Université du Québec
Institut National de la Recherche Scientifique
Centre Eau Terre Environnement

Simulation du niveau des eaux souterraines à l'aide de modèles hybrides produits par *transformation par ondelettes* et *décomposition par mode empirique d'ensemble*

Par

Ramin Bahmani

Mémoire présenté pour l'obtention du grade de
Maître ès sciences (M.Sc.)
en sciences de l'eau

Jury d'évaluation

Président du jury et examinateur interne	André St-Hilaire INRS-ETE
Examinateur externe	Younes Alila University of British Columbia
Directeur de recherche	Taha BMJ Ouarda INRS-ETE

AVANT-PROPOS

Ce mémoire comprend deux parties. La première est consacrée à la synthèse. Elle présente brièvement l'introduction, les méthodes utilisées, l'étude de cas ainsi que les résultats et les conclusions. La seconde partie présente trois articles dans lesquels sont fournies des informations détaillées sur ce mémoire.

Le titre et les auteurs de l'article sont :

1. “Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform”

Auteurs: Ramin Bahmani, Abazar Solgi, Taha B. M. J. Ouarda

Abazar Solgi: Contribution des données de l’étude et contribution à la discussion des résultats et de leurs impacts.

Taha B. M. J. Ouarda: Contribution à l'élaboration de la méthodologie, à la revue de littérature, au fondement théorique, à la production et à l'interprétation des résultats, ainsi qu'à la rédaction de l'article.

2. “Groundwater level simulation using Gene Expression Programming and M5 Model Tree combined with Ensemble Empirical Mode Decomposition”

Auteurs: Ramin Bahmani, Taha B. M. J. Ouarda

Taha B. M. J. Ouarda: Contribution au développement de la méthodologie, à la revue de la littérature, aux fondements théoriques, à l'interprétation des résultats et à la rédaction de l'article.

3. “A Discussion on “The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series” by Du et al. (2017)”

Auteurs: Ramin Bahmani, Taha B. M. J. Ouarda , Abazar Solgi

Taha B. M. J. Ouarda: Contribution au développement de la méthodologie, à la revue de littérature, au fondement théorique, à la production et à l'interprétation des résultats et à la rédaction de discussions.

Abazar Solgi: Contribution des données de l’étude et contribution à la discussion des résultats.

REMERCIEMENTS

Je voudrais remercier mon directeur, le professeur Taha Ouarda, pour ses conseils et son encadrement. Ma gratitude va également à tous les membres du jury pour avoir accepté de juger ma thèse. Je rends grâce à feu mon père et je voudrais exprimer mes remerciements à ma famille: Huorie, Bahram, Maryam, Shahram, Shahryar et à tous mes amis: Abazar, Zam-Zam, Mehrdad, Mostafa, Amina, Adnan, Ehsan, Shitanshu qui m`ont toujours soutenu dans les moments difficiles.

RÉSUMÉ

Afin d'améliorer la connaissance et la gestion du stress hydrologique, il faut revoir le modèle de simulation du niveau des eaux souterraines pour plus de précision. Dans la présente recherche, les modèles *Gene Expression Programming (GEP)* et *M5 model tree (M5)* sont utilisés pour simuler les niveaux mensuels des eaux souterraines. Ces modèles ont été combinés à deux méthodes, *Wavelet transform (WT)* et *Ensemble Empirical Mode Decomposition (EEMD)*, pour produire des modèles hybrides: *Wavelet-GEP*, *Wavelet-M5*, *EEMD-GEP* et *EEMD-M5*. Les données du mois en cours correspondant au niveau des eaux souterraines, à la température et aux précipitations de trois puits d'observation et d'une station météorologique sont utilisées pour simuler le niveau des eaux souterraines pour le mois d'après. Les résultats indiquent que les modèles hybrides produits avec *WT* sont plus précis que les modèles *GEP* et *M5*. Le *GEP* combiné avec *WT* est connu comme le modèle le plus précis, alors que la combinaison des modèles *M5* et *EEMD* n'est pas recommandée pour la simulation en raison de sa faible performance.

Mots-clés: Ensemble Empirical Mode Decomposition, Groundwater level, Gene Expression Programming, Hybrid model, M5 Model tree, Wavelet transform.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

Adaptive Neuro-Fuzzy Inference System	ANFIA
Artificial Neural Network	ANN
Biais moyen	BIAS
Biais moyen relatif	rBIAS
Classification et de régression	CART
Coefficient de determination	R^2
Discrete Wavelet Transform	DWT
Empirical Mode Decomposition	EMD
Ensemble EMD	EEMD
Gene Expression Programming	GEP
Intelligence artificielle	IA
Least Square Support Vector Machine	ASSVM
M5 model tree	M5
Mode intrinsèque	IMF
Multivariate Adaptive Regression Splines	MARS
Neuro-Fuzzy	NF
Niveau d'eaux souterraines	ES
Racine carrèe de l'erreur quadratique moyenne relative	rRMSE
Racine carrée de l'erreur quadratique moyenne	RMSE
Supportive Vector Machine	SVM
Wavelet transform	WT

PARTIE I: SYNTHÈSE

1. INTRODUCTION

Les travaux de recherche présentés dans ce mémoire concernent la simulation du niveau d'eau souterraine (ES). Nos trois articles sont résumées dans cette partie.

Les eaux souterraines constituent l'une des principales sources d'approvisionnement en eau pour l'homme et constituent une variable importante pour les projets hydro-environnementaux (Nourani and Mousavi, 2016). L'étude des ES fournit aux chercheurs des informations pour gérer les ressources en eau (Barzegar et al., 2017). Il est donc nécessaire, il est nécessaire de simuler les ES pour étudier leur disponibilité pour les activités humaines.

Les modèles physiques et mathématiques ont été pris en compte dans de nombreuses études pour la simulation des ES. Ces modèles ont été largement utilisés, mais ils n'ont pas toujours fourni toutes les informations requises concernant la simulation des niveaux de ES (Nayak et al., 2006). Pour simuler les ES, les modèles physiques requièrent beaucoup de données de haute précision et ne conduisent pas toujours à de bonnes performances (Nayak et al., 2006; Shiri et al., 2013). Les modèles déterministes ont également besoin de beaucoup d'informations et peuvent mener à des équations difficiles à résoudre.

Pour adresser les limites des modèles physiques et mathématiques, un grand nombre de chercheurs ont utilisé des modèles d'intelligence artificielle (IA). Ces modèles ont relevé une capacité élevée de simulation des ES (Coppola et al., 2003; Lallahem et al., 2005; Nayak et al., 2006; Sahoo et al., 2005). Parmi une gamme relativement large de modèles d'intelligence artificielle, le "*Gene Expression Programming (GEP)*" et le "*M5 model tree (M5)*" ont montré une grande capacité à simuler des phénomènes hydrologiques tels que des ES (Shiri and Kisi, 2011; Shiri et al., 2013).

Shiri and Kişi (2011) ont étudié l'aptitude du GEP et du "*Adaptive Neuro-Fuzzy Inference System (ANFIS)*" à la prévision des fluctuations des ES et ont constaté que le GEP donnait de meilleurs résultats par rapport à l'ANFIS. Shiri et al. (2013) ont simulé les fluctuations des ES en utilisant les modèles GEP, "*Supportive Vector Machine (SVM)*" et ANFIS. Les auteurs ont examiné différentes combinaisons de valeurs de précipitation,

des ES et d'évaporation et ont conclu que le GEP conduit à de meilleures performances par rapport aux autres.

Solomatine and Xue (2004), pour la prévision des crues, et Bhattacharya and Solomatine (2005), pour la relation niveau d'eau-débit, ont appliqué M5 et "*Artificial Neural Network (ANN)*" pour comparer la précision des modèles. Les études ont montré que M5 avait la même précision que les ANN, mais avec un certain nombre d'avantages, tels que la transparence, la vitesse de formation élevée et la convergence rapide. Kisi (2015) et Kisi (2016) ont utilisé "*Least Square Support Vector Machine (LSSVM)*", "*Multivariate Adaptive Regression Splines (MARS)*" et M5 pour analyser la précision des modèles de simulation d'évaporation et d'évapotranspiration, respectivement. L'auteur a défini différents scénarios pour analyser la précision des modèles et a conclu que chaque modèle peut avoir une bonne précision en modifiant les variables d'entrée.

