XIX International Conference on Water Resources
CMWR 2012
University of Illinois at Urbana-Champaign
June 17-22,2012

IMPROVING GROUNDWATER FLOW MODEL PREDICTION USING COMPLEMENTARY DATA-DRIVEN MODELS

Tianfang Xu[1*], Albert J. Valocchi[1], Jaesik Choi[2] and Eyal Amir[2]

[1]Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[*]e-mail: txu3@illinois.edu

Key words: Uncertainty, data-driven models, machine learning

Summary. Current analyses of groundwater flow and transport typically rely on a physically-based model (PBM), which in its essence is a simplification of reality and is thus subject to error and uncertainty from multiple sources, such as parameter error, conceptual model error, and input data error. The model uncertainty can be difficult to quantify, and is propagated to the prediction. In this study, complementary data-driven models (DDMs) are used to improve prediction of PBM. Three machine learning techniques, instance-based weighting (IBW), support vector regression (SVR) and clustering are employed to build DDMs. We use a real-world case study to demonstrate that the framework effectively reduces the head prediction error of a regional groundwater flow model.

1 INTRODUCTION

The inherent uncertainty in groundwater modeling has been widely recognized in the literature.[8,9] Model uncertainty comes from multiple sources, including the model structural error due to the misrepresentation and simplification of site characteristics and hydrologic processes, uncertainty in parameter values, inaccuracy of input data as well as measurement error. As a result, the prediction made by the model contains both systematic and random error, which is typically represented as the mismatch between the observed quantity of interest and its simulated counterpart. A common practice to achieve better prediction by reducing parameter uncertainty is to use regression-based inverse method, or calibration. However, as suggested by Doherty,[5] in the process of calibration, the model parameters might be over-adjusted to compensate for the model structure defects, which, according to Beven,[1] could be a significant cause of prediction uncertainty. The perspective of equifinality,[1] that satisfactory agreement with observations can be achieved by many potential model and parameter combinations rather than

a single best calibrated model, gives rise to Monte-Carlo based methods that allow for comparing model structures and exploration in the parameter space. For groundwater models which usually require relatively long running times, the computational expense becomes a drawback of these methods.

Demissie et al.[3,4] proposed a complimentary framework where separately-developed data-driven models (DDMs) were used to correct the head prediction error of the physically based MODFLOW model (PBM) in a hypothetical case study. This framework suggests that, if structure (i.e. temporal and spatial correlation, etc.) is presented in the discrepancy between the MODFLOW model and the real groundwater system, then the discrepancy can be learned by data-driven models that are trained on the historical errors of the PBM. This framework is not restricted to any particular type of model errors and hence is advantageous if the sources of prediction error are multiple and not easily identifiable. In addition, it does not invoke any statistical assumption about the error distribution. Furthermore, unlike calibration and Monte-Carlo based approaches, the framework only runs the PBM once, thus making it suitable for PBMs with long running time. In this study, we improve the DDMs, and introduce clustering into this framework to make it more robust, flexible and computationally efficient, as demonstrated in a real-world case study of a regional groundwater flow model.

2 METHODOLOGY

The complimentary framework, as shown in Figure 1, recognizes the uncertainty of the PBM as a lumped bias, then models this bias with DDMs based on machine learning techniques. When forecasting, the prediction of the PBM is adjusted with the bias predicted by the DDMs. The remaining part of this section is a short review of machine learning techniques used to build DDMs.

