Cross-validation in an iterative ensemble smoother: Stopping earlier for better

Xiaodong Luo, Norwegian Research Centre (NORCE), Norway

Paulo Ranazzi, Universidade de São Paulo, Brazil

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf
University of Stavanger

UNIVERSITY OF BERGEN

IFE

NORCE

Similarities and differences between reservoir data assimilation and supervised machine learning (regression)

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

Differences

	Reservoir data assimilation (RDA)	Supervised machine learning (SML)
Ideal goal	Uncover the ground-truth reservoir model $\boldsymbol{m}^{truth}$, with $\boldsymbol{d}^{obs} = \boldsymbol{g}(\boldsymbol{m}^{truth}) + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{C}_d)$ (ignoring model errors here)	Learn a ground-truth mapping $\boldsymbol{g}^{truth}$ so that $\boldsymbol{d}_j = \boldsymbol{g}^{truth}(\boldsymbol{m}_j) + \boldsymbol{\epsilon}$ for all j (often $\boldsymbol{\epsilon}$ absent)
Available information	Field data $\boldsymbol{d}^{obs}$ as observations, sample size = 1	$\{(\boldsymbol{d}_j, \boldsymbol{m}_j)\}_{j=1}^{N_s}$: Dataset containing IID input-output pairs $(\boldsymbol{d}_j, \boldsymbol{m}_j)$, sample size = $\boldsymbol{N}_s$
Practical strategy	Find one or more reservoir models $\{\boldsymbol{m}_j\}_{j=1}^{N_e}$ so that $\{\boldsymbol{g}(\boldsymbol{m}_j)\}_{j=1}^{N_e} \to \boldsymbol{d}^{obs}$	Choose a class of parameterized functions $\boldsymbol{g}(\cdot, \boldsymbol{\theta})$ so that $\boldsymbol{g}(\boldsymbol{m}_j, \boldsymbol{\theta}) \to \boldsymbol{d}_j \ \forall\, j$

Similarities and differences between reservoir data assimilation and supervised machine learning (regression)

Similarities

- Using the trick of data augmentation, redefine in SML

$$\boldsymbol{d}^{obs} \equiv \left[\boldsymbol{d}_1^T, \boldsymbol{d}_2^T, \ \dots, \boldsymbol{d}_{N_s}^T\right]^T$$

$$\boldsymbol{g}(\boldsymbol{m}, \boldsymbol{\theta}) \equiv \left[\boldsymbol{g}(\boldsymbol{m}_1, \boldsymbol{\theta})^T, \boldsymbol{g}(\boldsymbol{m}_2, \boldsymbol{\theta})^T, \dots, \boldsymbol{g}\left(\boldsymbol{m}_{N_s}, \boldsymbol{\theta}\right)^T\right]^T$$

- Also define a common form of the forward simulator

$$\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{m}, \boldsymbol{\theta})$$

for both RDA and SML problems.

- Remarks:
  - In SML: estimating $\boldsymbol{\theta}$ (parameters of SML model) but keeping $\boldsymbol{m}$ (model input) constant
  - In RDA: estimating $\boldsymbol{m}$ (e.g., petro-physical parameters) but keeping $\boldsymbol{\theta}$ (e.g., well configuration parameters model) constant

Similarities and differences between reservoir data assimilation and supervised machine learning (regression)

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

Similarities

- Both RDA and SML formulated as a minimum-average-cost (MAC) problem*

$$\underset{\{\boldsymbol{v}_j^{i+1}\}}{\operatorname{argmin}} \frac{1}{N_e} \sum_j L\left(\boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}}, \boldsymbol{c}\right), j = 1, 2, \ldots, N_e$$

$$L\left(\boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}}, \boldsymbol{c}\right) \equiv \frac{1}{2}\left(\boldsymbol{d}^{obs} - \boldsymbol{g}\left(\boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}}, \boldsymbol{c}\right)\right)^T \boldsymbol{C}_d^{-1}\left(\boldsymbol{d}^{obs} - \boldsymbol{g}\left(\boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}}, \boldsymbol{c}\right)\right) + \frac{\gamma^i}{2}\left(\boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}} - \boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}}\right)^T (\boldsymbol{C}_v^i)^{-1}(\boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}} - \boldsymbol{v}_{\boldsymbol{j}}^{\boldsymbol{i+1}})$$

  - In SML: variable $\boldsymbol{v} = \boldsymbol{\theta}$, constant $\boldsymbol{c} = \boldsymbol{m}$
  - In RDA: variable $\boldsymbol{v} = \boldsymbol{m}$, constant $\boldsymbol{c} = \boldsymbol{\theta}$
  - i: iteration index; j: ensemble member index
  - γ: regularization parameter
  - $\boldsymbol{C}_d$ / $\boldsymbol{C}_v$: sample error covariance matrix of observations and model variables, respectively

*Luo, X., Stordal, A. S., Lorentzen, R. J., & Nævdal, G. (2015). Iterative ensemble smoother as an approximate solution to a regularized minimum-average-cost problem: theory and applications. *SPE Journal*, *20*(05), 962-982.