Malgré la large application des modèles, ceux-ci risquent de ne pas être aussi efficaces pour modéliser des séries chronologiques comportant de fortes fluctuations non stationnaires (Nourani et al., 2009a). Ainsi, certaines méthodes ont été développées pour améliorer la capacité des modèles à simuler des séries temporelles hautement non stationnaires (Lee and Ouarda, 2012). Des méthodes telles que "*Wavelet transform (WT)*", "*Empirical Mode Decomposition (EMD)*" et "*Ensemble EMD (EEMD)*" ont été recommandées comme outil de prétraitement de la série chronologique afin d'améliorer la précision des modèles basés sur l'IA (Kisi and Shiri, 2011; Nourani et al., 2009b; Solgi et al., 2017; Wang et al., 2015).

WT a été développé pour pré-traiter un signal. Cette méthode a été adaptée par un grand nombre d'hydrologues pour décomposer une série chronologique en sous-séries (Oh et al., 2017; Ouachani et al., 2013; Shoaib et al., 2015; Sivapragasam et al., 2015). Shoaib et al. (2015) ont prévu le ruissellement en utilisant le modèle GEP combiné avec le WT. Les auteurs ont conclu que le WT améliore les performances du GEP.. Kisi and Shiri (2011) ont produit des modèles hybrides en combinant les modèles GEP et "*Neuro-Fuzzy (NF)*" pour la prévision des précipitations. Les résultats ont montré que WT améliore la précision des modèles et que GEP combiné à WT, est le modèle le plus précis.

EEMD est une méthode permettant de décomposer une série chronologique en composantes et convient au prétraitement de séries chronologiques non stationnaires (Wang et al., 2015). De nombreuses études prouvent que l'utilisation de EEMD augmente la précision de la modélisation (Lee and Ouarda, 2011; Lee and Ouarda, 2012; Masselot et al., 2018). Huang et al. (2014) ont produit un modèle hybride conçu par EEMD pour simuler le débit mensuel et ont indiqué que EEMD pouvait améliorer les résultats de la simulation. Dans une autre étude, Jiao et al. (2016) ont confirmé qu'un modèle hybride produit par EEMD avait la capacité de simuler des données hydrologiques.

1.1. PROBLÈME

Comme les ES constituent une variable non linéaire et que les chercheurs n'ont pas accès à toutes les variables affectives sur les ES dans les conditions réelles, les modèles physiques et mathématiques peuvent ne pas aboutir à une simulation avec une bonne précision.

1.2. HYPOTHÈSE

Les modèles basés sur l'IA peuvent être utiles pour la simulation des ES et leurs combinaisons avec WT et EEMD pourrait améliorer leur précision.

1.3. OBJECTIFS

L'objectif principal de cette étude est de présenter un modèle permettant de simuler les ES avec une grande précision. Le modèle devrait dépasser les limites imposées à la simulation de séries chronologiques hautement non stationnaires.

1.4. ORIGINALITÉ

À la connaissance de l'auteur, il n'y a pas eu d'étude combinant GEP et M5 avec EEMD pour la simulation ES.

1.5. ORGANISATION DE LA SYNTHÈSE

La présente synthèse est divisée en cinq parties principales. La section 1 présente l'introduction. La section 2 explique les méthodes. La section 3 donne les informations sur l'étude de cas et les données utilisées. La section 4 indique les résultats. La section 5 présente les conclusions et recommandations.

2. MÉTHODES

Dans la présente étude, deux modèles appelés GEP et M5 et deux méthodes appelées WT et EEMD sont appliqués à la simulation des ESs.. Pour ce faire, on prend en compte les séries temporelles de précipitations mensuelles et de température de l'air d'une station météorologique, ainsi que les séries temporelles des niveaux d'ES de trois puits d'observation situés dans la plaine de Delfan, située au nord de la province de Lorestan, en Iran. Ensuite, chaque série temporelle est décomposée en sous-signaux par WT et EEMD. Les sous-signaux générés par WT sont utilisés pour produire des modèles hybrides appelés WGEP et WM5. Les sous-signaux produits par EEMD sont appliqués pour former des modèles hybrides appelés EEMD-GEP et EEMD-M5. Enfin, les valeurs de simulation utilisant des modèles hybrides et simples sont évaluées selon différents critères pour déterminer le modèle à grande précision.

Les algorithmes et les formules de modèles et de méthodes sont présentés dans les articles. Dans les parties suivantes, un résumé des modèles et méthodes utilisés est présenté.

2.1. GENE EXPRESSION PROGRAMING (GEP)

GEP a été introduit par Ferreira (2001) et fait partie des algorithmes évolutifs. Les algorithmes évolutifs imitent le mécanisme des organismes vivants. L'avantage de GEP par rapport à d'autres modèles tels que ANN est que la structure optimale du modèle est déterminée pendant le processus de formation.

Ferreira (2006) explique que l'algorithme GEP génère une population de solutions au hasard. Si l'algorithme atteint une solution attendue, l'algorithme est arrêté et la meilleure

solution est présentée. Dans le cas de non satisfaction de la solution attendue, l'évolution de la population continue à améliorer les solutions.

2.2. M5 MODEL TREE (M5)

Quinlan (1992) a présenté le M5 en développant un "*Decision Tree*". M5 établit une relation entre les variables dépendantes et indépendantes en présentant des règles incluant des régressions linéaires. L'avantage de M5 par rapport aux arbres de classification et de régression (CART) est que M5 produit un arbre plus petit et est plus précis (Quinlan, 1992).

M5 utilise une approche descendante pour aller d'un nœud en haut à une feuille en bas et son algorithme comporte deux étapes. Quinlan (1992) explique qu'à la première étape, un critère de scission est appliqué pour construire un arbre. Le critère de division explique le taux d'erreur au nœud et repose sur la minimisation de l'écart type des nœuds. À la deuxième étape, l'algorithme de Quinlan (1992) est utilisé pour élaguer les branches et les remplacer par des modèles de régression linéaire.

2.3. WAVELET TRANSFORM (WT)

WT a été présenté par Grossmann and Morlet (1984) dans la discipline des géosciences. Le WT est une méthode de traitement permettant d'extraire des informations cachées d'une série chronologique. Pour des applications pratiques, les hydrologues ont accès à des signaux de temps discrets (Nourani et al., 2014). Ainsi, Mallat (1989) a introduit une extension discrète de WT appelée "*discrete wavelet transform (DWT)*".

Mallat (1989) a mis au point la technologie DWT en utilisant des filtres passe haut et passe bas. Dans le DWT, pour le premier niveau de décomposition, un signal est décomposé en une approximation (A) et en détail (d) par les filtres. L'approximation traitée est ensuite appliquée de façon itérative aux décompositions résultantes (Mallat, 1989).

2.4. EMPIRICAL MODE DECOMPOSITION (EMD)

Huang et al. (1998) ont introduit l'EMD en tant que méthode pour décomposer une série chronologique en composantes de fréquence. Les principales différences entre EMD et WT sont qu'EMD extrait les caractéristiques d'une série chronologique sans présupposition et est capable de reconnaître la fréquence instantanée de la série chronologique.

Les composantes de fréquence appelées fonctions de mode intrinsèque (IMF) ont deux caractéristiques (Lee and Ouarda, 2012). Premièrement, le nombre d'extrema locaux doit être égal au nombre de passages par zéro ou en diffèrer au plus par un. Deuxièmement, la moyenne des enveloppes supérieure et inférieure, calculée par les maxima et les minima locaux, devrait être égale à zéro.

Le processus de calcul des IMF est le suivant : d'abord, déterminer les minima et les maxima locaux. Deuxièmement, adapter les enveloppes supérieure et inférieure aux maxima et minima locaux, respectivement, par splines cubiques. Troisièmement, soustraire la moyenne des enveloppes supérieure et inférieure de la série temporelle à tout emplacement temporel pour calculer la première composante. Quatrièmement, calculer Dk comme critère d'arrêt. Avec $D_k=\frac{\sum_{t=0}^{T}|h_1^{k-1}(t)-h_1^k(t)|^2}{\sum_{t=0}^{T}|h_1^{k-1}(t)|^2}$ où k est le numéro de répétition additionnelle pour le processus de tamisage. Si Dk est supérieur à 0,2, la série temporelle est remplacée par le signal calculé à partir de la troisième étape. Ensuite, le processus est répété à partir de la première étape. Si Dk est inférieur à 0,2, le signal calculé est appelé le premier FMI. Cinquièmement, définir le résidu en soustrayant le signal calculé de la série temporelle. Si le résidu satisfait aux caractéristiques des IMF, il est considéré comme une série chronologique et les étapes 1 à 4 sont extraites pour produire le prochain FMI. En cas de non-satisfaction de la condition, l'algorithme est arrêté.