Instance-based weighting (IBW) is an extension of the widely-used k-Nearest Neighbor method (kNN) by introducing a weighting function

$$w_{\mathbf{x}'|\mathbf{x}_i} = \alpha exp(-||\mathbf{x}' - \mathbf{x}_i||^2/p^2), \tag{1}$$

where $w_{\mathbf{x}'|\mathbf{x}_i}$ denotes the weight of i-th neighbor, α is a scaling factor to ensure $\sum_{i=1}^{n} w_{\mathbf{x}'|\mathbf{x}_i} = 1$, and p is a parameter optimized by cross validation. For a query $\boldsymbol{x'}$, IBW first finds its n nearest neighbors in the training set. The estimation of the query ($\hat{\epsilon}$) is calculated as a weighted average of the neighbors' target values ($\epsilon(\mathbf{x}_i)$)

$$\hat{\epsilon}(\mathbf{x}') = \sum_{i=1}^{n} w_{\mathbf{x}'|\mathbf{x_i}} \epsilon(\mathbf{x}_i). \tag{2}$$

In the case study, Eqn. (1) proved to perform better than the inverse-distance weighting used by Demissie et al.[3]

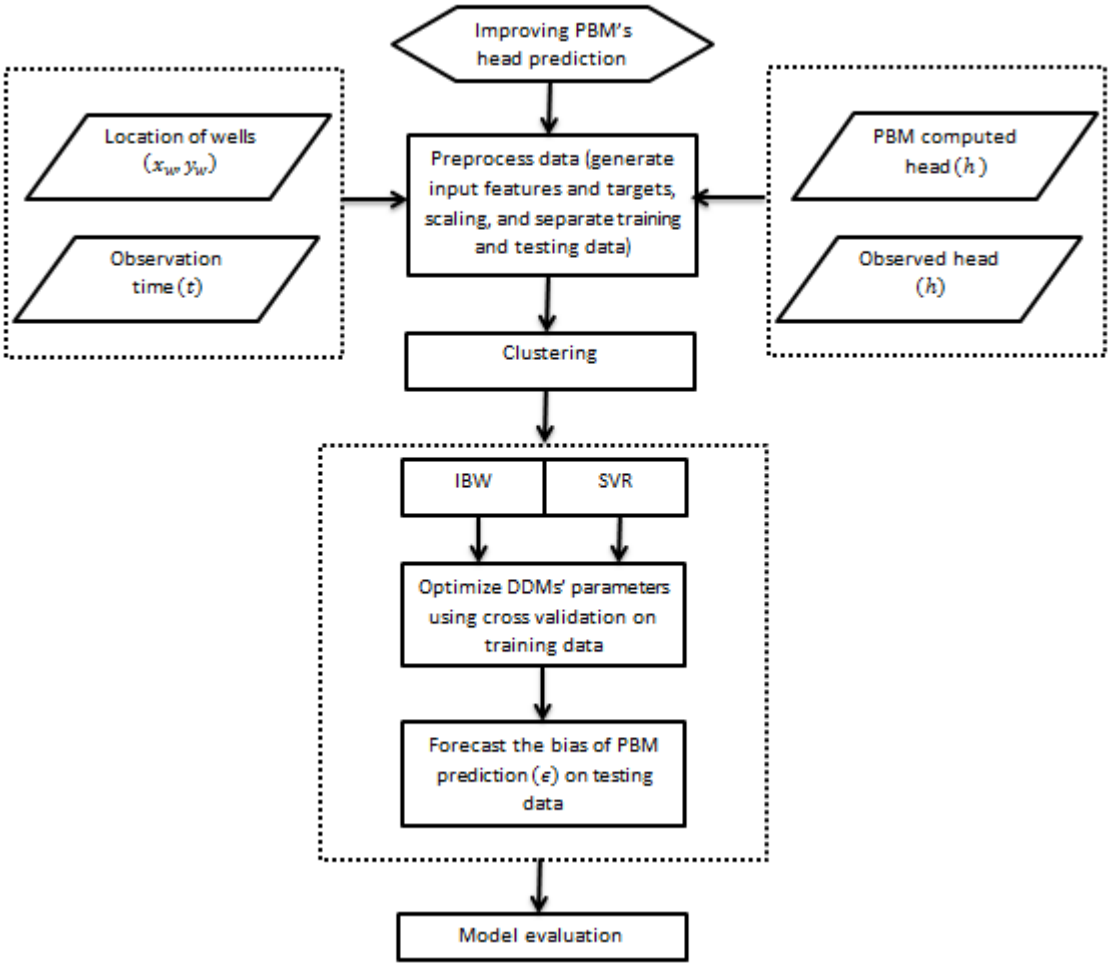


Figure 1: The framework of using complimentary DDMs to improve head prediction of PBM.

