown. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation is 300 model runs.

5.2.4 Prediction metric 4: Maximum vertical flux

The prediction of maximum flux was found to be most sensitive to the vertical *K* of the upper aquitard sequence K_V 05 (Figure 5-10). Other model parameters such as the vertical *K* of the Pilliga Sandstone aquifer, K_V 04, and the horizontal *K* of the Namoi Alluvium aquifer K_H 01 are not significant (Figure 5-10). Indeed, only K_V 05 shows sensitivity metrics significantly larger than all the other parameters, i.e. δ =0.414 and S_1 =0.651 (Figure 5-11). For parameters K_V 04 and K_H 01, the sensitivity metrics are not significant (i.e. δ =0.112 and S_1 =0.033 for K_V 04 and δ =0.108 and S_1 =0.066 for K_H 01).



Figure 5-10. Relative rankings of ten parameters used by the revised Gunnedah-Surat basins groundwater flow model (most sensitive = red, second most sensitive = blue, third most sensitive = green; less sensitive parameters are shown in grey) to which the prediction of the maximum change in vertical flux in the Pilliga Sandstone aquifer was most sensitive. Parameter rankings were based on two global sensitivity analysis metrics: (a) the Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) the first order Sobol' metric (S_1 ; Sobol', 2001).

The uncertainty of this prediction did not vary significantly with increasing sample size (Figure 5-3). Predicted values ranged from 18 m³/d to 38 m³/d. The 10th percentile value was consistently 23 m³/d; the median value ranged from 25 m³/d to 26 m³/d; and the 90th percentile value ranged from 33 m³/d to 34 m³/d.



Figure 5-11 Global sensitivity analysis metrics of 20 model parameters in relation to maximum vertical flux. (a) Delta Moment-Independent Measure (δ ; Plischke et al., 2013) and (b) first order Sobol' metric (S_1 ; Sobol', 2001). Sample size for metric computation is 300 model runs.

5.3 Prediction uncertainty analysis – Results and Discussion

As discussed in Section 3.6, the uncertainty of predictions generated using the CDM Smith (2016) Gunnedah-Surat basins model is described using summary statistics as follows. Differences between maximum and minimum values do not provide robust estimates of prediction uncertainty, as these are highly sensitive to the presence of outlying values. Instead, prediction uncertainty is quantified here using three statistics: (1) standard deviation, which is commonly used as a measure of spread in linearised prediction uncertainty analyses (though typically implemented as variance); (2) the interquartile range (i.e., the difference between the 25th and 75th percentiles), which is most robust to outliers; and 3) the 90 % confidence interval (i.e., the difference between the 5th and 95th percentiles).

The calculated prediction uncertainties were compared between the first (initially uninformed) and second (data-driven) parameterisation. The approach did not constrain parameters by observations of state variables such as hydraulic heads. Rather, the stochastic approach followed the Bayesian paradigm, in which prior beliefs and estimates of parameters and conceptualisation are iteratively updated when new data and information becomes available. Note that head observations for the considered depressurisation scenario are currently not available; an attempt to reduce uncertainty by constraining model simulations to heads beyond the initially calibrated model is not an option. In other words, the approach adopted avoids possible introduction of bias in the predictions due to calibration.

5.3.1 Prediction 1: Magnitude of maximum drawdown

Predictions of the magnitude of maximum drawdown ranged from < 1.0 m to 17.8 m, with a median value of 1.2 m (Figure 5-12). This reflected an increase in the median predicted value by 0.5 m. Both the standard deviation and 90 % confidence interval of predicted values remained unchanged. The interquartile range of predicted values reduced by 0.9 m to 3.9 m. As mentioned previously (Section 3.5.1), it should be noted that the numerical convergence criterion with regards to hydraulic head (i.e., 0.01 m) should be considered as an acceptance threshold for this metric. The magnitude of maximum drawdown was calculated as \leq 0.1 m in 68 of the 300 model runs.





5.3.2 Prediction 2: Timing of maximum drawdown

Predictions of the timing of maximum drawdown ranged from < 1 years to 160 years, with a median value of 155 years (Figure 5-13). This reflected an increase in the median predicted value by a single year. Both the standard deviation and 90 % confidence interval of predicted values remained unchanged. The interquartile range of predicted values increased by 2 years to 48 years. As described previously (Section 3.5.2), it should be noted that the maximum simulation time specified for the model (i.e., 160 years after the cessation of CSG extraction) significantly affected this prediction type. Over one third of all model runs (i.e., 107 runs) resulted in a time of maximum drawdown of 160 years. Such results indicate that the total simulation time of these models was not sufficiently large to capture the peaks of drawdown responses. Due to large model run times it was not feasible to repeat the sampling process using an extended model duration within the duration of the present study. This factor is also the cause of the non-asymptotic behaviour of the cumulative density function for this prediction (Figure 5-13a).



Figure 5-13. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the timing of maximum drawdown, based on a sample size of 300 model runs.

5.3.3 Prediction 3: Spatial extent of drawdown propagation

Predictions of the spatial extent of drawdown propagation ranged from zero cells to 9 990 cells, with a median value of zero cells (Figure 5-14). This reflected no change in the median predicted value. The standard deviation of predicted values increased by 275 cells to 2 104 cells. The interquartile range of predicted values increased by 1 376 cells to 2 872 cells. The 90 % confidence interval increased by 52 cells to 400 cells. Note that the cumulative density function presented here for the 'revised' case (red) is a truncated distribution, in which approximately 60 % of the model runs featured 'zero impact' (i.e., for which the maximum reported drawdown for Pilliga Sandstone aquifer cells was less than 2 m).

Spatial maps with isolines of drawdowns are shown in Figure 5-15 for the 5th, 50th, 90th, 95th, and 99th percentiles, based on 300 calculated values of drawdown for each model cell. The time at which the drawdown was reported corresponds to the median time to maximum drawdown across all 300 model runs. Isolines shown are for 1 and 2 m drawdown only (Figure 3-15), where drawdown values are taken at 155 years following cessation of gas extraction. For the 5th and 50th percentile there are no areas with a drawdown equal to or larger than 1 m. In other words, the expected (50th percentile) value shows that not a single area has a drawdown equal to or in excess of 1 m. Compared to the initial drawdowns (Figure 3-15), somewhat larger areas of drawdown larger than 1 and 2 m are observed for the 90th and 95th percentiles, while the 99th percentiles are similar. At the 95th percentile, the overall maximum drawdown across all 300 model runs is 7.9 m.



Figure 5-14. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the spatial extent of drawdown in excess of 2 m, based on a sample size of 300 model runs.



Figure 5-15. Percentiles of spatial distributions of 1 and 2 m drawdown (solid black lines), observed at the median time of maximum drawdown (i.e., 155 years after the cessation of coal seam gas extraction); purple = Pilliga Sandstone aquifer cells; blue = upper aquitard sequence cells. Drawdowns < 1 m are not shown. Percentiles shown are (a) 5th, (b) 50th, (c) 90th, (d) 95th and (e) 99th. Distributions are based on a set of 300 model runs using the 'revised' model.

5.3.4 Prediction 4: Maximum vertical flux

Predictions of the maximum vertical flux ranged from 306 m³/d to 1432 m³/d, with a median value of 636 m³/d (Figure 5-16). This reflected a reduction in the median predicted value by 101 m³/d. The standard deviation of predicted values increased by 1 m³/d to 242 m³/d. The 90 % confidence interval increased by 0.2 m³/d to 46 m³/d. The interquartile range of predicted values reduced by 76 m³/d to 288 m³/d.



Figure 5-16. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of changes in vertical groundwater fluxes, based on a sample size of 300 model runs.

These results indicate that uncertainties associated with three of the four prediction metrics (i.e., maximum drawdown, maximum extent of drawdown ≥ 2 m, and maximum vertical flux) were slightly reduced after the improved characterisation aquitard vertical hydraulic conductivity. Characterisation was undertaken by upscaling porosity–permeability relationships identified from core samples through combination with wireline log data. The median prediction of maximum drawdown in the Pilliga Sandstone aquifer increased by 0.4 m. Similarly, the number of model cells affected by drawdown ≥ 2 m increased by a single cell. Conversely, the median prediction of maximum vertical flux between the Pilliga Sandstone aquifer and underlying units reduced by 101 m³/d. The median prediction of the time at which maximum drawdown will occur did not reduce significantly. This may be attributed to the total temporal extent represented by the model (i.e., 186 years) which, in many cases, was not sufficiently large to capture the time of peak induced drawdown. Due to large model run times it was not feasible to repeat the sampling process using an extended model duration within the duration of the present study.

Table 5-1. Median values of four prediction metrics based on sets of 300 model runs for each of the 'initial' and 'revised' models.

Prediction	'Initial' model	'Revised' model	Difference
 Maximum drawdown (MXD)	0.7 m	1.2 m	+0.5 m
Timing of maximum drawdown (tMXD)	154 y	155 y	+1 γ
Spatial extent of drawdown $\ge 2 \text{ m}$ (NDD)	0 cells	0 cells	0 cells
Maximum vertical flux (MXQ)	737 m³/d	636 m³/d	-101 m³/d

The 95th percentile values for each prediction before and after the inclusion of aquitard K_V data (i.e., the 'initial' and 'revised' models, respectively) are summarised in Table 5-2. The increase in maximum drawdown by 0.7 m and number of cells exhibiting a drawdown > 2 m is consistent with the larger range in K_V values used to run the 300 models (two orders of magnitude for the initial

data set to four orders of magnitude for the updated set). The maximum flux, on the other hand, decreases by 47 m³/d. This indicates that the larger range in K_V values resulted in a more extreme (i.e. larger) K_V values that generated a larger drawdown area (e.g. the area with > 2 m drawdown increased by 861 cells) and a larger maximum drawdown. Because of this, the maximum flux decreases slightly as water is now flowing across a larger area towards the aquitard, or in other words, the total flux through the aquitard increases which slightly decreases the maximum flux at one particular cell.

Table 5-2. 95th percentiles of four prediction metrics based on sets of 300 model runs for each of the 'initial' and 'revised' models.