L'algorithme EMD peut causer des problèmes de mélange de modes (Lee and Ouarda, 2010; Wang et al., 2015). Pour résoudre le problème, Wu and Huang (2009) ont développé EMD en ajoutant des bruits blancs finis à la série chronologique ciblée et l'ont

appelé ensemble EMD (EEMD). Dans le système EEMD, le nombre d'ensembles ajoutés (ε) et l'amplitude du bruit devraient être prescrits.

2.5. CRITÈRE D'ÉVALUATION

Dans la présente étude, le coefficient de détermination (R^2), la racine carrée de l'erreur quadratique moyenne (RMSE), la racine carrée de l'erreur quadratique moyenne relative (rRMSE), le biais moyen (BIAS) et le biais moyen relatif (rBIAS) sont utilisés pour évaluer les performances des modèles.

3. ZONE D'ÉTUDE

Les méthodes sont appliquées aux données de la plaine de Delfan, située au nord de la province du Lorestan, en Iran. La plaine est située entre 47°, 21' et 48°, 21' E à 33°, 48' et 34°, 22' N, entre 1700 et 2100 mètres d'altitude. Les valeurs mensuelles de ES (m), de précipitation (mm) et température de l'air (°C) sont utilisées pour simuler ES un mois à l'avance. Dans la présente étude, des décalages sont appliqués aux variables d'entrée. Le décalage dans le temps signifie l'utilisation des valeurs des variables correspondant aux mois précédents. Le décalage temporel 1 fait référence aux variables correspondant au mois précédent, le décalage temporel 2 aux variables correspondant à il y a deux mois.

La série d'observations est divisée en deux parties. La première partie, comprenant 75% de la série, est utilisée pour l'étalonnage et la seconde partie, comprenant 25%, est utilisée pour la validation.

4. RÉSULTATS

Les résultats de cette étude sont résumés dans le paragraphe suivant. En premier lieu nous présentons les résultats des modèles simples, GEP et M5. Ensuite nous résumons ceux des modèles hybrides.

La performance des modèles simples indique une meilleure performance de GEP par rapport au M5 pour les puits 1 et 3. La performance des deux modèles est faible pour le

puits 2. Les résultats indiquent que, pour différentes combinaisons de variables, les précipitations ne sont pas une variable efficace pour la simulation, tandis que la température de l'air et les ES jouent un rôle important dans la simulation.

Pour tous les puits d'observation, les modèles hybrides produits par WT donnent de meilleures performances que GEP et M5. Les performances de WGEP et WM5 sont proches. Pour le puits 1, les résultats de calibration sont proches, mais pour la validation, WM5 performe légèrement mieux que WGEP. Pour le puits 2, les résultats de validation de WGEP montrent une meilleure performance que WM5. Pour le puits 3, WGEP est un peu meilleur que WM5 pour l'étalonnage et la validation.

Pour le puits 1, les performances du modèle EEMD-GEP produisent la meilleure précision en utilisant $\varepsilon = 0{,}3$. Pour les puits 2 et 3, $\varepsilon = 0{,}2$ indique la plus grande précision. Les critères d'évaluation de EEMD-M5 indiquent que ses performances sont trop faibles et que le modèle n'est pas adapté à la simulation.

On peut comprendre de la comparaison des modèles que les modèles hybrides générés par WT ont de meilleures performances que les modèles hybrides générés par la méthode EEMD. Pour le puits 1, WM5 peut être sélectionné comme le meilleur modèle pour la simulation. En ce qui concerne le puits 2, le WGEP indique les résultats les plus appropriés. Pour le puits 3, WGEP et WM5 ont des performances proches et sont meilleures que EEMD-GEP et EEMD-M5.

5. CONCLUSIONS ET RECOMMENDATIONS

Dans la présente étude, les modèles GEP et M5 sont utilisés pour la simulation. Des modèles WT, ainsi que EEMD, sont appliqués pour produire des modèles hybrides. Cette étude inclut également une discussion sur l'article de Du et al. (2017) sur l'utilisation correcte de l'application de WT. La conclusion de la discussion est également résumée à la fin de cette section. Cette partie présente le résumé des articles.

En se basant sur les résultats des modèles, on peut conclure que, pour prétraiter et simuler les séries temporelles de ES, le WT et le GEP, respectivement, produisent des résultats plus précis

Les résultats de modèles simples démontrent que la sélection d'un temps de latence approprié pour les valeurs d'entrée joue un rôle important dans la simulation. Si un grand nombre de temps de latence est utilisé, le temps de simulation augmente, alors que les temps de latence ne sont pas toujours utiles pour améliorer la précision.

Les résultats des modèles hybrides à différents niveaux de la décomposition indiquent que la sélection d'un niveau de décomposition approprié affecte la précision des modèles hybrides.. Un niveau élevé de décomposition n'est pas toujours utile pour augmenter la précision du modèle. Un niveau de décomposition optimal doit être sélectionné.

Les résultats des modèles hybrides intégrés à EEMD présentent l'importance de ε dans la méthode EEMD et son impact sur la précision des modèles hybrides.

En outre, on peut conclure que les valeurs de température de l'air et des ES sont des variables plus efficaces que les valeurs de précipitation pour la simulation des ES dans cette région de l'Iran. Cela signifie que la sélection de variables hydrologiques en tant qu'entrées dans les modèles est très spécifique à la région d'étude et au climat en question, et affecte de manière significative les performances des modèles.

Dans le présent travail, la simulation des ES a été envisagée à l'aide de modèles simples et hybrides. Il serait intéressant d'évaluer la capacité de ces modèles à prédire à long terme les ES. Ceci est important car la gestion efficace d'un certain nombre de problèmes environnementaux (tels que les sécheresses) nécessite un plan à long terme.

Pour améliorer les méthodes d'EEMD, de nouvelles techniques telles que "*Complementary Ensemble Mode Decomposition*" et "*Partly Ensemble Mode Decomposition*" ont été développées. Il est suggéré d'étudier la capacité des nouvelles techniques à produire des modèles hybrides pour la simulation des ES.

Dans la discussion, les mêmes données de Du et al (2017) sont utilisées pour vérifier la validité des affirmations des auteurs. La seule différence entre l'analyse effectuée dans la présente discussion et dans l'article initial est l'utilisation des décalages temporels appliqués aux entrées des modèles. Les résultats montrent que si les décalages temporels 1 à 7 sont utilisés comme entrées dans les modèles, la transformation de la

série chronologique par WT ne donne pas de faux résultats. Par conséquent, l'affirmation de Du et al. (2017) n'a pas pu être vérifiée pour DWT-ANN.

6. RÉFÉRENCES

[1] Barzegar R, Fijani E, Asghari Moghaddam A, Tziritis E (2017) Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models. Science of The Total Environment. 599-600:20-31. doi:https://doi.org/10.1016/j.scitotenv.2017.04.189

[2] Bhattacharya B, Solomatine DP (2005) Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing. 63:381-396. doi:https://doi.org/10.1016/j.neucom.2004.04.016

[3] Coppola E, Szidarovszky F, Poulton M, Charles E (2003) Artificial Neural Network Approach for Predicting Transient Water Levels in a Multilayered Groundwater System under Variable State, Pumping, and Climate Conditions. Journal of Hydrologic Engineering. 8:348-360. doi:10.1061/(ASCE)1084-0699(2003)8:6(348)

[4] Ferreira C (2001) Gene Expression Programming: A New Adaptive Algorithm for Solving Problems vol 13. Complex Systems Publications, Angra do Heroísmo, Portugal.