Support vector regression (SVR) comprises a robust class of learning algorithm,[10] featuring 1) use of kernels to transform the feature space into higher dimensional Hilbert space; 2) introduction of insensitive loss function and regularization to prevent overfitting. Given appropriate hyperparameters, SVMs provided good results on benchmark datasets.[10] In this study, the Gaussian radial basis function kernel was used, and the hyperparameters were chosen by five-fold cross validation and following the recommendation of Cherkassky and Ma[2] .

Clustering is a class of unsupervised data mining algorithms that partition data into groups with the goal of maximizing the similarity of data within the same group and minimizing the similarity among groups. The agglomerative hierarchical clustering was

used in this study. For more information about this algorithm, the readers are referred to Hastie et al.[6] and Mitchell.[7]

3 CASE STUDY

3.1 RRCA model and data overview

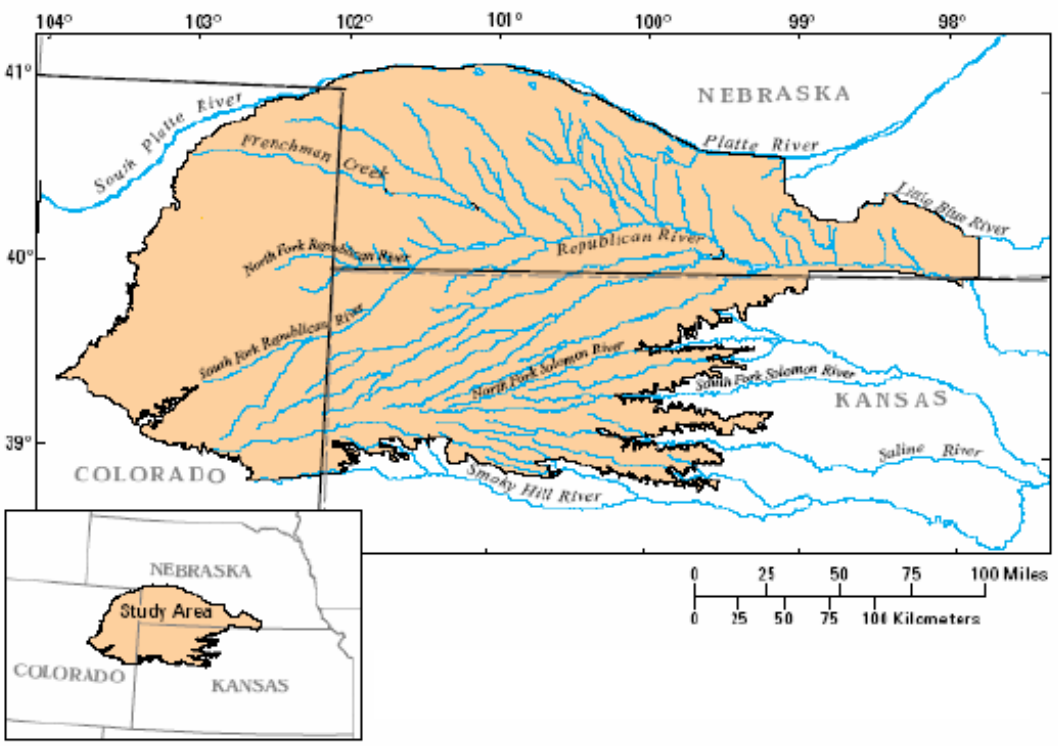


Figure 2: Republican River Basin covering portions of eastern Colorado, northwest Kansas and southwest Nebraska.