Similarities and differences between reservoir data assimilation and supervised machine learning (regression)

National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

Similarities

- Iterative ensemble smoother (IES) provides an approximate solution to the MAC problem, in the following form:

$$\boldsymbol{v}_j^{i+1} = \boldsymbol{v}_j^i + \boldsymbol{K}^i \left(\boldsymbol{d}^{obs} - \boldsymbol{g}\left(\boldsymbol{v}_j^i, \boldsymbol{c}\right)\right), j = 1,2, \dots, N_e$$

$\boldsymbol{K}^i$: Kalman-gain-like matrix

- In RDA problems, IES often equipped with the so-called localization technique, so that

$$\boldsymbol{v}_j^{i+1} = \boldsymbol{v}_j^i + \left(\boldsymbol{T}\left(\ell^i\right) \circ \boldsymbol{K}^i\right) \left(\boldsymbol{d}^{obs} - \boldsymbol{g}\left(\boldsymbol{v}_j^i, \boldsymbol{c}\right)\right), j = 1,2, \dots, N_e$$

$\boldsymbol{T}$: tapering matrix

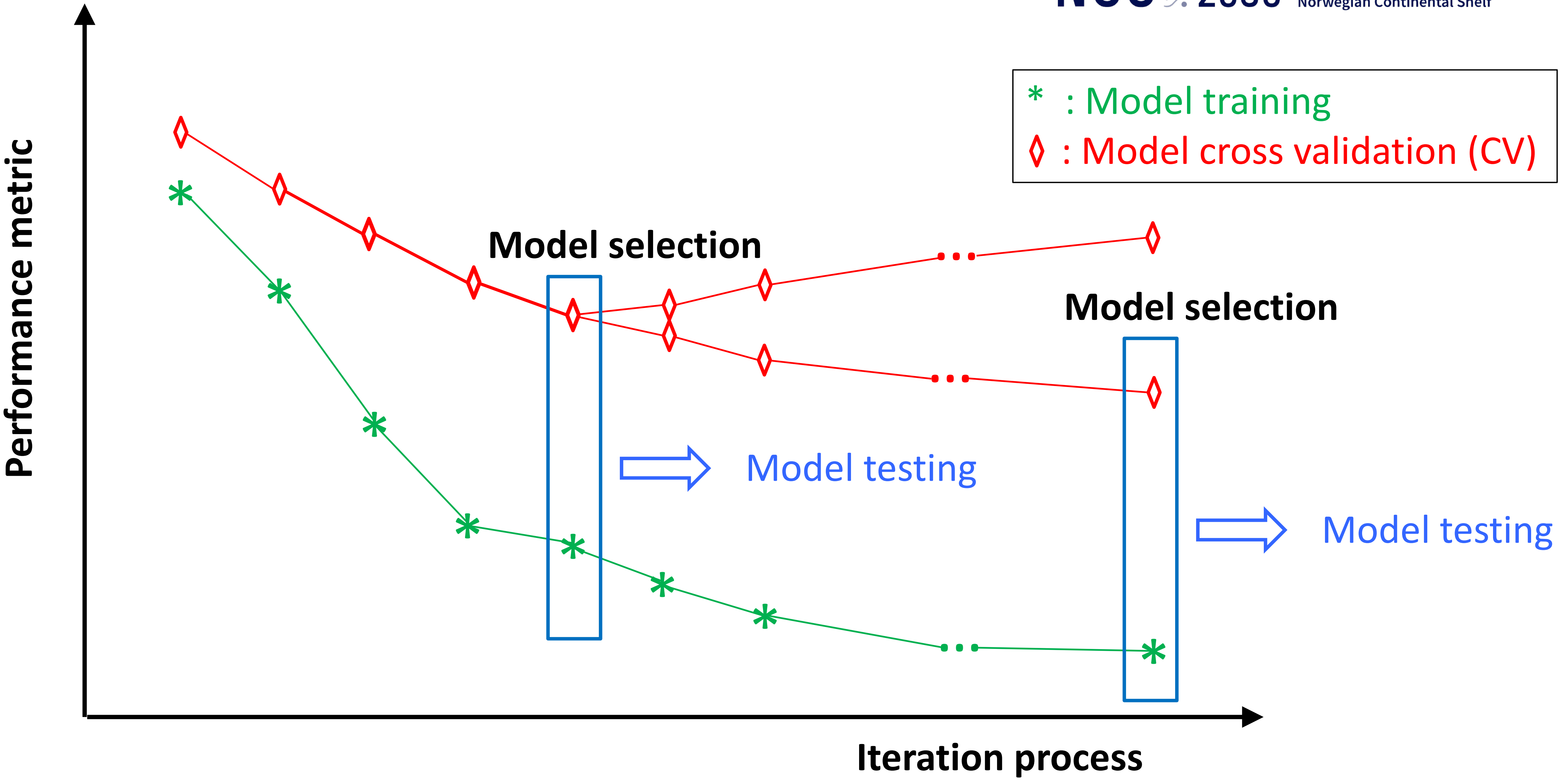
operator $\circ$: Schur product

ℓ^i: algorithmic hyper-parameter

Performance evaluation in SML

- Splitting the whole dataset $\boldsymbol{D} \equiv \{(\boldsymbol{d}_j, \boldsymbol{m}_j)\}_{j=1}^{N_s}$ into three disjoint sub-sets:
  - **Dataset** $\boldsymbol{D}^{tr} \equiv \{(\boldsymbol{d}_j, \boldsymbol{m}_j)\}_{j=1}^{N_s^{tr}}$ for model training
  - **Dataset** $\boldsymbol{D}^{cv} \equiv \{(\boldsymbol{d}_j, \boldsymbol{m}_j)\}_{j=1}^{N_s^{cv}}$ for model cross validation (CV)
  - **Dataset** $\boldsymbol{D}^{ts} \equiv \{(\boldsymbol{d}_j, \boldsymbol{m}_j)\}_{j=1}^{N_s^{ts}}$ for model testing
- Training and CV happening at the same time, testing after training and CV

Performance evaluation in SML



Performance evaluation in RDA

SML		RDA
Model training	⟺	History matching/Data assimilation/Inversion/Model calibration
Model cross validation (CV)		???
Model testing		Model QA, QC/Model diagnostics/Model criticism

Take-away messages:

- CV procedure typically absent in RDA