[5] Ferreira C (2006) Gene Expression Programming Mathematical Modeling by an Artificial Intelligence. Studies in Computational Intelligence, vol 21. Springer-Verlag Berlin Heidelberg, Germany. doi:10.1007/3-540-32849-1

[6] Grossmann A, Morlet J (1984) Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape SIAM. Journal on Mathematical Analysis 15:723-736. doi:10.1137/0515056

[7] Huang NE et al. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences. 454:903-995. doi:10.1098/rspa.1998.0193

[8] Huang S, Chang J, Huang Q, Chen Y (2014) Monthly streamflow prediction using modified EMD-based support vector machine. Journal of Hydrology. 511:764-775 doi:https://doi.org/10.1016/j.jhydrol.2014.01.062

[9] Jiao G, Guo T, Ding Y (2016) A New Hybrid Forecasting Approach Applied to Hydrological Data: A Case Study on Precipitation in Northwestern China Water. 8:367

[10] Kisi O (2015) Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology. 528:312-320. doi:https://doi.org/10.1016/j.jhydrol.2015.06.052

[11] Kisi O (2016) Modeling reference evapotranspiration using three different heuristic regression approaches Agricultural. Water Management. 169:162-172 doi:https://doi.org/10.1016/j.agwat.2016.02.026

[12] Kisi O, Shiri J (2011) Precipitation Forecasting Using Wavelet-Genetic Programming and Wavelet-Neuro-Fuzzy Conjunction Models. Water Resources Management. 25:3135-3152. doi:10.1007/s11269-011-9849-3

[13] Lallahem S, Mania J, Hani A, Najjar Y (2005) On the use of neural networks to evaluate groundwater levels in fractured media. Journal of Hydrology. 307:92-111. doi:http://dx.doi.org/10.1016/j.jhydrol.2004.10.005

[14] Lee T, Ouarda TBMJ (2010) Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. Journal of Geophysical Research: Atmospheres. 115: D13107. doi:10.1029/2009JD012801

[15] Lee T, Ouarda TBMJ (2011) Prediction of climate nonstationary oscillation processes with empirical mode decomposition. Journal of Geophysical Research: Atmospheres. 116: D06107. doi:10.1029/2010JD015142

[16] Lee T, Ouarda TBMJ (2012) Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition. Water Resources Research. 48: W02514. doi:10.1029/2011WR010660

[17] Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 11:674-693. doi:10.1109/34.192463

[18] Masselot P, Chebana F, Bélanger D, St-Hilaire A, Abdous B, Gosselin P, Ouarda TBMJ (2018) EMD-regression for modelling multi-scale relationships, and application to weather-related cardiovascular mortality. Science of The Total Environment. 612:1018-1029. doi:https://doi.org/10.1016/j.scitotenv.2017.08.276

[19] Nayak PC, Rao YRS, Sudheer KP (2006) Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. Water Resources Management. 20:77-90. doi:10.1007/s11269-006-4007-z

[20] Nourani V, Alami MT, Aminfar MH (2009a) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Engineering Applications of Artificial Intelligence. 22:466-472. doi:https://doi.org/10.1016/j.engappai.2008.09.003

[21] Nourani V, Hosseini Baghanam A, Adamowski J, Kisi O (2014) Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. Journal of Hydrology. 514:358-377. doi:https://doi.org/10.1016/j.jhydrol.2014.03.057

[22] Nourani V, Komasi M, Mano A (2009b) A Multivariate ANN-Wavelet Approach for Rainfall–Runoff Modeling. Water Resources Management. 23:2877. doi:10.1007/s11269-009-9414-5

[23] Nourani V, Mousavi S (2016) Spatiotemporal groundwater level modeling using hybrid artificial intelligence-meshless method. Journal of Hydrology. 536:10-25. doi:https://doi.org/10.1016/j.jhydrol.2016.02.030

[24] Oh Y-Y, Yun S-T, Yu S, Hamm S-Y (2017) The combined use of dynamic factor analysis and wavelet analysis to evaluate latent factors controlling complex groundwater level fluctuations in a riverside alluvial aquifer. Journal of Hydrology. 555:938-955. doi:https://doi.org/10.1016/j.jhydrol.2017.10.070

[25] Ouachani R, Zoubeida B, Taha O (2013) Power of teleconnection patterns on precipitation and streamflow variability of upper Medjerda Basin International. Journal of Climatology. 33:58-76. doi:doi:10.1002/joc.3407

[26] Quinlan JR (1992) Learning with Continuous Classes Proceedings of Australian. Joint Conference on Artificial Intelligence, Hobart, Australia, World Scientific, Singapore:343-348

[27] Sahoo GB, Ray C, Wade HF (2005) Pesticide prediction in ground water in North Carolina domestic wells using artificial neural networks. Ecological Modelling 183:29-46. doi:http://dx.doi.org/10.1016/j.ecolmodel.2004.07.021

[28] Shiri J, Kişi Ö (2011) Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Computers & Geosciences. 37:1692-1701 doi:http://dx.doi.org/10.1016/j.cageo.2010.11.010

[29] Shiri J, Kisi O, Yoon H, Lee K-K, Hossein Nazemi A (2013) Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques. Computers & Geosciences. 56:32-44. doi:https://doi.org/10.1016/j.cageo.2013.01.007

[30] Shoaib M, Shamseldin AY, Melville BW, Khan MM (2015) Runoff forecasting using hybrid Wavelet Gene Expression Programming (WGEP) approach. Journal of Hydrology. 527:326-344. doi:http://dx.doi.org/10.1016/j.jhydrol.2015.04.072

[31] Sivapragasam C, Kannabiran K, Karthik G, Raja S (2015) Assessing Suitability of GP Modeling for Groundwater Level. Aquatic Procedia. 4:693-699 doi:https://doi.org/10.1016/j.aqpro.2015.02.089

[32] Solgi A, Pourhaghi A, Bahmani R, Zarei H (2017) Pre-processing data using wavelet transform and PCA based on support vector regression and gene expression programming for river flow simulation. Journal of Earth System Science. 126:65. doi:10.1007/s12040-017-0850-y

[33] Solomatine DP, Xue Y (2004) M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. Journal of

Hydrologic Engineering. 9:491-501. doi:doi:10.1061/(ASCE)1084-0699(2004)9:6(491)

[34] Wang W-c, Chau K-w, Xu D-m, Chen X-Y (2015) Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition. Water Resources Management. 29:2655-2675. doi:10.1007/s11269-015-0962-6

[35] Wu Z, Huang NE (2009) ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD. Advances in Adaptive Data Analysis. 01:1-41. doi:10.1142/s1793536909000047

PARTIE II: ARTICLES

CHAPITRE II: GROUNDWATER LEVEL SIMULATION USING GENE EXPRESSION PROGRAMMING AND M5 MODEL TREE COMBINED WITH WAVELET TRANSFORM

Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform

Ramin Bahmani[1], Abazar Solgi[2], Taha B. M. J. Ouarda[1]

[1] Canada Research Chair in Statistical Hydro-climatology, INRS-ETE, Québec (Québec), Canada.

[2] Department of Water Resources Engineering, Faculty of Water Sciences Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

Journal : submitted to the Journal of Hydrology

Abstract

In order to understand and adequately manage hydrological stress, it is necessary to simulate the groundwater level accurately. In the present research, Gene Expression Programming (GEP) and M5 model tree (M5) are used to simulate monthly groundwater levels. The models are combined with wavelet transform to produce two hybrid models: Wavelet Gene Expression Programming (WGEP) and Wavelet M5 model tree (WM5). For the simulation, groundwater level, temperature and precipitation values from three observation wells and one meteorological station, located in the Delfan plain (Nourabad), Iran, are used. The results indicate that the hybrid models, WGEP and WM5, have better performance than simple models, GEP and M5. The performances of the two hybrid models are similar. It is also observed that selecting a suitable time lag for inputs plays an important role in the accuracy of the simple models. The selection of a suitable decomposition level strongly affects the accuracy of hybrid models.

Keywords: Groundwater level; Gene Expression Programming; Hybrid model; M5 Model tree; Wavelet transform.

Abbreviations

GEP	Gene Expression Programming
M5	M5 model tree
WGEP	Wavelet Gene Expression Programming
WM5	Wavelet M5 model tree
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
LSSVM	Least Square Support Vector Machine
MARS	Multivariate Adaptive Regression Splines
WT	Wavelet Transform
NF	Neuro-Fuzzy
RMSE	Root Mean Squared Error
rRMSE	Relative Root Mean Squared Error
BIAS	Mean Bias
rBIAS	Relative Mean Bias

1. Introduction

Measuring groundwater levels, using observation wells, is one of the main resources for investigating hydrological stresses and is necessary to understand groundwater fluctuations in arid and semi-arid areas (Barzegar et al., 2017; Ebrahimi and Rajaee, 2017). The investigation of groundwater levels provides key information to managers for decision making during extreme conditions such as droughts (Barzegar et al., 2017; Reghunath et al., 2005).

To simulate groundwater levels, physical models have been considered by a number of researchers. Although these models have been widely used, they have not always provided all required information concerning groundwater level simulation (Bierkens, 1998; Nayak et al., 2006). Compared to statistical approaches, physical models require much more high accuracy data for simulating groundwater levels and do not always lead to good performances in establishing non-linear relationships between variables (Nayak et al., 2006; Shiri et al., 2013).