The framework described in the previous section is applied to the Republican River Compact Association (RRCA) Model, a regional groundwater flow MODFLOW model of the entire Republic River Basin (Figure 2). The model was calibrated based on head measurements at over 10,000 wells and baseflow observations at 65 gages from Jan.1918 to Dec.2000. Since 2000, it has been run each year using new input data. Head predictions until 2007 are available via the RRCA official website (www.republicanrivercompact.org). The dataset used in this study includes over 300,000 water level measurements from 1918 to 2007 at all 3,078 wells within the model boundary that have no fewer than 10 observations and absolute mean error less than 100 ft.

3.2 DDMs implementation

Residual analysis revealed temporal and spatial correlation in the prediction error (ϵ) of the MODFLOW model's head prediction ($\hat{h}$) (results are not included here). This satisfies the premise of the effectiveness of the DDMs.[3] The framework first built DDMs to forecast the prediction error (ϵ), then output the updated head prediction $h^{new} = \hat{h} + \epsilon$. The DDMs took as inputs the location of the wells where the head measurements were taken as well as the head computed by the MODFLOW model. The DDMs were trained on a total of 301,861 historical data (1918-2000), and validated on 10,161 heads during the prediction period (2001-2007).

Using the clustering technique described in section 2, the 3,078 wells in the database were clustered according to their spatial locations, and first and second moments of the head error. The dataset was then divided according to the well clusters into 10 subsets. Each subset was comprised of a training dataset containing data during the calibration period and a validation dataset during the prediction period. One IBW model and one SVR machine were developed for each subset. The benefit of clustering in this study was two fold: First, residual analysis showed local patterns within the dataset. Rather than developing a global DDM, we tuned the parameters of "localized" DDMs based on the data within that cluster to allow for additional flexibility and robustness. Second, dividing the dataset containing over 300,000 samples into smaller subsets improved the computation efficiency, and made model selection by cross validation more feasible.

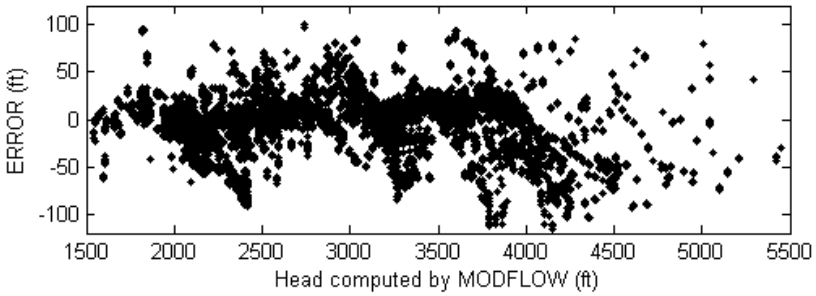
4 RESULTS

To be concise, only the global performance averaged among subsets is reported. The mean error (ME) and root mean squared error (RMSE) of the head prediction during 2001-2007 by MODFLOW ($\hat{h}$) and DDMs-corrected MODFLOW (h^{new}) are shown in Table 1. Both IBW and SVR effectively improved the accuracy of head prediction in the MODFLOW model, reducing the RMSE by over 80%. SVR almost eliminated the global bias, reducing the ME to almost zero. Figure 3(a) shows the head forecast error of MODFLOW during 2001 to 2007, while figures 3(b) and (c) plot the error after correcting with IBW and SVR respectively. The magnitude of residual significantly shrinks after DDMs updating, and the bias is largely removed.