Prediction	'Initial' model	'Revised' model	Difference
Maximum drawdown (MXD)	7.2 m	7.9 m	+0.7 m
Timing of maximum drawdown (tMXD)	160 y	160 у	0 у
Spatial extent of drawdown $\ge 2 \text{ m}$ (NDD)	4 490 cells	5 350 cells	+861 cells
Maximum vertical flux (MXQ)	1 310 m³/d	1 260 m³/d	-47 m³/d

6 Analyses of the spatial variability of aquitard hydraulic properties

The original CDM Smith (2016) Gunnedah-Surat basins model parameterised aquitard vertical hydraulic conductivities using spatially uniform (i.e., homogeneous) values. The characterisation of aquitard K_V described in Section 4 resulted in upscaled estimates at numerous locations distributed across the Gunnedah-Surat basins. Such observations allow for the characterisation of the spatial variability of aquitard K_V using geostatistical methods. This was undertaken in the present study as a basis for the stochastic generation of spatially distributed aquitard K_V parameter sets. These were then used as the basis for a non-comprehensive assessment of the effects of the heterogeneity of aquitard K_V on modelled hydraulic predictions. For this approach, the spatial variability of K_V was incorporated explicitly when quantifying uncertainty associated with predictions.

6.1 Spatial distributions of aquitard vertical hydraulic conductivity

Spatial distributions of upscaled numerical and harmonic (as well as arithmetic and geometric) mean values for the upper aquitard sequence are shown in Figure 6-1. The data set contains 54 data points for the numerical and 64 data points for the harmonic mean. Journel and Hujbregts (1978) advise practitioners that a minimum of 30 to 50 paired comparisons (or number of lag differences, see further) should suffice to construct a semi-variogram (see further in Section 6.2). In our case the number of data pairs for the harmonic mean K with 64 data points (n) equals $n \times (n - 1)$ 1)/2= 64×63/2=2016. The condition of minimum paired comparisons is thus more than satisfied, and the semi-variogram can be relatively accurately described. Of notice is the higher density of data points across the focus study area (arrow in Figure 6-1 and Figure 6-2), which will result in a relatively accurate description of the spatial structure or spatial dependency in K_V. Note that the spatial dependency is only calculated in two dimensions (i.e., in the horizontal direction), as any vertical variability in K_V is represented by a single upscaled value. Outside of the study area, very few data points exist which will result in a much less accurate description of spatial heterogeneity. This is acceptable, as the major impacts will be centred on the study area, and much less so on the broader model domain (see Section 5.3.3 for details). In case the regional impact exceeds the area that is most accurately characterised, additional characterisation across the broader domain can potentially be considered. Upscaled K_V values are very similar for the two averaging methods.

Spatial distributions of upscaled numerical and harmonic (as well as arithmetic and geometric) mean values for the lower aquitard sequence are shown in Figure 6-2. For the lower aquitard the number of data points are 69 for the numerical and 78 for the harmonic mean. The same observations are made as for the lower aquitard: higher data density across the study area, with good consistency in upscaled K_V for the two averaging methods.

Note that the reason for different number of wells between upper and lower aquitard sequences was discussed in Section 4.2.2. For the analytical upscaling, wells needed to include at least one of the contributing aquitards. An additional requirement for the numerical upscaling was that the

contributing aquitards needed to have hydraulic conductivity data for interburden layers that would connect the two aquitards within a single sequence.



Figure 6-1. Spatial distributions of upscaled (a) harmonic and (b) numerical mean values for the upper aquitards (Jurassic age). Grey shaded area is extent of aquitard. Arrow indicates focus study area with largest data density.



Figure 6-2. Spatial distributions of upscaled (a) harmonic and (b) numerical mean values for the lower aquitards (Permian age). Grey shaded area is extent of aquitard. Arrow indicates focus study area with largest data density.

6.2 Variogram analyses

The spatial variability of a given property can be characterised by the variance between a set of spatially distributed observations, which is expressed as a function of the distance (i.e., lag) between observation locations. The variance between observations increases with distance until reaching a maximum value (i.e., sill) at a given maximum distance (i.e., range). In some cases, a non-zero variance will apply at a separation distance of zero; this can be quantified using a nugget parameter. Increases in spatial variability with respect to increasing distance are expressed using a semi-variogram. Closed-form parametric models (e.g., spherical, exponential, Gaussian) can be fitted to semi-variogram data and subsequently used to interpolate values away from observation locations. Such models can also be used as the basis for the stochastic generation of property fields. The approaches to the characterisation of spatial variability described here are commonly known as two-point geostatistics (Journel and Huijbregts, 1978).

Differences between upscaled aquitard K_V values and their corresponding locations of origin (X, Y in the horizontal plane) were compared using a two-point geostatistical approach, described as follows. First, squared differences between each pair of values were calculated. All pairs featuring zero difference in either value or location were then excluded. The remaining datasets (all of which contained at least 2 900 paired differences) are shown as small grey closed circles in Figure 6-3a (upper aquitard sequence) and Figure 6-3b (lower aquitard sequence). Distances between pairs ranged from zero to 250 km. Differences between paired values (i.e., semi-variances) ranged from zero $(m/d)^2$ to 15 $(m/d)^2$. In all cases, small differences between values (i.e., semi-variance values $\leq 10 [m/d]^2$) are present for separation distances up to 60 km.

The two data sets (one for each aquitard sequence) were each divided into an arbitrary number of bins, each of which contained 100 sample pairs. Experimental or empirical semi-variograms, $\hat{\gamma}(h)$, (EVs; closed red circles and red lines) were then derived through the calculation of (arithmetic) mean differences in value and location for each data bin (Cressie, 1985; Kitanidis, 1997):

$$\hat{\gamma}(h) = \frac{1}{2 n(h)} \sum_{i=1}^{n(h)} [z(x_i + h) - z(x_i)]^2$$
(7)

where z is a data value at a particular location, h is the lag-distance between ordered data, and n(h) is the number of paired data at a distance of h. The semi-variance is half the variance of the increments $z(x_i+h)-z(x_i)$. For a data set with m observations, there are m(m-1)/2 unique pairs of data with a lag-distance h.

Both EVs appear to approach asymptotic conditions at separation distances of between 100 km and 200 km, with the exception of the final EV data bin for the upper aquitard sequence (harmonic mean approach). Also shown in Figure 6-3 are correlograms (blue filled circles and lines), which represent the reduction in autocorrelation between binned data values with increasing lag distance. Autocorrelation values (*A*) were calculated as:

$$A(h) = 1 - \hat{\gamma}(h) / \sigma_{\gamma(h)}^2 \tag{8}$$

Autocorrelation measures similarity of a property or parameter between sites and is often modelled as an exponential decay with distance.



Figure 6-3. Experimental variograms (red, based on semi-variance values) and correlograms (blue, based on autocorrelation values) for the (a) upper aquitard sequence and (b) lower aquitard sequence, using vertical hydraulic conductivity values that were upscaled using numerical model averaging.

Two parametric models were used to characterise the EVs: a spherical model and an exponential model, both of which were parameterised using range, sill and nugget values. The range is the lag distance at which the semi-variogram reaches the sill value; the autocorrelation is essentially zero beyond the range. The sill is the semi-variance value at which the variogram levels off. The nugget value represents variability at distances smaller than the typical sample spacing, including measurement error. The spherical model is given as (Cressie, 1985; Kitanidis, 1997):

$$\gamma(h) = c \left[\frac{3h}{2a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right] + n \quad , h < a$$

$$\gamma(h) = c + n \qquad , h \ge a$$
(9)

while the exponential model is given as (Cressie, 1985; Kitanidis, 1997):

$$\gamma(h) = c [1 - e^{-h/a}] + n$$
(10)

where γ is semi-variance, *h* is lag distance, *c* is sill, *a* is range, and *n* is the nugget. The spherical model reaches the specified sill value, *c*, at the specified range, *a*. Conversely, the exponential model approaches the sill asymptotically, where *a* represents the "practical range"; i.e., the distance at which the semi-variance reaches 95 % of the sill value (Figure 6-4).

The spherical and exponential models were, in part, selected as they are suitable for use in subsequent Sequential Gaussian Simulation of spatial parameter distributions. The models were fitted to EVs through the use of the Python language least squares algorithm *curvefit*. Parameters of the fitted models and associated coefficient of determination (R^2) values are summarised in Table 6-1.

Both variogram models fitted the experimental variogram equally well up to a lag distance of approximately 100 km (Figure 6-5). For larger distances, the spherical model reached a constant value (i.e., the sill) and remained close to the experimental variogram. In comparison, exponential model values continued to increase beyond this lag distance and approached the sill asymptotically. These differences between the two variogram models explained the consistently

larger sill and range values for the exponential model. In addition, the nugget values of both variogram models are very similar since both models fitted the exponential variogram equally well at short (i.e., <100 km) lag distances.



Figure 6-4 Theoretical variogram models (spherical, exponential, Gaussian) with indication of the practical range (Bohling, 2005).



Figure 6-5. Geostatistical analysis of (a) upper aquitard sequence and (b) lower aquitard sequence K_V values upscaled using a 1-D numerical groundwater flow model. Grey closed circles = data points $[z(x_i+h)-z(x_i)]^2$, red closed circles and lines = experimental variograms, solid green line = spherical variogram model, and dashed blue line = exponential variogram model.

Close inspection of the experimental variogram in Figure 6-5 for the upper aquitard sequence shows signs of short-range variability up to about 30 km lag distance in combination with large-range variability for lag distance up to 100 km. These double-hump or multiple-hump features have previously been discussed by Ababou et al. (1989), Gelhar (1993, 1989) and others, and have been related to spatial variability at different scales. Indeed, one can envisage scales of heterogeneity (Figure 6-6): i) at the regional scale due to different geological formations (scale \mathcal{L}), ii) at the scale of an individual geological formation due to variation in facies (scale of the flow domain, L), and iii) at the scale of an individual facies due to differences in cementation of pores (scale λ) (Weber, 1986). Typically, the regional scale \mathcal{L} could be on the order of 50 -100 km, the scale of the flow domain L could be on the order 1 – 10 km, and the scale of fluctuation λ on the order of 10 – 100 m. As shown in Figure 6-6, a composite semi-variogram can be assumed which describes the different spatial scales. The increasingly higher sill values are consistent with the higher variability in ln K_V that is expected when increasingly larger spatial domains (i.e. greater variability in geological strata) are sampled.