To solve the problem, Artificial Intelligence (AI) models have been considered for simulating groundwater levels and several studies have reported the good ability of AI-based models to simulate groundwater levels (Coppola et al., 2003; Lallahem et al., 2005; Nayak et al., 2006; Sahoo et al., 2005). Among a relatively wide range of AI models, Gene Expression Programming (GEP) has particularly shown a high ability to efficiently simulate groundwater levels. For instance, Shiri and Kişi (2011) investigated the ability of GEP and the Adaptive Neuro-Fuzzy Inference System (ANFIS) for forecasting water table fluctuations and found that the models have a good ability to forecast the fluctuations, with the GEP leading to better results than the ANFIS. Again, Shiri et al. (2013) simulated groundwater level fluctuations by considering the meteorological effects using GEP, Supportive Vector Machine, and ANFIS models. The authors used different combinations of precipitation, groundwater level, and evaporation values for simulation and concluded that GEP leads to better performances than the other methods.

AI models have shown a high performance for simulating natural variables (Chokmani et al., 2008; Eissa et al., 2013; Solgi et al., 2017c). However, they have been criticized for

being black box models with which it is hard to understand the dynamics and the effects of different variables on the simulation output (Liao and Sun, 2010; Solomatine and Xue, 2004). Hence, transparent models such as M5 model tree (M5), which produces explicit equations, have been proposed for simulation purposes (Keshtegar et al., 2016; Raza, 2015).

The M5 model has been successfully used for simulating hydrological variables. For instance, Liao and Sun (2010) used the improved decision tree model to predict water quality. The authors found the decision tree method to be more accurate and easier to use than Artificial Neural Networks (ANNs). They concluded that improved decision tree models, which integrate some of the ANN characteristics in the main decision tree methods, were suitable for their research field. Solomatine and Xue (2004) used M5 and ANN to compare the accuracy of models for flood forecasting. The study reported that M5 has the same accuracy than ANNs but with a number of advantages, such as transparency, high speed of training, and fast convergence. In another study, Bhattacharya and Solomatine (2005) used ANN and an M5 to develop models of the water level-discharge relationship. They concluded that both ANN and M5 models were better than the traditional rating curve, using the same data. In more recent research, Kisi (2015) used Least Square Support Vector Machine (LSSVM), Multivariate Adaptive Regression Splines (MARS) and M5 to analyze the accuracy of the models for simulating pan evaporation. The author defined different scenarios to analyze the accuracy of the models and concluded that, when the local input and output data are used, LSSVM has the best results, whereas MARS leads to a better performance when the local input and output data are not considered. Kisi (2016) used the same models to simulate evapotranspiration and concluded that LSSVR, MARS, and M5 are suitable for simulation with the local input and output, without local input, and without the local input and output values respectively. Kisi and Parmar (2016) used the same models to simulate river water pollution and concluded that MARS and LSSVM are suitable to model river water pollution.

The literature indicates that GEP and M5 have a high ability to adequately simulate hydrologic time series. However, when hydrologic phenomena are highly non-stationary AI-based models may not be able to properly simulate the hydrologic time series (Nourani

et al., 2014). In this case, wavelet transform (WT) can be recommended as a tool for pre-processing the time series in order to increase the accuracy of AI-based models (Nourani et al., 2009a; Solgi et al., 2017c). WT helps hydrologists decompose a time signal into sub-signals allowing the models to capture information with different resolution levels (Nourani et al., 2009b).

WT was presented by Grossmann and Morlet (1984) in the field of geoscience and has been adapted for the pre-processing of time series in hydrology (Oh et al., 2017; Ouachani et al., 2013; Shoaib et al., 2015; Sivapragasam et al., 2015; Solgi et al., 2017a). Partal and Kişi (2007) used WT to decompose daily precipitation time series to sub-signals. Then, the sub-signals were entered as inputs to a Neuro-Fuzzy (NF) model to forecast precipitation. The authors reported that, when WT was used, the NF model led to a better fit with observed values. Shoaib et al. (2015) combined GEP with WT to predict runoff and concluded that using WT improves the performance of GEP. Kisi and Shiri (2011) used GEP and NF models to forecast precipitation, and also combined the models with WT to produce hybrid models. The authors concluded that hybrid models were more accurate than simple models and GEP combined with WT led to best results.

Very few studies have used GEP and M5 in combination with WT to simulate groundwater levels. The first objective of the present study is to simulate monthly groundwater levels using GEP and M5. The models are also combined with WT (the resulting models are called WGEP and WM5) to produce hybrid models for groundwater level simulation. For the simulation using simple models, different combinations of groundwater level, precipitation and temperature time series are used. For the simulation using hybrid models, the time series is decomposed to sub-signals using WT. Then, the sub-signals are used as inputs to the GEP and M5 models. Finally, the performances of simple and hybrid models are compared to evaluate their performance.

The present paper is structured as follows. Section 2 provides a review of the theoretical background of the models used in this study, as well as the methodology adopted to produce the hybrid models and to compare the performances of the various models. Section 3 presents the case study. The results are presented and discussed in section 4. Finally, the conclusions and recommendations for future work are presented in section 5.

2. Methods

2.1. Gene Expression Programing (GEP)

GEP was first introduced by Ferreira (2001) and is part of what is called “evolutionary algorithms”. Evolutionary algorithms, inspired by the Darwinian evolution theory, are algorithms that imitate the mechanism of living organisms. GEP differs from other models such as ANN, in that the structure, including the input variables, target and position function, are firstly defined. Then, the optimal structure of the model and coefficients are determined during the training process, whereas in the ANN model, the structure must be determined and only the coefficients of the model are obtained during the training process. This helps provide more flexibility to the GEP model and generally represents an advantage. The general process of solving a problem with GEP is briefly illustrated by the following steps.

1. Select a terminal set, which includes dependent and independent values. In this step, the root mean square error, as the fitness function, is used.

2. Choose a set of functions which includes arithmetic operators, test functions, and Boolean functions. The 10 used mathematical operators are multiplication, division, addition, subtraction, exponential, square root, natural logarithm, x to the power of 2, cube root, and inverse. The arithmetic operators of addition, subtraction, and multiplication are the most common types.

3. Obtain the accuracy index of the model which indicates the model ability to solve a specific problem.

4. Select stopping criteria, which are measures for stopping the program and obtaining the results, such as generated numbers or the maximum fixation.

5. Determine the control components, including numerical values and qualitative variables. These are used for controlling the program.

The present paper does not delve into the detailed theory of GEP. For more information, the reader is referred to Ferreira (2001), Ferreira (2002), and Ferreira (2006).

2.2. M5 Model Tree (M5)

The M5 Model Tree (M5) was introduced by Quinlan (1992) based on the Decision Tree method. The main advantage of M5 is that it presents a set of linear regressions that show the relationships between dependent and independent variables and they are interpretable (Raza, 2015). The M5 model lies between linear models, like ARIMA, and non-linear models, like ANNs (Solomatine and Xue, 2004). M5 includes a root, branches, nodes, and leaves like a tree and is drawn using top-down approach. The root as the first node is on the top and the tree grows down by nodes and branches to reach leaves. In M5, every leaf is a linear regression. Building the M5 model involves two steps. The first step consists in building a tree using a splitting criterion. The second step consists in pruning branches and replacing them with linear regression models.

2.2.1 Splitting criterion

The splitting criterion explains the rate of errors at the node and is based on minimizing the standard deviation of nodes. Because of the splitting process, the standard deviation in a given node, called child node, is generally smaller than the standard deviation in the previous node, called parent node (Kisi, 2015). A node ends in a leaf when minimizing standard deviation in the node is not possible. For the calculation of the reduction in standard deviation, the following equation is used (Quinlan, 1992):

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \quad (1)$$

where *SDR* is the reduction of standard deviation in a given node, *T* is a set of training data that reaches the parent node; T_i is a subset of the training data having the i^{th} outcome of the potential set, and sd is the standard deviation (Kisi, 2015). The M5 model selects the minimum expected error for each attribute at that node. For more details concerning the splitting criterion, the reader is referred to Quinlan (1986) and Quinlan (1992).

2.2.2 Pruning

The splitting process produces a huge tree that may cause overfits to training data. Therefore, it is necessary to prune the branches. The recommended algorithm for pruning is the Quinlan algorithm. The algorithm first allows the tree to grow as much as possible.

Secondly, it prunes the branches which do not increase the accuracy of the model. Finally, after the pruning, smoothing is needed to smooth sharp dissociation (Bhattacharya and Solomatine, 2005; Solomatine and Xue, 2004). For more details concerning the general M5 model, the reader is referred to (Quinlan, 1992).