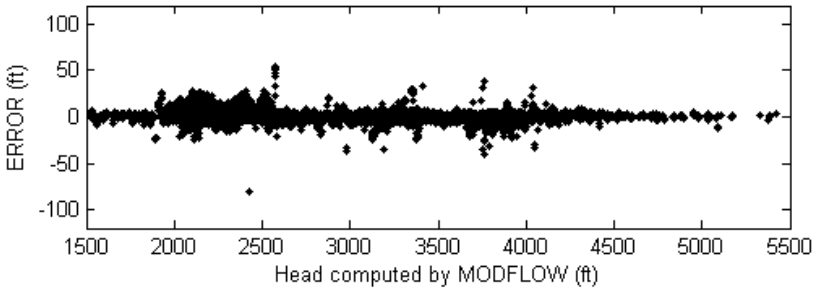
Figure 4 presents the hydrographs of several representative wells forecasted by MODFLOW and by the complimentary framework. In general, the DDMs significantly improve the prediction accuracy. For those wells where the MODFLOW model predicts the trend of water level correctly but has bias, the DDMs "shift" the MODFLOW prediction to correct the bias (Figure 4 (a)). In cases where MODFLOW makes an incorrect prediction of the trend, the DDMs can still compensate, as shown in (b), (c) and (d), however the

Table 1: The error of head prediction before and after corrected by DDMs.

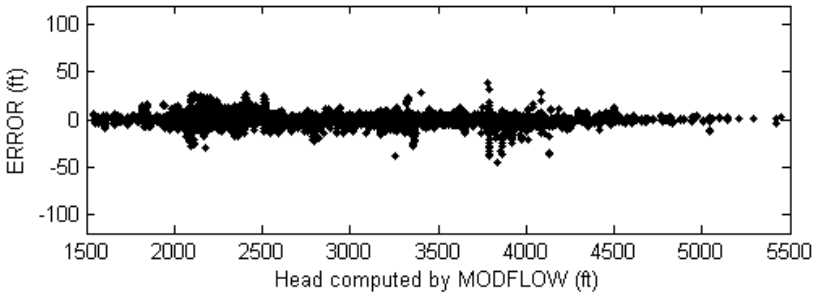
	MODFLOW	MODFLOW+IBW	MODFLOW+SVR
ME (ft)	-2.29	0.81	0.03
RMSE (ft)	30.23	5.32	5.16



(a)



(b)



(c)

Figure 3: Residual plot before (a) and after correcting using IBW (b) and SVR (c).

effectiveness of DDMs for trend correction varies for each well.

In general, both DDMs yield relatively smooth prediction compared with the fluctuating measurements, because the DDMs do not account for measurement error. It is also worth noting that SVR tends to yield a smoother prediction surface than IBW. The latter model is highly localized, producing a complex prediction surface, especially when few neighbors are used to calculate a query. The SVR, on the other hand, is a global model (within each subset) with regularization to keep the prediction surface smooth and less complex.