Figure 6-6 Multiple scales of heterogeneity. (top) Identification of scales: regional scale (\mathcal{L}), scale of the flow domain (L), and local-scale correlation scale (λ). (bottom) Hypothetical multi-scale semi-variogram corresponding with multiple scales of heterogeneity (modified from Ababou et al. [1989]).

The question of how many heterogeneous random fields suffice in a stochastic analysis has been addressed by Ababou (1988) and Ababou et al. (1989). These authors used, as an extreme case, a single realization and developed some guidance in regards to minimum acceptable size of the model domain and maximum grid size, for the model to be statistically meaningful. When a single realization approach is used, inference of statistical moments of the distribution of relevant variables requires a flow domain that is sufficiently large compared with the correlation scales of the pertinent formation properties. In addition, in order to preserve details of the spatial structure of the formation properties, the size of the numerical cells must be small compared with the characteristic length scale of the heterogeneity of the relevant formation properties. Ababou (1988) and Ababou et al. (1989) suggested the following two criteria: i) the domain size should be at least 10 to 50 times the correlation length, and ii) at least four nodes or grid cells per correlation length.

Table 6-1. Variogram model parameter (i.e., range, sill and nugget) values and associated coefficient of determination (R^2) values for power and logarithmic variogram models used to characterise the spatial correlation between upscaled K_V values for the upper and lower aquitard sequences. *Note that "Practical range" refers to the distance at which semi-variance values generated using an exponential variogram model reach 95 % of the sill value.

Aquitard sequence	Model type	Upscaled data type	Range	Practical range*	Sill	Nugget	R ²
			(km)	(km)	(m/d)²	(m/d) ²	(-)
Upper	spherical	harmonic	129	N/A	0.764	0.327	0.964
		numerical	100	N/A	0.987	0.269	0.964
	exponential	harmonic	298	283	2.570	0.337	0.964
		numerical	133	126	2.000	0.290	0.960
Lower	spherical	harmonic	319	N/A	1.320	0.133	0.947
		numerical	128	N/A	0.811	0.175	0.985
	exponential	harmonic	43 400	41 300	262	0.136	0.947
		numerical	255	242	2.420	0.183	0.984

A re-analysis of the experimental variogram for both the upper and lower aquitard sequences was undertaken in which data points featuring lag distances in excess of 30 km were excluded. In this way the local-scale correlation length was determined which included mainly data from the focus study area with the highest data density (see Figure 6-1). Based on the numerically upscaled data, the spherical semi-variogram model revealed a correlation length of 20 km for the upper aquitard and 30 km for the lower aquitard. With a model domain that is approximately 230×230 km², the first criterion of at least 25 correlation lengths within the modelling domain is not entirely satisfied. Nevertheless, with about half the required number of correlation lengths included in the flow model, the number of realizations required to derive meaningful statistics is not expected to be excessive. Therefore, in the subsequent section, a total of 50 realizations has been generated for subsequent flow modelling. As to the second criterion, there should be four nodes every 20 km (i.e. correlation length). For the minimum grid size of 1×1 km² there were 20 nodes per correlation

length, thus satisfying the second criterion. The criterion is also satisfied for the largest grid size, i.e., 5×5 km².

6.3 Stochastic generation of spatially distributed parameter fields

Geostatistical characterisation of the spatial variability of the upper and lower aquitard sequences, as described above, enabled the stochastic generation of spatial distributions of vertical hydraulic conductivity. In comparison to the generation of uniform parameter fields, these heterogeneous fields provided a means of investigating the effects of spatial parameter variability on the four predictions of interest. Heterogeneous parameter fields were generated according to the relevant spherical variogram model and were conditioned to the relevant observations of vertical hydraulic conductivity. A conditional Sequential Gaussian Simulation algorithm, which used ordinary kriging to interpolate between populated cell values, was used to generate multiple and equally probable parameter fields or realisations. The conditional simulation approach takes into account the spatial variation of actual data at sampled locations, while the variation of estimates at unsampled locations is interpolated using kriging. Conditional stochastic simulation reproduces the statistics of a given sample set (e.g., the histogram of values and the specified model of spatial correlation, such as a semi-variogram model) while, for cases in which the nugget value is zero, honouring data values at their sampled locations. In cases where the nugget value is non-zero, values at sampled locations will vary according to the degree of inherent variability described by the nugget effect. The software utility FIELDGEN (WNC, 2016) was used to undertake spatial field generation.

Two example aquitard K_V spatial distributions are presented in Figure 6-7. Note that only lateral spatial heterogeneity is represented for each aquitard sequence: the vertical heterogeneity is replaced by the upscaled K_V value. It should be noted that individual stochastic realisations may differ significantly from one another, even while the underlying statistical correlation structure remains the same. Statistics of the equivalent aquitard K_V values of the 50 spatial distributions generated for the upper and lower aquitard sequences are summarised in Table 6-2.

Aquitard sequence	min	10 th	25 th	50 th	mean	75 th	90 th	max
Upper	-6.0	-5.5	-5.4	-5.3	-5.3	-5.1	-5.0	-4.7
Lower	-5.6	-5.2	-5.1	-5.0	-5.0	-4.9	-4.8	-4.5

Table 6-2. Summary statistics of 50 spatial distributions of equivalent log₁₀ aquitard vertical hydraulic conductivity (m/d) generated for the upper and lower aquitard sequences using Sequential Gaussian Simulation.



Figure 6-7. Heterogeneous spatial distributions of log_{10} vertical hydraulic conductivity ($log_{10} K_V$) for the (a) upper and (b) lower aquitard sequences generated using Sequential Gaussian Simulation based on spherical variogram models and conditioned to values upscaled using numerical model averaging (red = low values; blue = high values; white = non-Pilliga Sandstone aquifer cells).

6.4 Prediction uncertainty analysis using spatially distributed parameter fields

Subsequent sections describe analyses of the effects of incorporating a spatially variable parameterisation $\log_{10} K_V$ on the uncertainty of predictions. Such a conceptualisation is more physically realistic than the spatially uniform conceptualisation assumed previously. However, in practice, this approach requires a higher level of data support. Furthermore, predictions of interest may or may not be sensitive to spatially variable hydraulic properties; indeed, boundary conditions may, in some circumstances, be the primary control on predictions. The uncertainty quantification presented here was focused solely on the influence of heterogeneity in aquitard K_V values on the four prediction metrics discussed previously.

6.4.1 Prediction metric 1: Magnitude of maximum drawdown

Predictions of the magnitude of maximum drawdown ranged from 1.3 m to 7.2 m, with a median value of 3.4 m. This reflected an increase in the median predicted value by 2.7 m. The standard deviation of predicted values reduced by 1.4 m to 1.6 m. The interquartile range of predicted values reduced by 0.2 m to 2.8 m. The 90 % confidence interval increased by 0.1 m to 0.7 m.



Figure 6-8. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the magnitude of maximum drawdown, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous').

6.4.2 Prediction metric 2: Timing of maximum drawdown

Predictions of the timing of maximum drawdown ranged from 11 to 137 years, with a median value of 35 years (Figure 6-9). This reflected a reduction of the median predicted value by 119 years. The standard deviation of predicted values reduced by 17 to 25 years. The 90 % confidence interval increased by 3 to 11 years. The interquartile range of predicted values reduced by 15 years to 32 years. The 10–90 percentile range reduced by 43 to 54 years.



Figure 6-9. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the timing of maximum drawdown, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous').

These results are unlikely due to the smaller number of model runs (i.e., 50 compared to 300), based on the condition that the scale of the model is larger than the correlation scales of the

pertinent formation properties (i.e. scale of the small-scale heterogeneity, Section 6.3). As illustrated in Figure 6-10, the heterogeneous model results in much better constrained model outputs due to Sequential Gaussian Simulation used to generate, and constrain, the heterogeneous K_V fields. Indeed, the time of maximum drawdown is constrained to a relatively narrow band between 11 and 80 years, with only one exception at 137 years. Time of maximum drawdown for the homogeneous model is not constrained, with values covering the full modelled extent of 160 years, and some that have not yet reached their maximum drawdown.





Of further interest is the decreasing trend in drawdown with increasing time to maximum drawdown. This trend extends beyond the data points for the heterogeneous case, and links into at least a subset of the homogeneous data points. In other words, the bulk of the data behaves as expected, where high K_V values would allow for a higher drawdown which is obtained relatively sooner due to quicker reaction time of the system.

6.4.3 Prediction metric 3: Spatial extent of drawdown propagation

Predictions of the spatial extent of drawdown propagation ranged from zero cells to 726 cells, with a median value of 63 cells. This reflected an increase in the median predicted value by 63 cells. The standard deviation of predicted values reduced by 1 636 cells to 194 cells. The interquartile range of predicted values reduced by 1 217 cells to 279 cells. The 90 % confidence interval reduced by 257 cells to 90 cells.



Figure 6-11. (a) Cumulative density function and (b) frequency histogram of initial (black) and revised (red) predictions of the spatial extent of drawdown in excess of 2 m, based on sample sizes of 300 model runs ('Initial') and 50 model runs ('Heterogeneous').

Figure 6-12 shows spatial maps with isolines of drawdowns for the 5th, 50th, 90th, 95th, and 99th percentiles, based on 50 calculated values of drawdown for each model cell. Isolines shown are for 1 and 2 m drawdown (Figure 6-12), where drawdown values are taken at 155 years following cessation of gas extraction. For the 5th and 50th percentile there are no areas with a drawdown equal to or larger than 1 m. In other words, the expected (50th percentile) value shows that not a single area has a drawdown equal to or in excess of 1 m. Compared to the initial model drawdowns (Figure 3-15) and revised model drawdowns (Figure 5-15), considerably smaller areas of drawdown larger than 1 and 2 m are observed for all percentiles. At the 95th percentile, the overall maximum drawdown across all 50 model runs is 6.1 m.



Figure 6-12. Percentiles of spatial distributions of 1 and 2 m drawdown (solid black lines), observed at the median time of maximum drawdown (i.e., 155 years after the cessation of coal seam gas extraction); purple = Pilliga Sandstone aquifer cells; blue = upper aquitard sequence cells. Drawdowns < 1 m are not shown. Percentiles shown are (a) 5th, (b) 50th, (c) 90th, (d) 95th and (e) 99th. Distributions are based on a set of 50 model runs that featured spatially variable parameterisation of aquitard vertical hydraulic conductivity values (i.e., the 'heterogeneous' model).