- In RDA algorithms like an IES, both continuous hyper-parameters (e.g., localization length scale) and discrete ones (e.g., stopping step) influencing model qualities and potentially causing overfitting
- Without CV, possibility of taking worse reservoir models

Performance evaluation in RDA

Essential features in SML		Situations in RDA

Performance evaluation in RDA

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

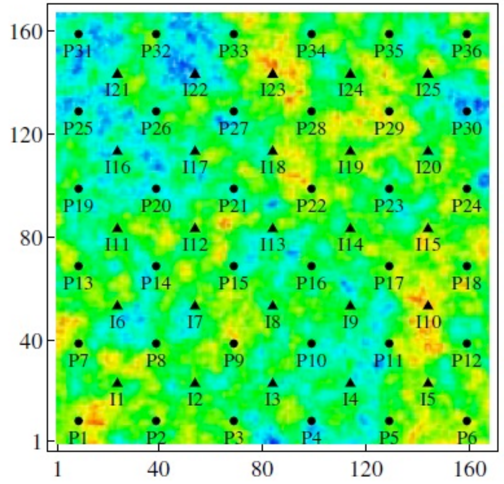
- The conditional independence in RDA making differences in CV/testing procedures from those in SML, e.g.,
  - randomly splitting observations may not work in general
  - K-fold CV may not work in general

- Rigorous treatment of the issue of conditional independence (vs. marginal independence) perhaps infeasible

- Empirical approaches possibly still useful for improving the performance of RDA, by reducing the marginal dependences among observations for training/CV/testing, e.g.,
  - Cross correlation (CC) for selecting CV data from a number of wells, with the corresponding CC between wells being the minimum ones
  - Domain knowledge (e.g., info. of reservoir compartmentalization/zonation/fluid dynamics)

- Subtle difference: The focus here on testing the CV procedure, not the selected model (so the performance metric of testing calculated at each iteration step)

Numerical example I: Synthetic 2D case*



Experimental settings	
Model information	167 x 167; 36 producers + 25 injectors; Uncertain parameters: PERMX
Production data used for RDA	WOPR, WWPR, WBHP, WWIR total number = 1098
RDA algorithm	IES with/without localization (ensemble size = 100)
Data for RDA/CV/testing	CV: data from 10 wells (180 data points) RDA: data from remaining wells (918 data points) Testing: Reference model
Performance metric	RDA: Average data mismatch (DM) in observation space CV: Average data mismatch (DM) in observation space Testing: Average root mean squared error (RMSE) in model space

*Chen, Y. and Oliver, D.S., 2010. Cross-covariances and localization for EnKF in multiphase flow data assimilation. *Computational Geosciences*, *14*(4), pp.579-601.