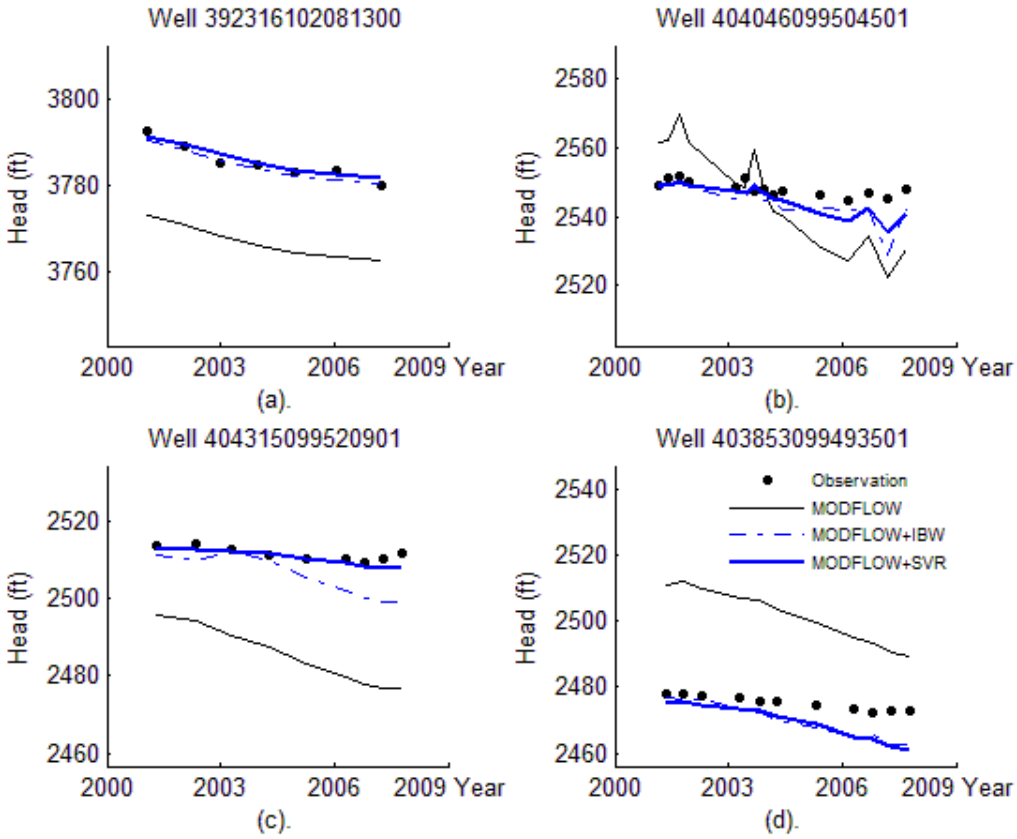


Figure 4: Measurements, MODFLOW predicted and DDMs-updated hydrographs at representative well locations during the prediction period.

5 CONCLUSIONS

This paper presents an extension of the complimentary data-driven framework developed by Demissie et al.[4] It is assumed that DDMs can discover the PMB's defects (uncertainty in parameters, model structure and input data) via machine learning techniques from the historical error, and can then estimate the tendency of the PBM to make

biased forecasts. We improved the DDMs and introduced clustering to make the framework more robust and efficient. The strength of the presented approach lies in: 1) it works for error from multiple sources, and does not require assumptions about the error distribution; 2) it is computationally efficient compared with calibration and Monte-Carlo simulation based models, especially for large, complex PBMs; and 3) it is straightforward to assimilate newly available data.

This framework is shown to successfully improve the head prediction of a regional groundwater flow model. The magnitude and bias of the prediction uncertainty (quantified by error) are sufficiently reduced. Extensions of this approach include applying the DDMs to other types of prediction and PBMs, and using them as tools to assess the prediction uncertainty.

REFERENCES

[1] K. Beven and A. Binley. The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6(3):279–298, 1992.

[2] V. Cherkassky and Y. Ma. Practical selection of *svm* parameters and noise estimation for *svm* regression. *Neural Networks*, 17(1):113–126, 2004.

[3] Y.K. Demissie. *Data-driven models to enhance physically-based groundwater model predictions*. PhD thesis, University of Illinois at Urbana-Champaign, 2008.

[4] Yonas K. Demissie, Albert J. Valocchi, Barbara S. Minsker, and Barbara A. Bailey. Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. *Journal of Hydrology*, 364(3-4):257–271, 2009.

[5] J. Doherty and S. Christensen. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research*, 47, 2011.

[6] Trevor. Hastie, R. Tibshirani, and JH Friedman. *The elements of statistical learning*. Springer, 2001.

[7] T.M. Mitchell. Machine learning. wcb. *Mac Graw Hill*, page 368, 1997.

[8] Catherine Moore and John Doherty. The cost of uniqueness in groundwater model calibration. *Advances in Water Resources*, 29(4):605 – 623, 2006.

[9] SP Neuman and PJ Wierenga. A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites. university of arizona. Technical report, Report NUREG/CR-6805, 2003.

[10] V. Vapnik. Statistical learning theory, 1998.