2.3 Wavelet transform (WT)

The application of the wavelet transform (WT) in the geoscience discipline was introduced by Grossmann and Morlet (1984). The WT is an effective digital signal processing technique to extract hidden information form a signal. The WT was formulated as the continuous wavelet transform (CWT) by Mallat (1989) as follow:

$$CWT = \int_{-\infty}^{+\infty} f(t).\psi^*(t).d(t) \tag{2}$$

where *f(t)* is the continuous signal, $\psi(t)$ is the wavelet function or mother wavelet and * refers to the complex conjugate. A mother wavelet has three characteristics: 1) limited number of fluctuations, 2) quick return to zero in both positive and negative directions in its domain and 3) zero average (Thuillard, 2001).

A mother wavelet is defined as (Oh et al., 2017):

$$\psi_{(a,b)}(t) = |a|^{-0.5}\psi\left(\frac{t-b}{a}\right), \quad a \in R, b \in R, a \neq 0 \tag{3}$$

where *a* is a dilation factor, *b* is a temporal translation, and *R* is the domain of real numbers.

For practical applications, hydrologists and modelers in general often have access to discrete time signals (Nourani et al., 2014). Therefore, a discrete extension of equation (2) called “discrete wavelet transform” (DWT) was proposed. The DWT is defined by (Kisi and Shiri, 2011):

$$DWT\ (a,b) = 2^{-\frac{j}{2}} \int_{j=1}^{j=J} \psi^*(\,2^{-\frac{j}{2}} - k).f(t).dt \tag{4}$$

where *f(t)* is a discrete time series at any time t, . *j* and *k* are integers which respectively control the wavelet dilation and the translation.

By using high-pass and low-pass filters, Mallat (1989) developed a way to implement the DWT. In the DWT, for the first decomposition level, a signal is decomposed to an approximation (A) and detail (d). The processed approximation is then applied to consequent decompositions iteratively (Mallat, 1989).

2.3.1 Choice of a mother wavelet

The main consideration for selecting a mother wavelet is the type of the time series. The main features of a mother wavelet include the region of support, associated with the wavelet span length, and the number of vanishing moments, limiting the ability of a wavelet to show information in a time series. In the present work, the Haar, Coif1, Sym3, Db4, and Db2 wavelets, which are well-known and commonly used in hydrological studies, are adopted for decomposing time signals(Nourani et al., 2014; Shoaib et al., 2015; Solgi et al., 2017b). Fig. 1 illustrates the graph of the mother wavelets.

2.4 Hybrid Wavelet M5 model tree (WM5) and Wavelet Gene Expression Programming Models (WGEP) development

To develop hybrid models with the wavelet transform, the following steps are carried out. The groundwater level, temperature and precipitation time series are decomposed to sub-signals using DWT for different decomposition levels. Then, the sub-signals are used as inputs into the GEP and M5 models, hence called WGEP and WM5. Fig. 2 presents the schematic diagram for producing WGEP and WM5 models. In this figure, P(t), T(t), and H(t) refer to the precipitation, temperature, and groundwater level time series respectively. The sub-signals of P_a, T_a, and H_a refer to the approximation of the final decomposition level. The sub-signals of Pd1 to Pd_n, Td1 to Td_n and Hd1 to Hd_n refer to the details of the decomposition from level 1 to the last decomposition level.

To determine the level of decomposition (L), the suggested equation of Nourani et al. (2009b) is used.

$$L = \text{Int}[\log(N)] \quad (5)$$

where L is the proposed decomposition level and N is the number of observed values.

2.5 Evaluation Criteria

Evaluation criteria serve to estimate the errors of the models according to the observed and simulated values. In the present paper, the following criteria are used to evaluate the models (Chebana et al., 2014):

a- The coefficient of determination or R^2:

$$R^2 = 1 - \frac{\sum(H_{obs} - H_{est})^2}{\sum(H_{obs} - \bar{H})^2} \quad (6)$$

b- The root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum(H_{obs} - H_{est})^2}{n}} \quad (7)$$

c- The relative root mean square error (rRMSE):

$$rRMSE = 100\sqrt{\frac{\sum(\frac{H_{obs} - H_{est}}{H_{obs}})^2}{n}} \quad (8)$$

d- The mean bias (BIAS):

$$BIAS = \frac{\sum(H_{obs} - H_{est})}{n} \quad (9)$$

e- The relative mean bias (rBIAS)

$$rBIAS = 100 * (\frac{\sum\left(\frac{H_{obs} - H_{est}}{H_{obs}}\right)}{n}) \quad (10)$$

where H_{obs} refers to the observed values, H_{est} refers to the estimated values, $\bar{H}$ is the average of observed values and n is the number of observations.

3. Case study and data used

The methods are applied to the data of the Delfan plain, located in the north of the Lorestan Province, Iran. The plain is situated between 47°, 21' and 48°, 21' E to 33°, 48' and 34°, 22' N (see Fig. 3), 1700 to 2100 meters above sea level, in the Zagros mountain chain region. The total area of the plain is about 300 Km^2. The high lands of the region are formed by thick layers of limestone and the lowland areas are formed by a combination of igneous, metamorphic, and calcareous structures mixed with metamorphic sheets. The morphological units of the plain consist of different alluvial structures, such as fans and debris. The average annual rainfall in the plains is 480 mm/year. Table 1 presents the features of three observation wells and one meteorological station located in the plain.

Monthly groundwater level (m), precipitation (mm), and temperature ($^{\circ}C$) values are used to simulate groundwater level one month ahead. In general, groundwater level responds to the input variables with a delay. Therefore, in the present paper, the input values are considered with time lags. A time lag refers to the value of a variable corresponding to previous months. Table 2 presents different combinations of variables with time lags. In Table 2, H_t, P_t, and T_t refer to the groundwater level, precipitation, and temperature for the present time, respectively. H_{t-1}, P_{t-1}, and T_{t-1} represent the variable values with one time lag, which means the values of the variables corresponding to the previous month.

The observation series are divided into two parts. The first part, including 75% of the series, is used for calibration and the second part, including 25%, is used for validation.

In Eq. 5, to produce hybrid models L can take the value of 2. However, to investigate the effect of the decomposition level on simulation, values of L= 1 to 4 are considered in the present paper.

4. Results and Discussion

4.1 Results of GEP model

Different combinations are used for the calibration and validation of the GEP model. Results of the evaluation criteria are presented in Table 3, for the calibration, and Table 4, for the validation.

According to Tables 3 and 4, for well 1, the combination 15 can be selected as the best combination for the GEP model. For combination 15, the R^2 is respectively equal to 0.64 and 0.63 for the calibration and validation. For the combination, RMSE and rRMSE are small and the rBIAS value is zero which indicates that the model is unbiased. For well 1, although the R^2 of combination 2 is equal to 1 for calibration, the combination is not selected as the best one because its R^2 decreases to 0.56 for the validation.

For well 2, GEP shows a weak performance. For calibration, most of the combinations show high R^2 values and small RMSE and BIAS, whereas for validation, R^2 is low and the combinations indicate large RMSE and BIAS values. For well 2, combination 13 is selected as the best combination.

For well 3, combination 13 is also selected as the best combination because it leads to a high R^2 value and zero rBIAS for calibration and has the highest R^2 value for validation.

According to the results, it is concluded that using variables with time lags produces more accurate results. It is important to mention that the results of the GEP model for combinations 4, 5 and 6, for which only precipitation values are used, is too weak for all observation wells.

4.2 Results of M5 model

To calibrate and validate the M5 model, different combinations described in Table 2 are considered as well. The results of the best combinations for calibration and validation are presented in Tables 5 and 6.

According to Tables 5 and 6, combination 10 can be selected for groundwater level simulation at well 1. The evaluation criteria of combination 12 show better results for

calibration, but it is not selected as the best model because its R^2 decreases from 0.80 in calibration to 0.32 for validation, and its RMSE and BIAs increase to 0.62 and -0.18 respectively. For well 1, combinations 10, 11 and 12 are similar to combinations 13, 14 and 15 respectively because the M5 model does not use precipitation values in the simulation. For well 1, M5 leads to a better performance when groundwater level and temperature values are used. On the other hand, the model shows low performances when precipitation values are considered.