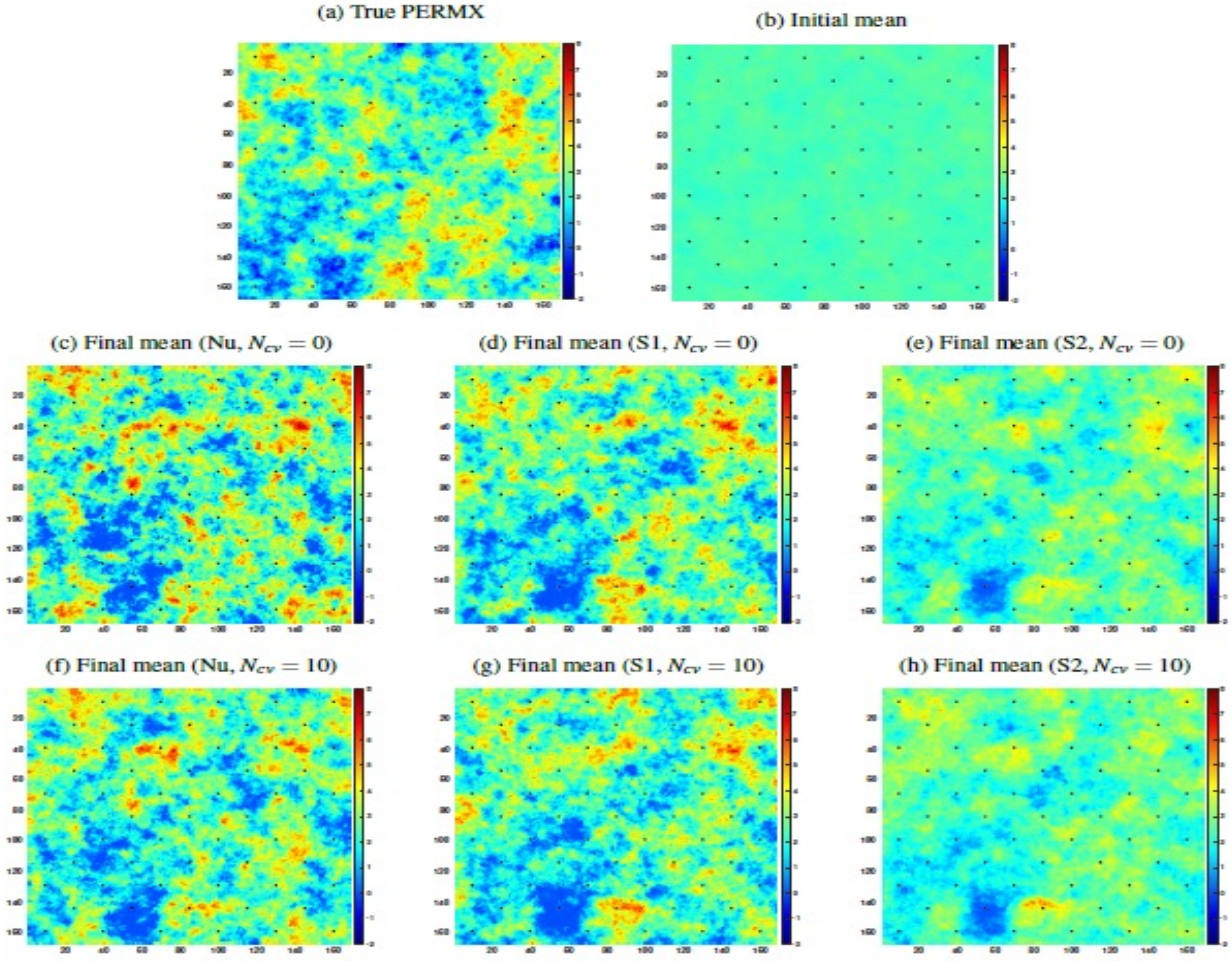
Numerical example I: Synthetic 2D case

National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

	No localization	Simple localization*	Sophisticated localization*
RDA (Average DM)			
CV (Average DM)			
Testing (RMSE)			

*Luo, X., Cruz, W. C., Zhang, X. L., & Xiao, H. (2023). Hyper-parameter optimization for improving the performance of localization in an iterative ensemble smoother. *Geoenergy Science and Engineering*, *231*, 212404.