6.4.4 Prediction metric 4: Maximum vertical flux

Predictions of the maximum change in vertical flux ranged from 548 m³/d to 1124 m³/d, with a median value of 751 m³/d. This reflected an increase in the median predicted value by 14 m³/d. The standard deviation of predicted values reduced by 50 m³/d to 191 m³/d. The interquartile range of predicted values reduced by 14 m³/d to 350 m³/d. The 90 % confidence interval increased by 43 m³/d to 89 m³/d.

A comparison of the 50th percentiles (or expected values) of four prediction metrics is available from Table 6-3; the expected value of maximum drawdown increases from 0.7 m for the initial model to 1.2 m for the revised model and 3.4 m for the heterogeneous model.

Table 6-3. 50th percentile (i.e., median) values of four prediction metrics based on sets of 300 model runs for each of the 'initial' and 'revised' models and 50 model runs with the 'heterogeneous' models.

Prediction	'Initial' model	'Revised' model	'Heterogeneous' model
Maximum drawdown (MXD)	0.7 m	1.2 m	3.4 m
Timing of maximum drawdown (tMXD)	154 y	155 у	35 у
Spatial extent of drawdown $\ge 2 \text{ m}$ (NDD)	0 cells	0 cells	63 cells
Maximum vertical flux (MXQ)	737 m³/d	636 m³/d	751 m³/d

The timing of the maximum drawdown occurs about five times earlier for the heterogeneous model compared to the homogeneous initial and revised models. The spatial extent of the drawdown > 2 m is 63 cells for the heterogeneous model compared to zero cells for the other two models. Finally, the maximum vertical flux is slightly higher for the heterogeneous model compared to the homogenous models. The result for the third metric is strongly influenced by the very truncated distributions of the initial and revised model, with nearly 60% of the model runs producing zero cells with a drawdown > 2 m.





A further comparison across the four prediction metrics is available for the 95th percentile (Table 6-4). Here, the heterogeneous model generates the lowest maximum drawdown (6.1 m versus 7.9 for the revised model), has a spatial extent of drawdown > 2 m that is one order of magnitude smaller than that of the two other models, and a considerably smaller maximum flux. These results are consistent with the spatial distribution of drawdown shown in Figure 6-12, illustrating that the heterogeneous model seems to be much better constrained resulting in a much smaller range of drawdowns. The better constrained K_V model results from using the conditioned Sequential Gaussian Simulation, which honours the observed K_V data. A second reason for the smaller range of drawdowns is the smaller number of model runs, 50 for the heterogeneous

model versus 300 for the homogeneous models. Therefore, it is likely that a smaller K_V parameter space is sampled in case of the heterogeneous model.

Table 6-4. 95th percentiles of four prediction metrics based on sets of 300 model runs for each of the 'initial' and 'revised' models and 50 model runs with the 'heterogeneous' models.

Prediction	'Initial' model	'Revised' model	'Heterogeneous' model
Maximum drawdown (MXD)	7.2 m	7.9 m	6.1 m
Timing of maximum drawdown (tMXD)	160 y	160 y	81 y
Spatial extent of drawdown $\ge 2 \text{ m}$ (NDD)	4 490 cells	5 350 cells	551 cells
Maximum vertical flux (MXQ)	1 310 m³/d	1 260 m³/d	1 074 m³/d

6.5 Prediction uncertainty and groundwater management

Groundwater impact assessments are increasingly being undertaken in a probabilistic framework whereby all sources of uncertainty (model parameters, model structure, boundary conditions, and calibration data) are taken into account (Vrugt et al., 2013; Guillaume et al., 2016). This has resulted in groundwater impact metrics being presented as probability density functions and/or cumulative distribution functions, spatial maps displaying isolines of percentile values for specific metrics, etc.

Groundwater management on the other hand typically uses single values (i.e., in a deterministic framework) to evaluate what decisions are required to protect groundwater resources. For instance, the third groundwater impact metric applied in this study considers the number of cells in the Pilliga Sandstone aquifer with drawdown greater than or equal to two meters. This nominal drawdown value of two metres was selected as this is consistent with trigger-level thresholds specified by the NSW Aquifer Interference Policy (NSW DPI OW, 2012). In many cases, when drawdowns induced by groundwater extraction exceed two metres, "make-good" provisions are enacted (such as the surrendering of extraction licences).

The information obtained from a quantitative uncertainty analysis (such as presented here) can be used to guide decision making in several ways. Two examples are discussed here: the first of which would not require modification of existing "deterministic" trigger or guideline values, whereas the second example assumes that the regulatory criteria are also expressed in probabilistic terms.

The first example is a straightforward interpretation of calculated percentile values for specific impact metrics. Consider a 5% lower confidence limit (the magnitude of the impact metric would typically be at the lower end of all possible impacts) that is above the regulatory standard of concern, then it is likely that the standard will be violated. Indeed, in this case 95% of all calculated impacts would exceed the standard: the decision that the standard is violated has a high degree of certainty. If, on the other hand, the 95% upper confidence limit (the magnitude of impact is at the upper end of all possible impacts) is below the standard, it is likely that the standard will not be violated. In this case there is only a very small chance (5%) for the metric to exceed the standard: the decision to accept the impact has a very small margin of error. If the 95% upper confidence limit exceeds the standard, but the 50th percentile is less than the standard, further study could be recommended for those parameters that most control the overall uncertainty. However, if the 50th percentile exceeds the standard, further study may still be recommended. Under some circumstances one may choose to proceed with regulatory action depending on the cost-effectiveness of measures for risk reduction.

The second examples goes a step further, as the previous deterministic thresholds do not currently allow for a probabilistic interpretation; e.g., there is no statement that "the probability of exceeding the threshold shall not be larger than 50%" or "the likelihood of exceeding the threshold should not exceed one chance in 100". It would be sensible to have a set of thresholds with an associated acceptable probability of exceedance (or probability of not exceeding a threshold) that decreases as the impact increases (Figure 6-14). In this way both the prediction uncertainty and management rules are expressed in a probabilistic framework.



Figure 6-14 Relationship between impact thresholds and acceptable probability of exceeding a given threshold.

The theoretical example of Figure 6-14 is subsequently applied to the results from the groundwater impact metrics discussed in Section 6.4.1. Note that the impact metrics were expressed in terms of their cumulative distribution function, e.g. Figure 6-8. First, this plot will be converted into complementary cumulative distribution function, or CCDF (see also Turnadge et al. 2018a). A complementary cumulative distribution function represents the probability of exceedance, i.e. Prob (X > x) = 1 - Prob ($X \le x$). CCDFs are commonly used to display the results of risk assessments for two reasons. First, CCDFs answer the question "how likely is an outcome to be this large or larger", which is typically the question of interest in risk assessment. Second, CCDFs facilitate displaying small probabilities associated with large consequences.



Figure 6-15 Cumulative distribution function (CDF) and complementary cumulative distribution function (CCDF) for variable *v* with a triangular distribution on [1, 10] and a mode at 7 (based on Helton et al. 2004).

Once the CCDFs have been calculated for one or several impact metrics, they can be compared with threshold boundaries. The latter define a probability-consequence limit line for assessing, or limiting, the risk to the public or the environment. Any probability-consequence points above the limit line have an unacceptable high risk, whereas points below the limit line have an acceptable level of risk. Different limit lines may be considered, including lines of constant risk, lines that account for stakeholders being risk averse towards high consequence values (i.e., the acceptable risk at high consequence values will be smaller than at low consequence values), or lines defined by a single requirement (Cox and Baybutt, 1981). Figure 6-16 illustrates the principle of boundary lines, with one continuous line, one boundary line defined by a single requirement (i.e., probability of exceeding the consequence value of 30 should be less than one in ten) and one defined by two requirements (i.e. probability of exceeding consequence value 20 should be less than 8% and the probability of exceeding consequence value 10 should be less than 60%). Because the CCDF does not exceed any of the boundary lines, the risk is acceptable.



Figure 6-16 Boundary line approach to specification of acceptable risk (modified from Helton and Breeding, 1993).

The principle of boundary line will be applied to the results from the groundwater impact metrics discussed in Section 6.4.1. The maximum drawdown will be considered as example. Figure 6-17 displays four different potential cases of boundary lines associated with either a single requirement or with multiple requirements. The single requirement cases all have a threshold of 2 m drawdown, but different probabilities of exceedance, i.e., 30%, 50% and 80% (Figure 6-17a, b, and c). Based on the three calculated CCDFs for maximum drawdown, only the models using the initial data set would be accepted for a 30% probability. When a 50% probability is considered, models based on initial and revised data set are acceptable; for the 80% probability, all three models are acceptable. Multiple requirements are shown in Figure 6-17d: the probability of exceeding the threshold of 0.2 m shall not be larger than 50%, not larger than 10% for the threshold of 2 m, and not larger than 1% for a threshold of 10 m. In this case all three models are not acceptable for the 0.2 and 2 m drawdown; only the heterogeneous *K* model is acceptable at the 10 m threshold.



Figure 6-17 Comparison of CCDF for maximum drawdown and single requirement boundary lines for 30% (A), 50% (B), and 80% (C) acceptable probability of exceeding 2 m drawdown. Multiple requirement boundary line considers 50% acceptable probability of exceeding for 0.2 m, 10% for 2 m and 1% for 10 m drawdown (D).

Many other combinations of thresholds and acceptable probabilities of exceedance can be developed, for single impact metrics or by combining several impact metrics. For example, the acceptable probability of exceedance can be made dependant on the surface area affected by a given drawdown. The larger the affected surface area, the more stringent the tolerable probability of exceedance.

Casting groundwater impact metrics in a probabilistic framework will have greatest benefits to groundwater management if management rules are also expressed in a probabilistic sense. Future research is recommended to explore how to optimally connect water management to probabilistic results from groundwater impact studies.

7 Data worth analysis

The sensitivity analysis presented in previous chapters identified which model parameters are most important to groundwater impact metrics such as drawdown and time to drawdown predictions. Section 7.1 goes one step further and provides an overview of value of information or data worth when carrying out groundwater investigations. In section 7.2 bootstrapping is applied to estimate the robustness of statistical measures of particular groundwater metrics such as predicted maximum drawdown.