For well 2, combination 15 can be selected as the best since it has the largest R^2 value for validation. The performance of M5, similarly to GEP, is not good for the groundwater level simulation of well 2. For calibration, M5 shows high R^2 and low RMSE and BIAS values, while for validation, the R^2 corresponding to all combinations decreases considerably.. For well 2, Similarly to well 1, when only precipitation data are used (combinations 4 to 6) the performance of the model decreases.

For well 3, combination 10 is selected as the best combination because it shows a high performance for both calibration and validation and it leads to similar values of R^2, RMSE, rRMSE, BIAS, and rBIAS for calibration and validation. Similarly to wells 1 and 2, when only precipitation data are used for the simulation, the model leads to a lower performance.

4.3 Results of WGEP model

The results of the evaluation criteria for the WGEP model are shown in Table 7 for calibration, and Table 8 for validation. Results indicate that, for well 1, the mother wavelet of Db4, with L=2, can be selected as the suitable structure for simulation. For well 1, Db2 with L=1 and L=4 also leads to a good performance for validation. However, these combinations are not selected because Db2 with L=4 has an overly complex structure and for Db2 with L=1, the R^2 decreases from 0.8 in calibration, to 0.7 in validation, which may indicate that the structure is not good for generalization.

The best result for well 2 is obtained with the mother wavelet of Db2 with L=3, for which a good performance is obtained with both calibration and validation. The best result for well 3 is obtained with the mother wavelet of Db4 with L=3.

The comparison of Tables 3 and 4 with Tables 7 and 8 shows that using the WT improves the performance of GEP. The improvement is considerable for well 2 in validation. Its R^2 increases form 0.4 for the best combination to 0.76 for the best mother wavelet.

4.4 Results of WM5 model

The results of the evaluation criteria for the WM5 model are presented in Table 9 for calibration, and Table 10 for validation. For well 1, the mother wavelets of Coif1 with L=3 and Haar with L=1 show high performances for validation. However, Coif1 with L=3 is selected as the best mother wavelet because its R^2 value is higher than Haar with L=1 for calibration. Also, its RMSE, rRMSE, BIAS and rBIAS are smaller than Haar with L=1.

For well 2, Db2 with L=4 leads to the best performance. For well 3, all mother wavelets show a good performance. However, the mother wavelet Db2 with L=1 is selected as the best model because the evaluation criteria for both calibration and validation are high and close to each other.

Regarding Tables 9 and 10, it can be concluded that using WT increases the accuracy of M5. Furthermore, it can be concluded that the use of a high level of decomposition does not always lead to an increase in the performance of the models. In a high decomposition level, the sub-signals are too smooth, which means that they do not carry much information. As an example, Fig. 4 presents sub-signals of temperature time series with the mother wavelet of Coif1 with L=8. Fig. 4 shows that when the decomposition level increases, the sub-signals become overly smooth.

4.5 Comparison of different models

Table 11 presents the R^2 and RMSE values for the best results of the M5, GEP, WM5 and WGEP models. In the table, rRMSE, BIAS and rBIAS values are not presented because their values are very similar and close or equal to zero.

The performances of M5 and GEP for different combinations indicate that precipitation is not an effective variable for simulation, unlike temperature and groundwater levels. The low effect of precipitation in simulation may be explained by the fact that the case study is located in a semi-arid region where precipitation is usually zero for several months,

during the end of the spring, the whole summer and the beginning of the fall. During these months, groundwater level decreases are observed despite the fact that precipitation is continuously null. Therefore, the M5 and GEP models cannot effectively establish a relationship between precipitation and groundwater level values.

According to Table 11, for all observation wells, hybrid models lead to better performances than simple models. The performances of WGEP and WM5 are close and it is hard to select one of them as the best model. For well 1, the calibration results are close, but for validation, WM5 is a little better than WGEP. For well 2, the validation results of WGEP show a better performance than WM5. Finally, for well 3, WGEP is a little better than WM5 for both calibration and validation, WM5 shows a slightly better performance than WGEP.

The simulated values for the validation are presented in Fig. 5 and compared to the observed values. In Fig. 5, it can be seen that the hybrid models lead to a better fit with observations in comparison with the simple models.

5. Conclusions and future work

The present paper proposes a few methods to simulate groundwater levels. The GEP and M5 models are used for the simulation and WT is used to produce hybrid models, WGEP and WM5. The results indicate that the performances of WGEP and WM5 are close and it is shown that hybrid models are more accurate than the simple ones. However, M5 and WM5 present two advantages. First, the generated models are simpler and more interpretable. It is possible to understand which independent variables have more effect on simulating the dependent variable. Second, running the M5 and WM5 models is faster than running the GEP and WGEP models respectively.

Based on the results of the present work it can be concluded that first, selecting a suitable lag time for input values plays an important role in groundwater level simulation. If a large number of lag times is used, the time for simulation increases, while the lag times are not always helpful to improve the accuracy. Therefore, an optimal number of lag times for input values must be identified. Second, selecting a suitable decomposition level affects

the accuracy of hybrid models. A high level of decomposition is not always helpful to increase the accuracy of the model and an optimal decomposition level must also be identified.

Furthermore, it was concluded that temperature and groundwater level values are more effective variables than precipitation values for groundwater level simulation in this arid region of Iran. This means that the selection of hydrological variables as inputs to the models is very specific to the region of study and the climate in question, and it affects significantly the performance of the models. Therefore, it is recommended to carry out exhaustive studies in the future in order to evaluate the effects of other hydrological variables on groundwater simulation and understand the dynamics involved, with the objective of finding suitable variables for groundwater modeling in different climates and conditions.

In the present work, the simulation of groundwater levels was carried out using simple and hybrid models. It would be of interest to evaluate the ability of these models, and other models, for prediction of groundwater levels in 3, 6, and 12 months ahead. This is important as the efficient management of a number of environmental issues (such as droughts) requires a long-term plan.

In the present paper, WT was used to improve the accuracy of the models. It is also suggested, for future studies, to evaluate the effect of using other techniques such as Empirical Mode Decomposition Lee and Ouarda (2011) for, as a tool for pre-processing hydrological time series for groundwater level simulation.

6. Acknowledgements

The presents work was partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

7. References

[1] Barzegar R, Fijani E, Asghari Moghaddam A, Tziritis E (2017) Forecasting of groundwater level fluctuations using ensemble hybrid multi-wavelet neural network-based models. Science of The Total Environment. 599-600:20-31. doi:https://doi.org/10.1016/j.scitotenv.2017.04.189

[2] Bhattacharya B, Solomatine DP (2005) Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing. 63:381-396. doi:https://doi.org/10.1016/j.neucom.2004.04.016

[3] Bierkens MFP (1998) Modeling water table fluctuations by means of a stochastic differential equation. Water Resources Research. 34:2485-2499. doi:10.1029/98WR02298

[4] Chebana F, Charron C, Ouarda TBMJ, Martel B (2014) Regional Frequency Analysis at Ungauged Sites with the Generalized Additive Model. Journal of Hydrometeorology. 15:2418-2428. doi:10.1175/jhm-d-14-0060.1

[5] Chokmani K, Ouarda TBMJ, Hamilton S, Ghedira MH, Gingras H (2008) Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques. Journal of Hydrology. 349:383-396. doi:https://doi.org/10.1016/j.jhydrol.2007.11.024

[6] Coppola E, Szidarovszky F, Poulton M, Charles E (2003) Artificial Neural Network Approach for Predicting Transient Water Levels in a Multilayered Groundwater System under Variable State, Pumping, and Climate Conditions. Journal of Hydrologic Engineering. 8:348-360. doi:10.1061/(ASCE)1084-0699(2003)8:6(348)

[7] Ebrahimi H, Rajaee T (2017) Simulation of groundwater level variations using wavelet combined with neural network, linear regression and support vector machine. Global and Planetary Change. 148:181-191. doi:https://doi.org/10.1016/j.gloplacha.2016.11.014

[8] Eissa Y, Marpu PR, Gherboudj I, Ghedira H, Ouarda TBMJ, Chiesa M (2013) Artificial neural network based model for retrieval of the direct normal, diffuse horizontal and global horizontal irradiances using SEVIRI images. Solar Energy. 89:1-16. doi:https://doi.org/10.1016/j.solener.2012.12.008

[9] Ferreira C (2001) Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. vol 13. Complex Systems Publications, Angra do Heroísmo, Portugal.