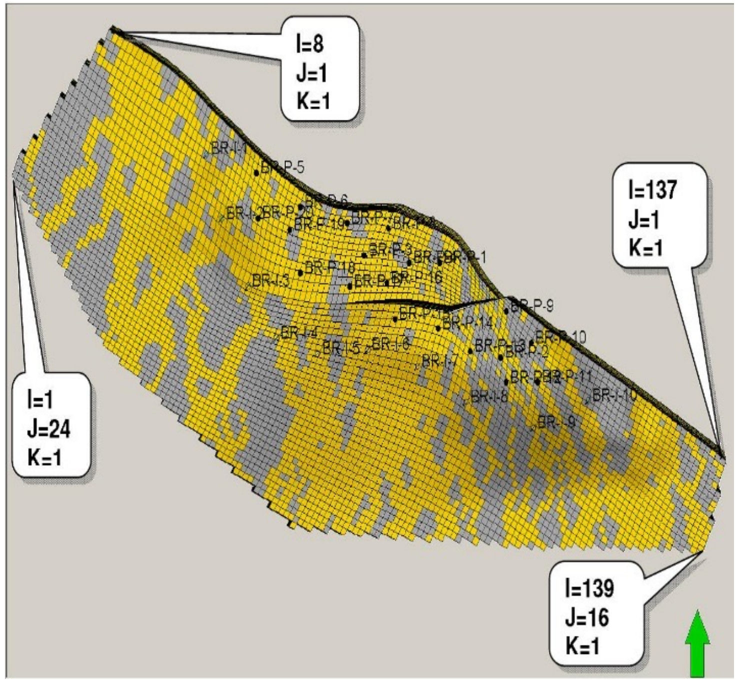
Figure 4: Mean PERMX maps of the final ensembles in the M-5Spots case, which are obtained by the IES at the suggested stopping steps in Table 3, with the "Nu", "S1", and "S2" localization strategies, and the corresponding number N_{cv} of CV wells being 0 and 10, respectively. For comparison, the PERMX map from the ground-truth model and the mean PERMX map of the initial ensemble are also present. In all the plots, the black dots indicate the locations of wells in the numerical reservoir model.



- Nu = No localization
  S1 = Simple localization
  S2 = Sophisticated localization
- N_{cv} = number of CV wells

Numerical example II: 3D Brugge benchmark

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf



Grid geometry of the Brugge field

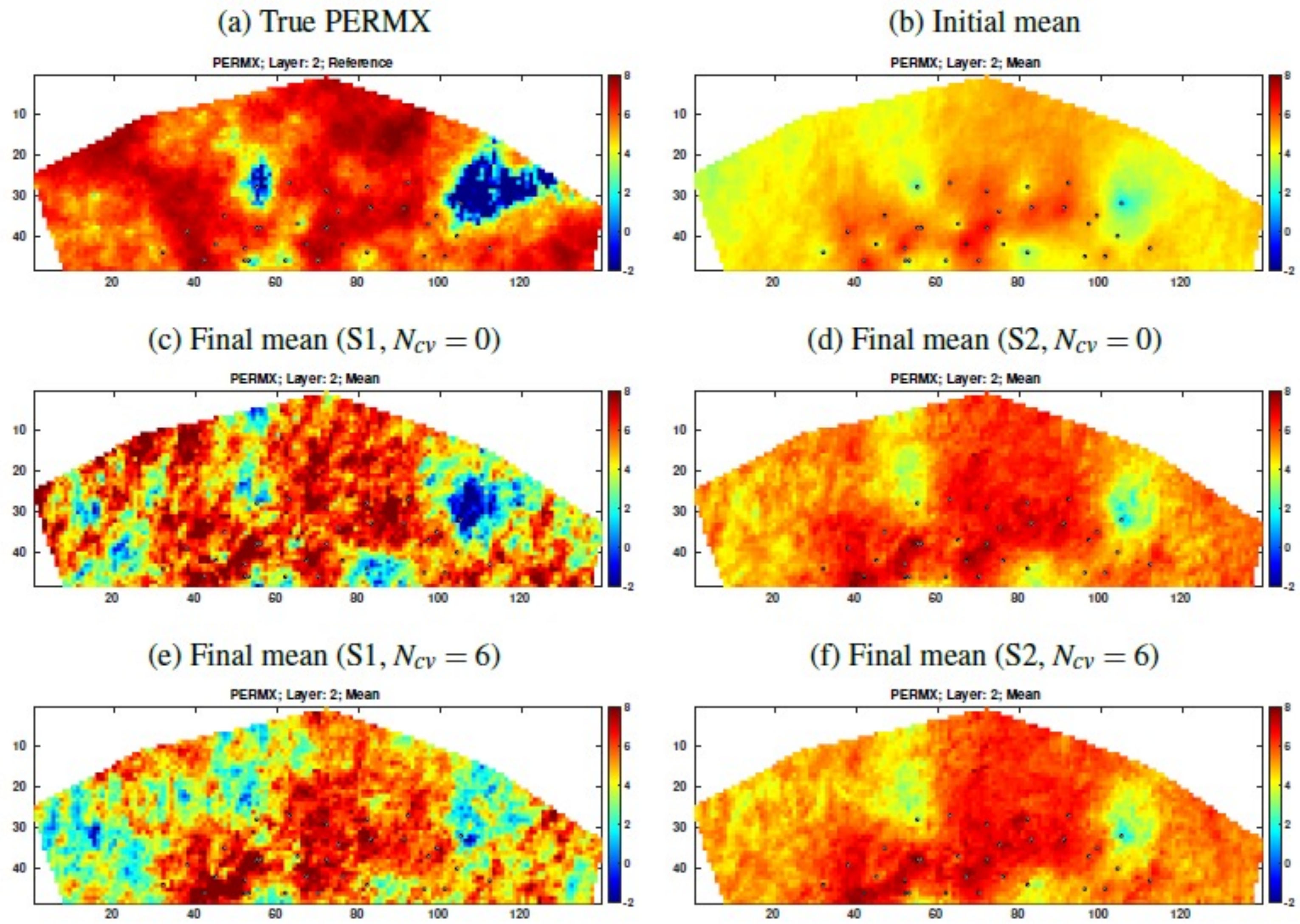
Experimental settings	
Model information	139 x 48 x 9; 20 producers + 10 injectors Uncertain parameters: PERMX, PERMY, PERMZ, PORO
Production data used for RDA	WOPR, WWCT, WBHP total number = 1400
RDA algorithm	IES with simple and sophisticated localization (ensemble size = 103)
Data for RDA/CV/testing	CV: data from 6 wells (300 data points) RDA: data from remaining wells (1100 data points) Testing: Reference model
Performance metric	RDA: Average data mismatch (DM) in observation space CV: Average data mismatch (DM) in observation space Testing: Average root mean squared error (RMSE) in model space