7.1 Literature review

Data worth analysis (also known as value of information analysis in the oil and gas industries) is targeted to the quantification of the impact of new potential measurements on the expected reduction of predictive uncertainty based on a given process model. The purpose is to identify how adding new (or removing existing data) would decrease (increase) model prediction uncertainty. Or in a management context, whether purchasing a new information source would result in making better decisions (Trainor-Guitton et al., 2011). Indeed, the high costs associated with collecting hydrogeological parameters, setting up a new or extending an existing monitoring network indicates the need to develop robust methodologies conducive to the identification of optimal strategies for the collection of future data which are potentially valuable for a specific environmental goal considered. Goal oriented data sets can assist to improve understanding of complex systems and minimise uncertainty while considering budget constraints (Dai et al., 2016). In other words, site investigations should only be carried out if the risk reduction it will achieve is greater than the cost of carrying it out.

Especially for large-scale site characterisations a careful selection has to be made prior to any data collection efforts about what the best combination of hard and soft data will be, depending on the management questions that needs to be addressed. Value of information or data worth analysis studies then become very useful to ensure the limited resources are used to provide data most suitable to resolve a particular question (Engelhardt et al., 2013; Freeze et al. 1992; Moore, 2005; Wallis et al., 2014). However, the value of information is not absolute; instead, it is always dependent on the management or research question of interest. Indeed, there is no intrinsic value in collecting data unless it can influence a specific decision goal, or reduce the risk of an undeniable event that has a cost associated with it.

For a data worth analysis to be practical, the acceptable model prediction uncertainty should be established (Finsterle, 2015). Developing criteria of success in terms of acceptable prediction uncertainty is critical to reduce the risk of undertaking extensive data collection efforts that could fail to support a specific scientific or technical question. In other words, a specific decision goal has to be formulated; reducing uncertainty on subsurface parameters for the sake of reducing uncertainty has little or no value. Once acceptable prediction uncertainty is defined, limits can be imposed on the acceptable level of uncertainty in the input parameters. Subsequently, a data collection campaign can be designed in such a way that it will result in input parameters with

acceptable uncertainties. A non-exhaustive summary of data worth or value of information applications in groundwater investigations and management is provided in Table 7-1.

Table 7-1 Examples of data worth or value of information applications in groundwater investigations and management.

Subject	Application	Reference
Data worth and its use for developing site investigations strategies	Reduction of uncertainty in aquitard continuity and the reduction of uncertainty in hydraulic-conductivity distribution in an aquifer	Freeze et al. (1992)
Data-worth analysis of potential new monitoring- well locations to improve models of groundwater/ surface-water interactions	Data-worth analysis of potential new monitoring-well locations by using a model. The relative worth of new measurements was evaluated based on their ability to increase confidence in model predictions of groundwater levels and base flows	Leaf et al. (2015)
Bayesian analysis of groundwater data worth	Assessment of the worth of collecting additional data on steady- state flow, with log hydraulic conductivity data found to be worth more than an equal number of corresponding head measurements	Xue et al. (2014)
Groundwater quality management under uncertainty	To identify optimal pumping and sampling strategies to minimise model uncertainty within the context of ground-water management	Wagner (1999)
Evaluating data worth for groundwater management	For a specified data collection budget, the monitoring network design model identifies, prior to data collection, the sampling strategy that will minimize model uncertainty when designing the containment of a plume of groundwater contamination through the installation and operation of pumping wells	Wagner et al. (1992)
Uncertainty and data worth analysis for the hydraulic design of funnel-and-gate systems in heterogeneous aquifers	Determine optimal sampling locations for additional conductivity measurements to guide design of funnel- and-gate systems as part of a permeable reactive barrier in a contaminated aquifer	Cirpka et al. (2004)
Estimation, optimization, and value of information for groundwater remediation.	Consideration of groundwater clean-up objectives, cost formulations, and sensitivity of costs to uncertainty in parameters, measurements, and the model itself to seek minimization of expected cost under conditions of incomplete information	Liu et al. (2012)
Optimisation of monitoring data for increased predictive reliability of regional water allocation models	Identify optimal data monitoring networks, where "optimal" is defined in terms of increased reliability of the particular prediction that underpins a management decision	Moore et al., (2011)
A decision tree model to estimate the value of information provided by a groundwater quality monitoring network	Estimation of the value of information provided by a groundwater quality monitoring network located in an aquifer whose water poses a spatially heterogeneous and uncertain health risk	Khader et al. (2013)
Impacts from re-injection of coal seam gas produced water	Explore the relative data-worth of injection tracer tests, pump tests and regional groundwater monitoring data in informing the dilution predictions made using a well field water quality model.	Sreekanth and Moore (2015)
Data worth analysis to determine cost effectiveness of airborne electromagnetic (AEM) data for defining hydraulic properties in a groundwater flow model	How cost-effective are airborne electromagnetic (AEM) data for refining the spatial variability of hydraulic conductivity and vertical aquifer boundaries within a groundwater flow model when compared to more traditional techniques	Magali et al. (2016)

7.2 Bootstrapping resampling

7.2.1 Methodology and application

The data worth analysis in this study is based on a bootstrap resampling methodology. The bootstrap resampling approach is a non-parametric way of calculating prediction confidence intervals and parameter uncertainty (Lall and Sharma, 1996; Burn, 2003). The resampling approach involves creating new samples from the original sample by a bootstrapping process which involves randomly selecting data points, with replacement, from the original sample and subsequently estimating prediction uncertainty from each of the resampled data sets.

The principle of bootstrap resampling is illustrated here based on the updated data set and its 300 parameter combinations generated in the previous section (Section 5). The 300 parameter combinations are randomly sampled from an unimodal log-triangular distribution that is based on the improved aquitard characterisation using upscaled well-log data combined with hydraulic conductivity measurements. The uncertainty around the maximum drawdown will be considered as an example.

The bootstrap procedure follows following scheme:

- 1. Take a random sample of 50 parameter combinations with resampling from the 300 available parameter combinations;
- 2. For each parameter combination, used the existing model runs to compute the 50th, 75th and 90th percentiles of maximum drawdown;
- 3. Repeat steps 1-2 a large number of time (1 000 in this example);
- 4. Summarize with histograms the variability in the percentiles of the parameters and model predictions (i.e. drawdown).

This type of analysis shows how robust the summary statistics (the percentiles) of the predictions are. The exercise may be repeated for different sample sizes (e.g. 50, 200, 300) which allows to compare the range of the percentiles, and if they are affected by the sample size. If the range of percentiles is small, the sampling density is considered high enough and the summary statistic is considered robust. If the range or uncertainty is high, however, the summary statistic can then not be considered robust and the data density is insufficient to characterise the parameter distribution according to an acceptable degree of uncertainty. The bootstrap resampling method does not assess the value of the data itself, e.g. how the prediction uncertainty would change as the number of core-based K_V values in the data set decreases or increases. Addressing such questions was beyond the scope of this study.

7.2.2 Results and discussion

Figure 7-1 and Figure 7-2 show the results of a 1 000 fold bootstrap resampling with respectively a sample size of 50 and 300. The bootstrap with sample size 50 shows that there is considerable spread in the percentiles of predicted drawdown, with the median varying between 0.4 and 2 m, while the 90th percentile varies between 5.2 m and 8 m. Figure 7-1 shows that by increasing the
sample size to 300 the range of predictions decreases considerably, with the range of the median between 0.9 and 1.6 m and the range of the 90th percentile now about 2 m (between 5.6 m and 7.8 m). The percentiles of the parameters show a similar decrease in spread at the 50th, 75th and 90th percentiles when the sampling size increases from 50 to 300. The smaller percentiles of the predictions are not shown, as they are close to zero and show no variation. It highlights once more that the predictive distribution is strongly skewed to the right.



Figure 7-1. Illustration of the bootstrapping approach to estimate parameter (a) and model prediction uncertainty (b) based on a sample size of 50.



Figure 7-2. Illustration of the bootstrapping approach to estimate parameter (a) and model prediction uncertainty (b) based on a sample size of 300.

In the next step the 90 % confidence interval ranges were derived for each percentile of the maximum drawdown prediction based on the results shown in Figure 7-1b and Figure 7-2b. The 90% confidence intervals were derived for increasing sample size from 50, 100, 200, and 300 (Table 7-2). Because the distribution of the percentiles was not necessarily normal, the 90% confidence intervals were derived from the cumulative distribution function of calculated percentiles (i.e. based on the values shown in Figure 7-1b and Figure 7-2b), leaving 5% in each of the tails of the distribution. As is clear from Table 7-2, the 90% confidence interval becomes smaller and nearly constant as the sample size increases (Figure 7-3). On the basis of bootstrapping it is thus demonstrated that having generated 300 data sets followed by 300 model runs is sufficient to provide robust estimators of summary statistics of groundwater impact metrics (here the maximum drawdown). Further increasing the sample size would have decreased the 90% confidence interval marginally compared to the gains obtained by increasing the sample size from 50 to 300.

Table 7-2 90%	<mark>ն confidence</mark> i	ntervals de	erived for	summary	statistics p50,	p75, and p90.
---------------	-----------------------------	-------------	------------	---------	-----------------	---------------

Sample size	p50	p75	p90
50	1.570	3.058	3.123
100	1.036	2.575	2.042
200	0.615	1.980	1.577
300	0.490	1.802	1.400



Figure 7-3 Effect of sample size on robustness (90 % confidence interval) of estimated percentiles of maximum drawdown.

In a final step the bootstrapping analysis was repeated for both the homogenous and heterogeneous model using a sample size of 50 (to have a consistent sample size across both model approaches). As Figure 7-4 shows, the percentiles for the groundwater impact metric maximum drawdown are much better defined (i.e. smaller data range) for the heterogeneous than for the homogeneous model.



Figure 7-4 Illustration of the bootstrapping approach to estimate model prediction uncertainty (maximum drawdown) based on a sample size of 50 for the homogeneous (a) and heterogeneous model (b).

Whether or not a sample size of 50 heterogeneous K_V fields and subsequent 50 model runs was sufficient to obtain robust estimations of the four groundwater impact metrics is illustrated in Figure 7-5. For all four impact metrics, all statistics (full data range, interquartile range, median) reach more or less constant values at a sample size of 50 indicating that a sample size of 50 was indeed sufficient.



Figure 7-5 Illustration of the bootstrapping approach to estimate model prediction uncertainty across different sample sizes (heterogeneous model). Box plots show full data range, interquartile range (box) and median (red line).