Numerical example II: 3D Brugge benchmark

National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

	Simple localization	Sophisticated localization
RDA (Average DM)		
CV (Average DM)		
Testing (RMSE)		

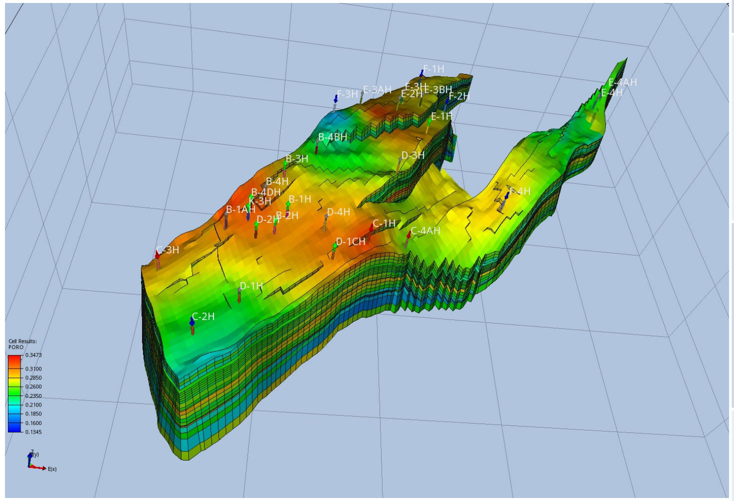
Figure 7: Mean PERMX maps of the final ensembles on Layer 2 of the Brugge model, which are obtained by the IES at the suggested stopping steps in Table 5, with the "S1" and "S2" localization strategies, and the corresponding number N_{cv} of CV wells being 0 and 6, respectively. For comparison, the PERMX map from the ground-truth model and the mean PERMX map of the initial ensemble are also present.



- S1 = Simple localization
  S2 = Sophisticated localization
- N_{cv} = number of CV wells

Numerical example III: Norne field case

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf



Experimental settings	
Model information	46 x 112 x 22; 22 producers + 14 injectors Uncertain parameters: PERMX, PORO, NTG + regional/zonal/scalar parameters
Production data used for RDA	WOPRH, WWPRH, WGPRH total number = 7260
RDA algorithm	IES using sophisticated localization, with/without CV (ensemble size = 100)
Data for RDA/CV/testing	CV: data from wells 'B-1BH', 'B-1H', 'B-2H’ and 'B-3H' (1320 data points) RDA: data from remaining wells (5940 data points) Testing: RFT data from well 'C-4AH’ (26 data points)
Performance metric	RDA: Average data mismatch (DM) in observation space CV: Average data mismatch (DM) in observation space Testing: Average data mismatch (DM) in observation space

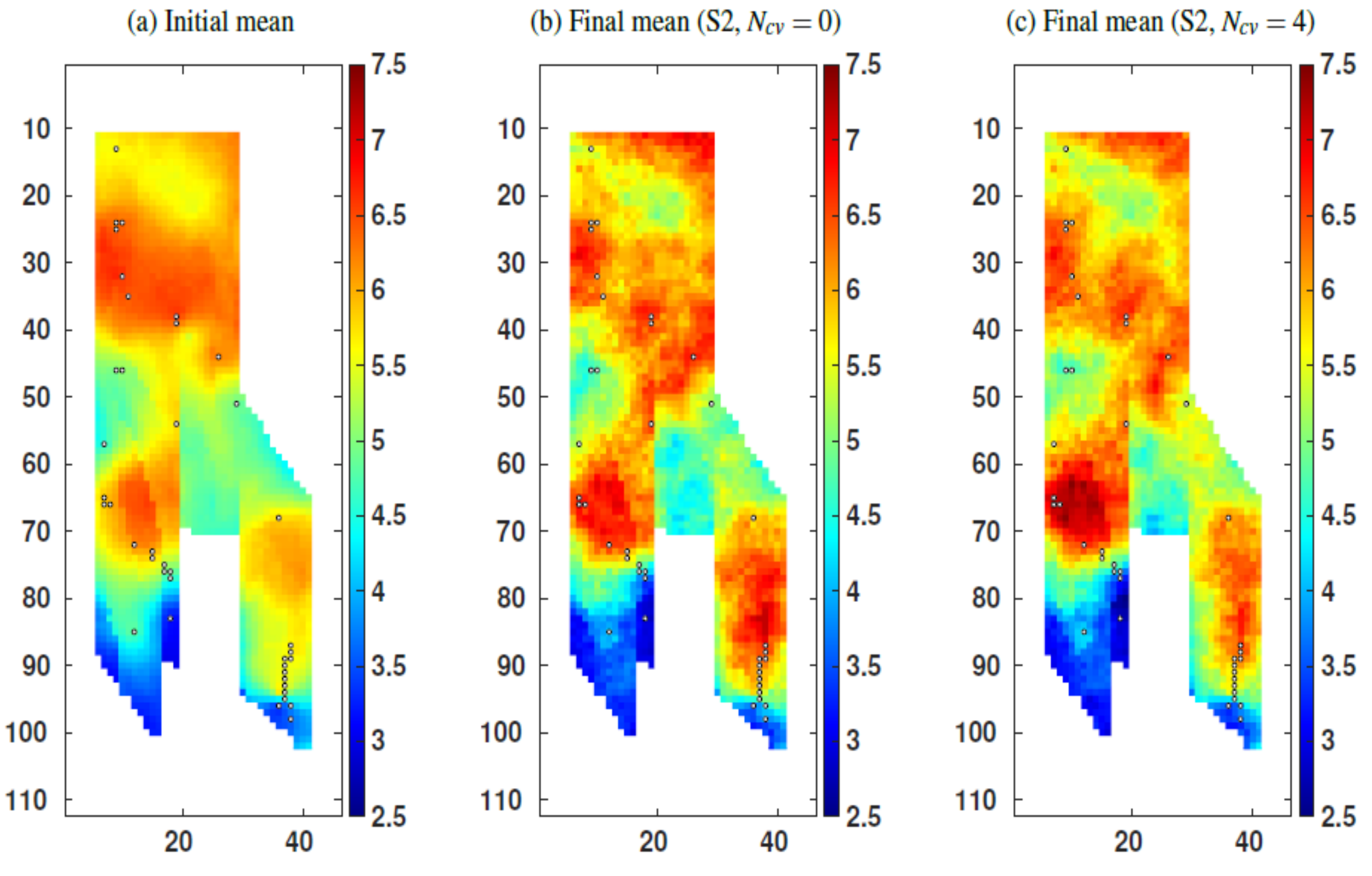
Numerical example III: Norne field case

National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

	Sophisticated localization no CV	Sophisticated localization with CV
RDA (Average DM)		
CV (Average DM)		
Testing (Average DM)		

Figure 9: Mean PERMX maps of the final ensembles on Layer 2 of the Norne field model, with the "S2" localization strategy and the corresponding number N_{cv} of CV wells being 0 and 4, respectively. For comparison, the mean PERMX map of the initial ensemble is also included. In addition, in all the maps, the white dots indicate the locations of various wells perforating Layer 2. In particular, the injector "C-4AH" containing the RFT data corresponds to the white dot at the coordinate (29, 51).

National Centre for
Sustainable Subsurface Utilization of the
Norwegian Continental Shelf



- S2 = Sophisticated localization
- N_{cv} = number of CV wells

Discussion and conclusion

- Similarities and differences identified in SML (regression) and RDA problems

- Similarities: performance evaluation procedures in SML => those in RDA
  - Cross validation (CV) typically missing in RDA

- Differences: non-straightforward extensions of CV and testing procedures from SML to RDA

- Empirical approach used to divide wells into distinct groups for model calibration/CV/testing
  - CV helping mitigate the problem of overfitting in synthetic case studies
  - CV also identifying a possible way to further improve RDA performance in the Norne field case
  - CV based criterion applicable to real-world problems, offering the possibility of stopping earlier for better

- More questions need to be answered
  - More rigorous way to split observation data for RDA/CV/testing?
  - Impacts of model errors on CV?
  - etc.

Acknowledgement

XL acknowledges the Research Council of Norway and the industry partners of NCS2030 – RCN project number 331644 – for their support.