The analysis also illustrates that in order to characterise the higher extremes of the distribution of model predictions, more effort needs to be expended in reliably characterising the right tail of the parameter distribution, i.e. the high values. Increasing the reliability of the left tail will have limited effect on the predictions. Having a right tail that is too wide or too narrow may result in a considerable over or under prediction respectively of maximum drawdown.

As such the bootstrap procedure outlined and applied above only tested the robustness of the summary statistics with regards to the sampling of a known log-triangular distribution. A possible next step in the data worth analysis would be to include the definition of the log-triangular parameter distribution in the bootstrap procedure (i.e. use actual data to define a distribution and its parameters). This can be achieved by taking random samples with replacement from the set of hydraulic property measurements; for the current study this would mean the second data set. Based on each random sample, a new log-triangular distribution would then be formulated which is subsequently sampled to run the groundwater flow model and generate the corresponding maximum drawdown. While the analysis presented in Figure 7-1 and Figure 7-2 is based on a post-processing of the existing 300 model runs, this additional analysis would require additional model runs as adequate sampling is no longer guaranteed based on the existing distribution, when the log-triangular distribution changes each time a new sample is generated.

8 Summary and conclusions

The present report described an example for the inclusion of upscaled aquitard parameter values in a numerical groundwater flow model. The values and associated uncertainties of four selected predictions generated by a numerical groundwater flow model were assessed prior to and following the incorporation of aquitard vertical hydraulic conductivity data. Sensitivity analyses with an initial data set based on 300 model runs illustrated the relative importance of the vertical hydraulic conductivity (K_V) of the upper aquitard as an influential parameter when calculating four groundwater impact metrics.

*K*_V data for the two example aquitards were measured at the core-scale by laboratory testing, correlated with wireline logging data and subsequently upscaled to the regional scale commensurate with the scale of a cellular groundwater flow model using analytical and numerical methods. Upscaled aquitard property observations were used to update prior parameter distributions which, after incorporation in the model, produced updated prediction (posterior) distributions. The workflow presented here is one example that is considered to be practicable and suitable for certain real-world applications.

Key findings from using updated aquitard parameterisations for modelling groundwater impact metrics are:

- Improved characterisation of aquitard *K*_V resulted in more credibly defining (i.e. based on a combination of field and lab-based data with a relatively high spatial density) the probability distribution for *K*_V, with the prior log-uniform distribution being replaced by an unimodal log-triangular distribution;
- For both aquitards, the updated (posterior) probability distribution for K_V was roughly two orders of magnitude wider than the prior distributions; mean K_V values were nearly identical for prior ($\log_{10} K_V = -5.0 \text{ m/d}$ for both upper and lower aquitards) and posterior distributions ($\log_{10} K_V = -5.2 \text{ m/d}$ for upper and = -4.7 m/d for lower aquitard, based on numerical upscaling). Although improving the characterisation of the aquitards resulted in an increase in parameter range (i.e., the uncertainty associated with K_V increased), the new data set has a solid evidence base and thus its credibility (and the model predictions based on it) has significantly improved the aquitard parameterisation compared to the initial parameterisation based on literature values and models;
- Regarding predicted groundwater impact metrics, the median of the maximum modelled drawdowns increases slightly (from 0.7 m to 1.2 m) when the first (prior) data set is replaced by the second data set (improved aquitard characterisation with *K*_V spatially uniform within each model layer). Maximum modelled drawdown further increases (from 1.2 m to 3.4 m) when the more realistic model with heterogeneous *K*_V is used. Overall, the calculations of the median modelled drawdown are fairly robust (in a statistical sense) for the three data sets tested. Improved aquitard characterisation does affect the predicted median drawdown, however for the model and data sets used here the effects are rather minor.

- Extreme drawdowns (magnitude and spatial extent of 95th percentile) are similar for the ٠ first and second data set (7.2 and 7.9 m, respectively), but much smaller with the heterogeneous K_V model (third data set, 6.1 m). The improved aquitard characterisation and use of an improved conceptualisation (heterogeneous versus homogeneous hydraulic conductivity distribution) thus has a major effect on such percentiles. The much smaller extreme drawdowns are also apparent in the maps of the spatial distribution of drawdowns (i.e., the 90th, 95th and 99th percentile of maximum drawdowns), illustrating that the heterogeneous K_V model is much better constrained resulting in a much smaller range of extreme drawdowns. For instance, the spatial extent of drawdowns > 2 m decreases from 5 350 for the second data set to 550 for the third data set. The extreme values (e.g., 95th percentiles) are materially affected (i.e. smaller) by using the heterogeneous K_V model. Main reasons for this result are: i) a better constrained K_V model owing to the use of conditioned Sequential Gaussian Simulation, which honours the observed K_V data; and ii) the smaller number of model runs, 50, for the heterogeneous model versus 300 for the homogeneous models, causing a smaller K_V parameter space to be sampled.
- Time to maximum drawdown decreased considerably when the heterogeneous model was implemented: the median value decreased from 155 (homogenous model) to 35 year (heterogeneous model), and the 95th percentile decreased from 160 (homogenous model) to 81 year (heterogeneous model). Unlike the homogeneous models, all of the heterogeneous model runs do achieve their maximum drawdown within the total model run time, yielding the smaller median value of 35 years. This result is believed to be due to the higher degree of connectivity within a heterogeneous model when both high and low conductivity zones are present in the model. These conditions are conducive to propagate the depressurisation occurring in the coal formation much faster than models that have a homogeneous conductivity model. The heterogeneous model thus provides a more realistic reaction time of the groundwater systems, and is a more accurate approach for obtaining a more robust estimate of likely timing of maximum drawdown.

Casting groundwater impact metrics in a probabilistic framework will be of greatest benefits to groundwater management if management rules are also expressed in a probabilistic sense. Future research is recommended to explore how to optimally connect water management to probabilistic results from groundwater impact studies.

Bootstrapping analysis demonstrated how robust the summary statistics (i.e., the percentiles) of the groundwater impact predictions are. Based on a bootstrap resampling with respectively a sample size of 50 and 300, the bootstrap with sample size 50 shows considerable spread in the percentiles of predicted drawdown. By increasing the sample size from 50 to 300 the range of predictions decreases considerably, demonstrating how robustness of the summary statistics improve as sample size increases.

Main learnings from this study relevant to other characterisation and groundwater modelling studies are as follows:

- Improved characterisation of aquitard *K*_V resulted in more credibly defining the probability distribution for *K*_V, with the rather arbitrarily chosen prior log-uniform distribution being replaced by a unimodal log-triangular distribution. The new data set has a solid evidence base and thus its credibility (and the model predictions based on it) has significantly improved. The model further provides a basis for subsequent investigations that aim to reduce model uncertainty once the key factors contributing to uncertainty have been identified. This underscores the need to improve more broadly characterisation efforts of hydrogeological parameters to progressively reduced predictive uncertainty to a level that both the modelling community and regulators are comfortable with. Building reliable groundwater models is an iterative process whereby uncertainties in the initial parameters and model components are reduced progressively through data collection, sensitivity and uncertainty analysis (Gedeon et al., 2013).
- Extreme drawdowns (magnitude and spatial extent of 95th percentile) are much smaller for the heterogeneous model compared to the homogeneous model. This illustrates that the heterogeneous model is much better constrained resulting in a much smaller range of extreme drawdowns. The better constrained *K*_V model results from using the conditioned Sequential Gaussian Simulation, which honours the observed *K*_V data.
- The heterogeneous model provides a more realistic reaction time of the groundwater systems, and is a more accurate approach for obtaining a more robust estimate of likely timing of maximum drawdown.
- The use of improved aquitard conceptualisations and parameterisation does not necessarily result in overall reduction in predictive uncertainty. However, incorporation of a more data-driven and spatially heterogeneous hydraulic conductivity parameterisation always results in a more technically defensible model and provides more credible model predictions.

References

- Ababou R (1988) *Three-Dimensional Flow in Random Porous Media*, PhD thesis, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2 vol.s, 833p.
- Ababou R, McLaughlin D, Gelhar LW, and Tompson AF (1989) Numerical simulation of three dimensional saturated flow in randomly heterogeneous porous media, *Transport in Porous Media* 4: 549-565.
- Arena A, Carnerup A, Deakin L, Golab A, Green C, Hussain F, Goodwin C, Rajan P, Dodd N, Khor J, Knackstedt M, Knuefing L, Larsson A, Sok R, Sommacal S, Young B, Zhang J (2016) Maximising the value of digital core analysis for carbon sequestration site assessment. Milestone 0.4: Final project report.
- Bakker M, Post VEA, Langevin CD, Hughes JD, White JT, Starn JJ and Fienen MN (2016) Scripting MODFLOW model development using Python and FloPy, *Ground Water* 54: 733–739.
- Bohling G (2005) Introduction to geostatistics and variogram analysis, Kansas Geological Survey, URL=<https://pdfs.semanticscholar.org/ef33/f1ef327b1b8bcf5e424f9f4c90b51fc78ae5.pdf>, accessed 22 March 2017, 20p.
- Borgonovo E (2007) A new uncertainty importance measure, *Reliability Engineering and System Safety* 92(6): 771-784.
- Bredehoeft JD (2005) The conceptualization model problem—surprise. *Hydrogeology Journal* 13: 37–46.
- Burn D H (2003) The use of resampling for estimating confidence intervals for single site and pooled frequency analysis, *Hydrological Sciences Journal* 48: 25-38.
- Carrera J, Alcolea A, Medina A, Hidalgo J and Slooten L J (2005) Inverse problem in hydrogeology Hydrogeology Journal, 13, 206-222 http://dx.doi.org/10.1007/s10040-004-0404-7
- CDM Smith (2016) *Narrabri Gas Project Groundwater Modelling Report,* report produced for Santos Ltd, 389p.
- Cirpka OA, Buerger CM, Nowak W and Finkel M (2004) Uncertainty and data worth analysis for the hydraulic design of funnel-and-gate systems in heterogeneous aquifers. Water Resources Research, Vol. 40, W11502, doi:10.1029/2004WR003352.
- Cook P, Miller A, Shanafield M and Simmons C (2016) Predicting Water Resource Impacts of Unconventional Gas Using Simple Analytical Equations Groundwater, http://dx.doi.org/10.1111/gwat.12489.
- Cox DC and Baybutt P (1981). Limit lines for risk, Nuclear Technology 57: 320-330.
- Cressie N (1985) Fitting variogram models by weighted least squares, *Mathematical Geology* 17(5): 563-586.

- CSIRO (2012) Water resource assessment for the Surat region. A report to the Australian Government from the CSIRO Great Artesian Basin Water Resource Assessment. CSIRO Water for a Healthy Country Flagship, Australia.
- Dai C, Xue L, Zhang D, and Guadagnini A (2016) Data-worth analysis through probabilistic collocation-based ensemble Kalman filter, *Journal of Hydrology* 540: 488-503.
- Der Kiureghian A, Ditlevsen O (2007) Aleatory or epistemic? Does it matter? Special Workshop on Risk Acceptance and Risk Communication March 26-27, 2007, Stanford University: 13 pp.
- Desbarats A (1987) Numerical estimation of effective permeability in sand-shale formations, *Water Resources Research* 23(2): 273-286.
- Distinguin M, Lavanchy JM (2007) Determination of hydraulic properties of the Callovo -Oxfordian argillite at the Bure site: Synthesis of the results obtained in deep boreholes using several in situ investigation technniques, *Phys. Chem. Earth*, 32: 379–392.
- Deutsch C (1989) Calculating effective absolute permeability in sandstone/shale sequences, SPE Formation Evaluation 4(3).
- Engelhardt I, Prommer H, Moore C, Schulz M, Scheuth C, and Ternes TA (2013) Suitability of temperature, hydraulic heads, and acesulfame to quantify wastewater-related fluxes in the hyporheic and riparian zone, *Water Resources Research* 49: 426–440.
- Ferretti F, Saltelli A and Tarantola S (2016) Trends in sensitivity analysis practice in the last decade, Science of the Total Environment 568: 666-670.
- Finsterle S (2015) Practical notes on local data-worth analysis, *Water Resources Research* 51: 9904-9924.
- Freeze RA and Cherry JA (1979) Groundwater, Prentice Hall, Englewood Cliffs, NJ, USA, 604p.
- Freeze RA, James B, Massmann J, Sperling T and Smith L (1992) Hydrogeological decision analysis:
 4. The concept of data worth and its use in the development of site investigation strategies, Groundwater 30(4): 574-588.
- Gedeon M, Mallants D, and Rogiers B (2013) Building a staircase of confidence in groundwater modeling: a summary of ten years data collection and model development, Modflow and More 2013: Translating Science into Practice, Golden, CA, 2-5 June, pp. 545-550.
- Gelhar LW (1993) Stochastic Subsurface Hydrology, Englewood Cliffs, NJ, Prentice Hall, 1993.
- Gelhar LW (1986). Stochastic subsurface hydrology: From Theory to Applications, *Water Resources Research* 22(9): 135s-145s.
- Giambastiani B M S, Kelly B F J, The C, Andersen M S, McCallum A M and Acworth R I (2009) 3D time and space analysis of groundwater head change for mapping river and aquifer interactions 18th World Imacs Congress and Modsim09 International Congress On Modelling and Simulation: Interfacing Modelling and Simulation With Mathematical and Computational Sciences, 3067-3073 http://www.mssanz.org.au/modsim09/I1/giambastiani.pdf
- Geoscience Australia (2016) Australian Stratigraphic Units Database, URL=http://dbforms.ga.gov.au/www/geodx.strat_units.int, last accessed 08/11/16.

- Gómez-Hernández J and Wen X (1998) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology, *Advances in Water Resources* 21(1): 47-61.
- Guillaume JHA, Hunt RJ, Comunian A, Blakers RS and Fu B (2016) Methods for Exploring Uncertainty in Groundwater Management Predictions, *Integrated Groundwater Management*, 711-737. http://dx.doi.org/10.1007/978-3-319-23576-9_28.
- Gupta HV, Clark MP, Vrugt JA, Abramowitz G and Ye M (2012) Towards a comprehensive assessment of model structural adequacy, *Water Resources Research* 48(8).
- Harbaugh AW (2005) MODFLOW-2005, The US Geological Survey Modular Ground-Water Model: The Ground-Water Flow Process, US Geological Survey Techniques and Methods report no. 6-A16, US Department of the Interior, Reston, VA, USA, 253p.
- Helton JC and RJ Breeding (1993) Calculation of reactor accident safety goals, *Reliability* Engineering and Systems Safety 39: 129-158.
- Helton JC, Johnson JD, Oberkampf WL (2004) An exploration of alternative approaches to the representation of uncertainty in model predictions, *Reliability Engineering and System Safety* 85: 39–71.
- Herman JD and Usher W (2016) SALib: An open-source Python library for sensitivity analysis, Journal of Open Source Software, http://dx.doi.org/10.21105/joss.00097.
- Hill MC (2003) Preconditioned Conjugate-Gradient 2 (PCG2): A Computer Program for Solving Ground-Water Flow Equations, 2nd printing, US Geological Survey Water Resources Investigations Report no. 90-4048, 25p.
- Hill Mc (2006) The Practical Use of Simplicity in Developing Ground Water Models. Ground Water 44(6): 775-781.
- Højberg AL, Refsgaard JC (2005) Model uncertainty parameter uncertainty versus conceptual models. *Water Science Technology* 52(6): 177–86.
- Hunt and Zheng (1999) Eos Transactions, AGU, p. 29, January 19, 1999.
- HydroGeoLogic (1996) *MODHMS/MODFLOW-SURFACT software documentation Volume 1: Groundwater flow modules*, HydroGeoLogic Inc., Reston, VA, USA.
- Janardhanan S, Crosbie R, Pickett T, Cui T, Peeters L, Slatter E, Northey J, Merrin LE, Davies P, Miotlinski K, Schmid W and Herr A (2018) Groundwater numerical modelling for the Namoi subregion. Product 2.6.2 for the Namoi subregion from the Northern Inland Catchments Bioregional Assessment. Department of the Environment and Energy, Bureau of Meteorology, CSIRO and Geoscience Australia, Australia, http://data.bioregionalassessments.gov.au/product/NIC/NAM/2.6.2.
- Journel AG, Deutsch C and Desbarats AJ (1986) Power averaging for block effective permeability, SPE California Regional Meeting, Society of Petroleum Engineers.
- Journel AG and Huijbregts CJ (1978) *Mining Geostatistics*, Academic Press, New York, USA, 600p.
- Khader AI, Rosenberg DE and McKee M (2013) A decision tree model to estimate the value of information provided by a groundwater quality monitoring network. Hydrol. Earth Syst. Sci., 17, 1797–1807.

Kitanidis PK (1997) Introduction to geostatistics, Cambridge University Press, Cambridge, UK, 249p.

- Klohn Crippen Berger (KCB) (2012) *Forecasting coal seam gas water production in Queensland's Surat and southern Bowen basins*, technical report for the Department of Natural Resources and Mines, Qld, Brisbane, Australia, 118p.
- Kolterman CE and Gorelick SM (1996) Heterogeneity in sedimentary deposits: A review of structure-imitating, process-imitating, and descriptive approaches, *Water Resources Research* 32(9): 2617-2658.
- Lall U and Sharma A (1996) A nearest neighbour bootstrap for resampling hydrologic time series, Water Resources Research 32: 679-693.
- Leaf AT, Fienen MN, Hunt RJ and Buchwald CA (2015) Groundwater/Surface-Water Interactions in the Bad River Watershed, Wisconsin: U.S. Geological Survey Scientific Investigations Report 2015–5162, 110 p., http://dx.doi.org/10.3133/sir20155162.
- Li L, Zhou H and Gomez-Hernandez J (2011) A comparative study of three-dimensional hydraulic conductivity upscaling at the macro-dispersion experiment (MADE) site, Columbus Air Force Base, Mississippi (USA), *Journal of Hydrology* 404(3-4): 278-293.
- Liu X, Lee J, Kitanidis PK, Parker J and Kim U (2012) Value of Information as a Context-Specific Measure of Uncertainty in Groundwater Remediation. *Water Resource Management* (2012) 26:1513–1535.
- Mallants D, Volckaert G, and Labat S (2001) *Parameter values used in the performance assessment* of the disposal of low level radioactive waste at the nuclear zone Mol-Dessel, vol. 1, 2 & 3, report no. R-3521, SCK•CEN, Mol, Belgium, 41p (vol. 1), 220p (vol. 2), 63p (vol. 3).
- Mazurek, M, Alt-Epping, P, Bath, A, Gimmi, T, Waber, HN, Buschaert, S, De Canniere, P, De Craen, M, Gautschi, A, Savoye, S, Vinsot, A, Wemeare, I and Wouters, L (2011) Natural tracer profiles across argillaceous formations, *Applied Geochemistry* 173: 219-240.
- McKenna SA and CA Rautman (1996) Scaling of Material Properties for Yucca Mountain: Literature Review and Numerical Experiments on Saturated Hydraulic Conductivity. SAND95-2338.
- Neuman SP and Di Federico V (1998) Correlation, flow, and transport in multiscale permeability fields, in: Scale Dependence and Scale Invariance in Hydrology, Cambridge University Press, Cambridge, UK, pp. 354-397.
- Neuman SP and Wierenga PJ (2003) A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites. NUREG/CR-6805 U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Washington, DC.
- New South Wales Department of Primary Industries Office of Water (NSW DPI OW) (2012) NSW Aquifer Interference Policy, State of New South Wales Department of Trade and Investment, Regional Infrastructure and Services, 34p,

URL=<http://www.water.nsw.gov.au/__data/assets/pdf_file/0004/549175/nsw_aquifer_int erference_policy.pdf>, accessed 27/02/17.

Niswonger RG, Panday S and Ibaraki M (2011) *MODFLOW-NWT: A Newton formulation for MODFLOW-2005*, US Geological Survey Techniques and Methods report no. 6–A37, 44p.