User partners & observers

equinor

vår energi

Funded by
The Research
Council of Norway

AkerBP

OKEA

HALLIBURTON
Landmark

NORWEGIAN PETROLEUM
DIRECTORATE

DNO
EST - 1971

wintershall dea

Schlumberger

petoro

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

University of Stavanger

IFE

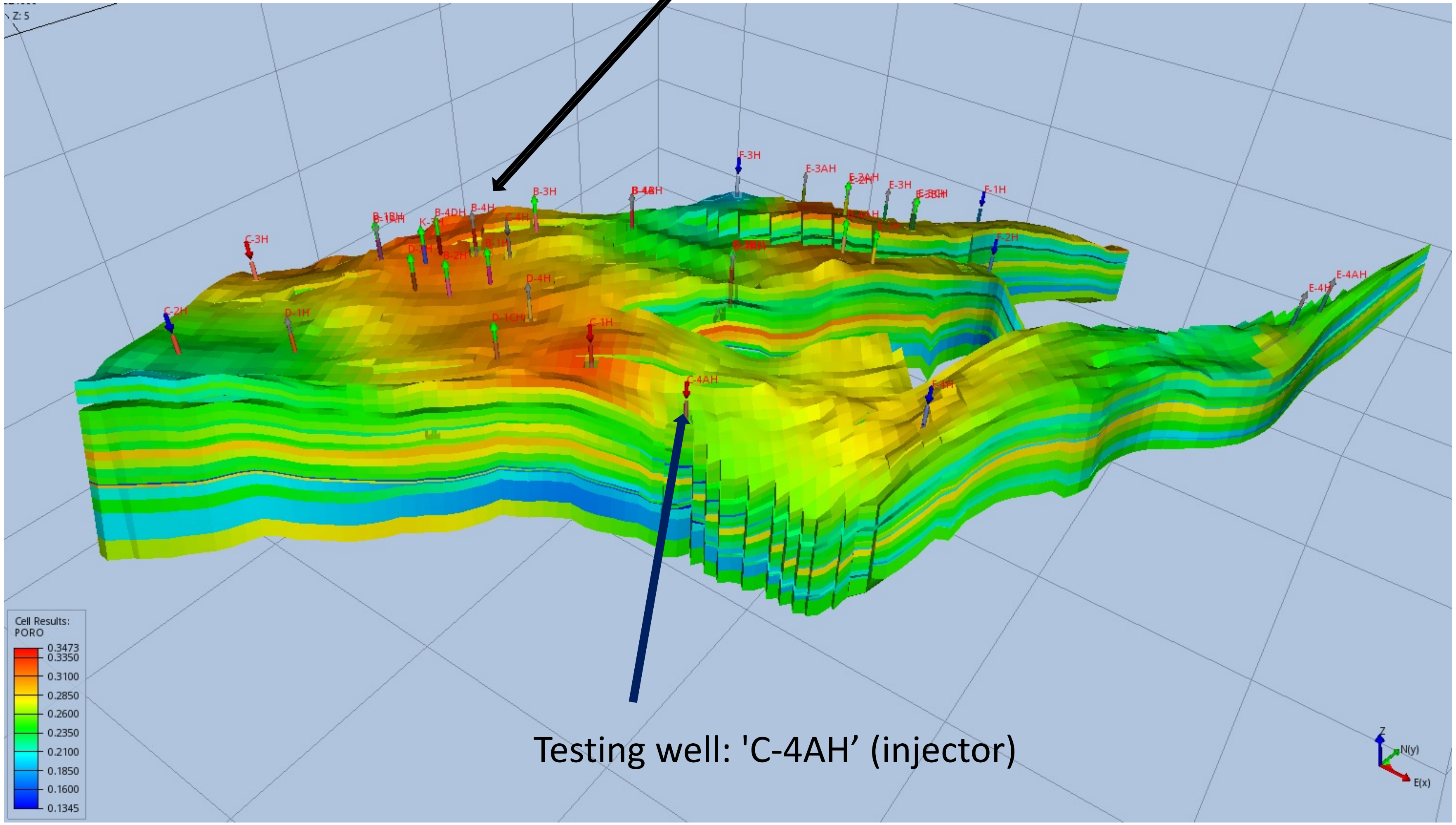
UNIVERSITAS BERGENSIS

NORCE

Numerical example III: Norne field case

NCS 2030 National Centre for Sustainable Subsurface Utilization of the Norwegian Continental Shelf

CV wells
'B-1BH', 'B-1H', 'B-2H' and 'B-3H' (producers)



Testing well: 'C-4AH' (injector)

- At least 3 different versions of reservoir models used in geophysical reports: 2002, 2006, and one earlier than 2002
- 2002 reservoir model used in RDA
- Many available geophysical reports inconsistent with the 2002 model
- Ending up with only one useful geophysical report for well 'C-4AH', from which RFT data extracted