- Moore CR (2005) *The Use of Regularized Inversion in Groundwater Model Calibration and Prediction Uncertainty Analysis*, PhD thesis, University of Queensland, St. Lucia, Australia.
- Moore CR and Doherty J (2006) The cost of uniqueness in groundwater model calibration, Advances in Water Resources 29(4): 605–623.
- Moore CR, Wohling T and Wolf L (2011) Optimisation of monitoring data for increased predictive reliability of regional water allocation models, 19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011.
- Moore CR, Doherty J, Howell S and Erriah L (2013). Upscaling hydraulic properties and processes in the Coal Seam Gas Context Heterogeneity and dual phase flow challenges in lithologically segregated upscaling, CSIRO client report, 129p.
- Moore CR, Doherty J, Howell and Erriah L (2015) Some challenges posed by coal bed methane regional assessment modeling, *Groundwater* 53(5): 737-747.
- Moore T (2012) Coalbed methane: a review, International Journal of Coal Geology 101: 36–81.
- Moreau M, Moore CR, Rawlinson Z (2016) Predictive uncertainty and data worth analysis to determine cost effectiveness of airborne EM data for defining hydraulic properties in a groundwater flow model used for predicting long term groundwater level drawdowns.
 International Association of Hydrogeologists. 43rd IAH CONGRESS 25-29th September, 2016 le Corum, Montpellier, France.
- Nossent J, Elsen P, Bauwens W (2011). Sobol' sensitivity analysis of a complex environmental model, *Environmental Modelling and Software* 26: 1515-1525.
- Office of Groundwater Impact Assessment (OGIA) (Qld) (2016) *Underground Water Impact Report for the Surat Cumulative Management Area*, Department of Natural Resources and Mines, 270p.
- Pianosi F, Beven K, Freer J, Hall JW, Rougier J, Stephenson DB, Wagener T (2016) Sensitivity analysis of environmental models: A systematic review with practical workflow, *Environmental Modelling and Software* 79: 21 -232.
- Plischke E, Borgonovo E and Smith CL (2013) Global sensitivity measures from given data, *European Journal of Operational Research* 226(3): 536-550.
- Ravestein JC, Griffiths CM, Dyt CP and Michael K (2015). Multi-scale stratigraphic forward modelling of the Surat Basin for geological storage of CO₂. *Terra Nova* 27: 346–355.
- Refsgaard JC, Christensen S, Sonnenborg TO, Seifert D, Højberg AL, Troldborg L (2012) Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. *Advances in Water Resources* 36: 36-50.
- Renard P and de Marsily G (1997) Calculating equivalent permeability: A review, Advances in Water Resources 20 (5-6): 253-278.
- Richardson JG, Harris DG, Rossen RH, and Van Hee G (1978) The effect of small, discontinuous shales on oil recovery, J. of Pet. Tec., November: 1531-1537.

- Rojas RKS, Peeters L, Batelaan O, Feyen L and Dassargues A (2010) Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling, *Journal of Hydrology* 394: 416-435.
- Saltelli A and Annoni P (2010) How to avoid a perfunctory sensitivity analysis, *Environmental Modelling and Software* 25(12): 1508-1517.
- Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S and Young P (2008) *Global Sensitivity Analysis: The Primer*, John Wiley and Sons, Chichester, UK, 292p.
- Saltelli A, Tarantola S and Campolongo F (2000) Sensitivity analysis as an ingredient of modelling, *Statistical Science* 15(4): 377-395.
- Sanchez-Vila X, Carrera J and Girardi J (1996) Scale effects in transmissivity, *Journal of Hydrology* 183(1-2): 1-22.
- Sanchez-Vila X, Girardi J and Carrera J (1995) A synthesis of approaches to upscaling of hydraulic conductivities, *Water Resources Research* 31(4): 867-882.
- Sanchez-Vila X, Guadagnini A and Carrera J (2006) Representative hydraulic conductivities in saturated groundwater flow, *Reviews of Geophysics* 44(3).
- Schulze-Makuch D, Carlson DA, Cherkauer DS and Malik P (1999). Scale Dependency of Hydraulic Conductivity in Heterogeneous Media, *Ground Water*, 37(6): 904-919
- Simmons CT and Hunt RJ (2012) Updating the Debate on Model Complexity. GSA Today 22(8).
- Smith SD, Mathouchanh E and Mallants D (2018) Characterisation of fluid flow in aquitards using helium concentrations in quartz, Gunnedah Basin, NSW, CSIRO, Australia.
- Sobol' IM (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Mathematics and Computers in Simulation* 55(1): 271-280.
- Sreekanth J and Moore C (2015) CSG water injection impacts: Modelling, uncertainty and risk analysis; Groundwater flow and transport modelling and uncertainty analysis to quantify the water quantity and quality impacts of a coal seam gas produced water injection scheme in the Surat Basin, Queensland. CSIRO, Australia.
- Srinivasan G, Tartakovsky DM, Robinson BA, and Aceves AB (2007) Quantification of uncertainty in geochemical reactions. Water Resources Research 43, W12415.
- Trainor-Guitton WJ, Caers JK, Mukerji T (2011) A methodology for establishing a data reliability measure for value of spatial information problems, *Mathematical Geosciences* 43: 929-949.
- Troldborg L, Refsgaard JC, Jensen KH, Engesgaard P (2007) The importance of alternative conceptual models for simulation of concentrations in multiaquifer system. *Hydrogeology Journal* 15:843–60.
- Turnadge C, Mallants D, Peeters L (2018a) Overview of aquitard and geological fault simulation approaches in regional scale assessments of coal seam gas extraction impacts, prepared by the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra.
- Turnadge C, Esteban L, Emelyanova I, Nguyen D, Pervukhina M, Han T and Mallants D (2018b) Multiscale Aquitard Hydraulic Conductivity Characterisation and Inclusion in Groundwater

Flow Models: Application to the Gunnedah Basin, New South Wales, prepared by the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Canberra, ACT, 66p.

- Vrugt JA, ter Braak CJF, Diks CGH, and Schoups G (2013) Advancing hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications, *Advances in Water Resources* 51, 457-478, doi:10.1016/j.advwatres.2012.04.002.
- Wagner JM, Shamir U, Nemati HR (1992) Groundwater quality management under uncertainty stochastic-programming approaches and the value of information. *Water Resources Research* 28(5):1233–1246. ISSN 0043-1397
- Wagner BJ (1999) Evaluating data worth for groundwater management under uncertainty. *Journal* of Water Resources Planning and Management, Vol. 125, No. 5, 281-288.
- Wallis I, Moore C, Post V, Wolf L, Martens E, Prommer H (2014). Using predictive uncertainty analysis to optimise tracer test design and data acquisition, *Journal of Hydrology* 515: 191–204.
- Warren J and Price H (1961) Flow in heterogeneous porous media, *Society of Petroleum Engineers Journal* 1(3): 153-169.
- Watermark Numerical Computing (WNC) (2016) Groundwater Data Utilities Part B: Program Descriptions, Watermark Numerical Computing, URL=<http://www.pesthomepage.org/getfiles.php?file=gwutil_b.pdf>, accessed 27/02/17.
- Weber K (1982) Influence of common sedimentary structures on fluid flow in reservoir models, J. of Pet. Tech., March: 665-672.
- Weber K (1986) How heterogeneity affects oil recovery, Reservoir Characterization 487: 544.
- Wen X and Gomez-Hernandez J (1996) Upscaling hydraulic conductivities in heterogeneous media: An overview, *Journal of Hydrology* 183(1-2): R9-R32.
- White J T, Doherty J E and Hughes J D (2014) Quantifying the predictive consequences of model error with linear subspace analysis Water Resources Research, 50, 1152-1173 http://dx.doi.org/10.1002/2013WR014767.
- Xue L, Zhang D, Guadagnini A, Neuman SP (2014) Multimodel bayesian analysis of groundwater data worth. Water Resources Research http://dx.doi.org/10.1002/2014WR015503.
- Yu L, Rogiers B, Gedeon M, Marivoet J, De Craen M and Mallants D (2013) A critical review of laboratory and in-situ hydraulic conductivity measurements for the Boom clay in Belgium, *Applied Clay Science* 75: 1-12.
- Zuidema P (1994) Validation: Demonstration of disposal safety requires a practicable approach, in GEOVAL 94, Validation through model testing, Proceedings of an NEA/SKI Symposium, Paris, France, 11-14 October 1994, pp. 35-42.

9 Appendix 1

The following appendices contain scatterplots of prediction versus parameter values for each of the four groundwater impact metrics.

9.1 Initial model: Magnitude of maximum drawdown prediction (MXD)



100 | Sensitivity and uncertainty analysis of a regional-scale groundwater flow model featuring coal seam gas extraction

9.2 Revised model: Magnitude of maximum drawdown prediction (MXD)



9.3 Initial model: Timing of maximum drawdown prediction (tMXD)



102 | Sensitivity and uncertainty analysis of a regional-scale groundwater flow model featuring coal seam gas extraction

9.4 Revised model: Timing of maximum drawdown prediction (tMXD)



9.5 Initial model: Number of model cells with drawdown > 2 m prediction (NDD)



104 | Sensitivity and uncertainty analysis of a regional-scale groundwater flow model featuring coal seam gas extraction

9.6 Revised model: Number of model cells with drawdown > 2 m prediction (NDD)



9.7 Initial model: Maximum vertical flux prediction (MXQ)



106 | Sensitivity and uncertainty analysis of a regional-scale groundwater flow model featuring coal seam gas extraction

9.8 Revised model: Maximum vertical flux prediction (MXQ)



CONTACT US

- t 1300 363 400 +61 3 9545 2176
- e csiroenquiries@csiro.au
- w www.csiro.au

AT CSIRO, WE DO THE EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help improve today – for our customers, all Australians and the world.

Our innovations contribute billions of dollars to the Australian economy every year. As the largest patent holder in the nation, our vast wealth of intellectual property has led to more than 150 spin-off companies.

With more than 5,000 experts and a burning desire to get things done, we are Australia's catalyst for innovation.

CSIRO. WE IMAGINE. WE COLLABORATE. WE INNOVATE.

FOR FURTHER INFORMATION

CSIRO Land and Water

- Chris Turnadge
- t +61 8 8303 8712
- e chris.turnadge@csiro.au
- w http://people.csiro.au/T/C/Chris-Turnadge

CSIRO Land and Water

- Dirk Mallants
- t +61 8 8303 8595
- e dirk.mallants@csiro.auw http://people.csiro.au/M/D/Dirk-Mallants

CSIRO Land and Water

- Luk Peeters
- t +61 8 8303 8405
- e luk.peeters@csiro.au
- w http://people.csiro.au/P/L/Luk-Peeters