[10] Ferreira C (2002) Gene Expression Programming in Problem Solving. In: Roy R, Köppen M, Ovaska S, Furuhashi T, Hoffmann F (eds) Soft Computing and Industry: Recent Applications. Springer London, London. pp: 635-653. doi:10.1007/978-1-4471-0123-9_54

[11] Ferreira C (2006) Gene Expression Programming Mathematical Modeling by an Artificial Intelligence. Studies in Computational Intelligence, vol 21. Springer-Verlag Berlin Heidelberg, Germany. doi:10.1007/3-540-32849-1

[12] Grossmann A, Morlet J (1984) Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape SIAM. Journal on Mathematical Analysis. 15:723-736. doi:10.1137/0515056

[13] Keshtegar B, Piri J, Kisi O (2016) A nonlinear mathematical modeling of daily pan evaporation based on conjugate gradient method. Computers and Electronics in Agriculture. 127:120-130. doi:https://doi.org/10.1016/j.compag.2016.05.018

[14] Kisi O (2015) Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology. 528:312-320. doi:https://doi.org/10.1016/j.jhydrol.2015.06.052

[15] Kisi O (2016) Modeling reference evapotranspiration using three different heuristic regression approaches. Agricultural Water Management. 169:162-172. doi:https://doi.org/10.1016/j.agwat.2016.02.026

[16] Kisi O, Parmar KS (2016) Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water

pollution. Journal of Hydrology. 534:104-112. doi:http://dx.doi.org/10.1016/j.jhydrol.2015.12.014

[17] Kisi O, Shiri J (2011) Precipitation Forecasting Using Wavelet-Genetic Programming and Wavelet-Neuro-Fuzzy Conjunction Models. Water Resources Management. 25:3135-3152. doi:10.1007/s11269-011-9849-3

[18] Lallahem S, Mania J, Hani A, Najjar Y (2005) On the use of neural networks to evaluate groundwater levels in fractured media. Journal of Hydrology. 307:92-111. doi:http://dx.doi.org/10.1016/j.jhydrol.2004.10.005

[19] Lee T, Ouarda TBMJ (2011) Prediction of climate nonstationary oscillation processes with empirical mode decomposition. Journal of Geophysical Research: Atmospheres. 116: D06107. doi:10.1029/2010JD015142

[20] Liao H, Sun W (2010) Forecasting and Evaluating Water Quality of Chao Lake based on an Improved Decision Tree Method Procedia. Environmental Sciences. 2:970-979. doi:http://dx.doi.org/10.1016/j.proenv.2010.10.109

[21] Mallat SG (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 11:674-693. doi:10.1109/34.192463

[22] Nayak PC, Rao YRS, Sudheer KP (2006) Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach. Water Resources Management. 20:77-90. doi:10.1007/s11269-006-4007-z

[23] Nourani V, Alami MT, Aminfar MH (2009a) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Engineering Applications of Artificial Intelligence. 22:466-472. doi:https://doi.org/10.1016/j.engappai.2008.09.003

[24] Nourani V, Hosseini Baghanam A, Adamowski J, Kisi O (2014) Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. Journal of Hydrology. 514:358-377. doi:https://doi.org/10.1016/j.jhydrol.2014.03.057

[25] Nourani V, Komasi M, Mano A (2009b) A Multivariate ANN-Wavelet Approach for Rainfall–Runoff Modeling. Water Resources Management. 23:2877. doi:10.1007/s11269-009-9414-5

[26] Oh Y-Y, Yun S-T, Yu S, Hamm S-Y (2017) The combined use of dynamic factor analysis and wavelet analysis to evaluate latent factors controlling complex groundwater level fluctuations in a riverside alluvial aquifer. Journal of Hydrology. 555:938-955. doi:https://doi.org/10.1016/j.jhydrol.2017.10.070

[27] Ouachani R, Zoubeida B, Taha O (2013) Power of teleconnection patterns on precipitation and streamflow variability of upper Medjerda Basin International. Journal of Climatology. 33:58-76 doi:doi:10.1002/joc.3407

[28] Partal T, Kişi Ö (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. Journal of Hydrology. 342:199-212. doi:https://doi.org/10.1016/j.jhydrol.2007.05.026

[29] Quinlan JR (1986) Induction of decision trees. Machine Learning. 1:81-106. doi:10.1007/bf00116251

[30] Quinlan JR (1992) Learning with Continuous Classes Proceedings of Australian Joint Conference on Artificial Intelligence, Hobart, Australia, World Scientific, Singapore:343-348

[31] Raza K (2015) M5 Model Tree and Gene Expression Programming for the Prediction of Metrological Parameters. Paper presented at the Proceeding of IEEE 2015 International Conference on Computers, Communications, and Systems (ICCCS-2015), Nov 2-3, 2015, Kanyakumari, India,

[32] Reghunath R, Murthy TRS, Raghavan BR (2005) Time Series Analysis to Monitor and Assess Water Resources: A Moving Average Approach. Environmental Monitoring and Assessment. 109:65-72. doi:10.1007/s10661-005-5838-4

[33] Sahoo GB, Ray C, Wade HF (2005) Pesticide prediction in ground water in North Carolina domestic wells using artificial neural networks. Ecological Modelling. 183:29-46. doi:http://dx.doi.org/10.1016/j.ecolmodel.2004.07.021

[34] Shiri J, Kişi Ö (2011) Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Computers & Geosciences. 37:1692-1701. doi:http://dx.doi.org/10.1016/j.cageo.2010.11.010

[35] Shiri J, Kisi O, Yoon H, Lee K-K, Hossein Nazemi A (2013) Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques. Computers & Geosciences. 56:32-44. doi:https://doi.org/10.1016/j.cageo.2013.01.007

[36] Shoaib M, Shamseldin AY, Melville BW, Khan MM (2015) Runoff forecasting using hybrid Wavelet Gene Expression Programming (WGEP) approach. Journal of Hydrology. 527:326-344. doi:http://dx.doi.org/10.1016/j.jhydrol.2015.04.072

[37] Sivapragasam C, Kannabiran K, Karthik G, Raja S (2015) Assessing Suitability of GP Modeling for Groundwater Level. Aquatic Procedia. 4:693-699. doi:https://doi.org/10.1016/j.aqpro.2015.02.089

[38] Solgi A, Nourani V, Bagherian Marzouni M (2017a) Evaluation of nonlinear models for precipitation forecasting. Hydrological Sciences Journal. 62:2695-2704. doi:10.1080/02626667.2017.1392529

[39] Solgi A, Pourhaghi A, Bahmani R, Zarei H (2017b) Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD). Ecohydrology & Hydrobiology. 17:164-175. doi:https://doi.org/10.1016/j.ecohyd.2017.02.002

[40] Solgi A, Pourhaghi A, Bahmani R, Zarei H (2017c) Pre-processing data using wavelet transform and PCA based on support vector regression and gene expression programming for river flow simulation. Journal of Earth System Science. 126:65. doi:10.1007/s12040-017-0850-y

[41] Solomatine DP, Xue Y (2004) M5 Model Trees and Neural Networks: Application to Flood Forecasting in the Upper Reach of the Huai River in China. Journal of Hydrologic Engineering. 9:491-501. doi:doi:10.1061/(ASCE)1084-0699(2004)9:6(491)

[42] Thuillard M (2001) Wavelets in Soft Computing. vol 25. Siemens Building Technologies, Switzerland.

8. Figure Captions

9. Table Captions

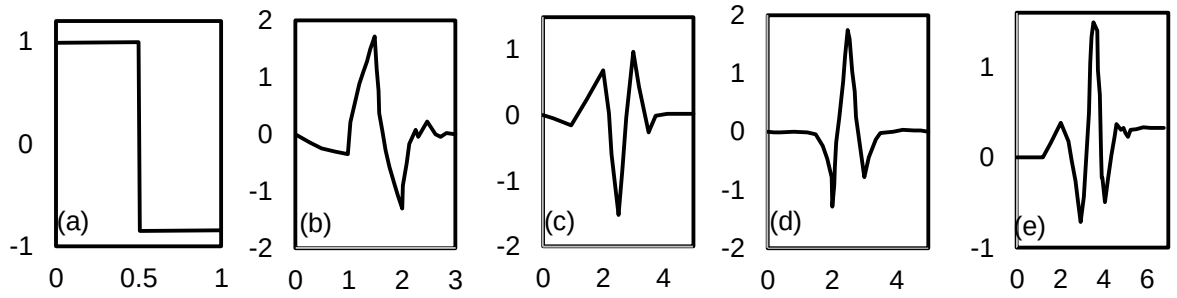


Fig. 3.1: a) Haar wavelet. b) Db2 wavelet. c) Sym3 wavelet. d) Coif1 wavelet. e) Db4 wavelet

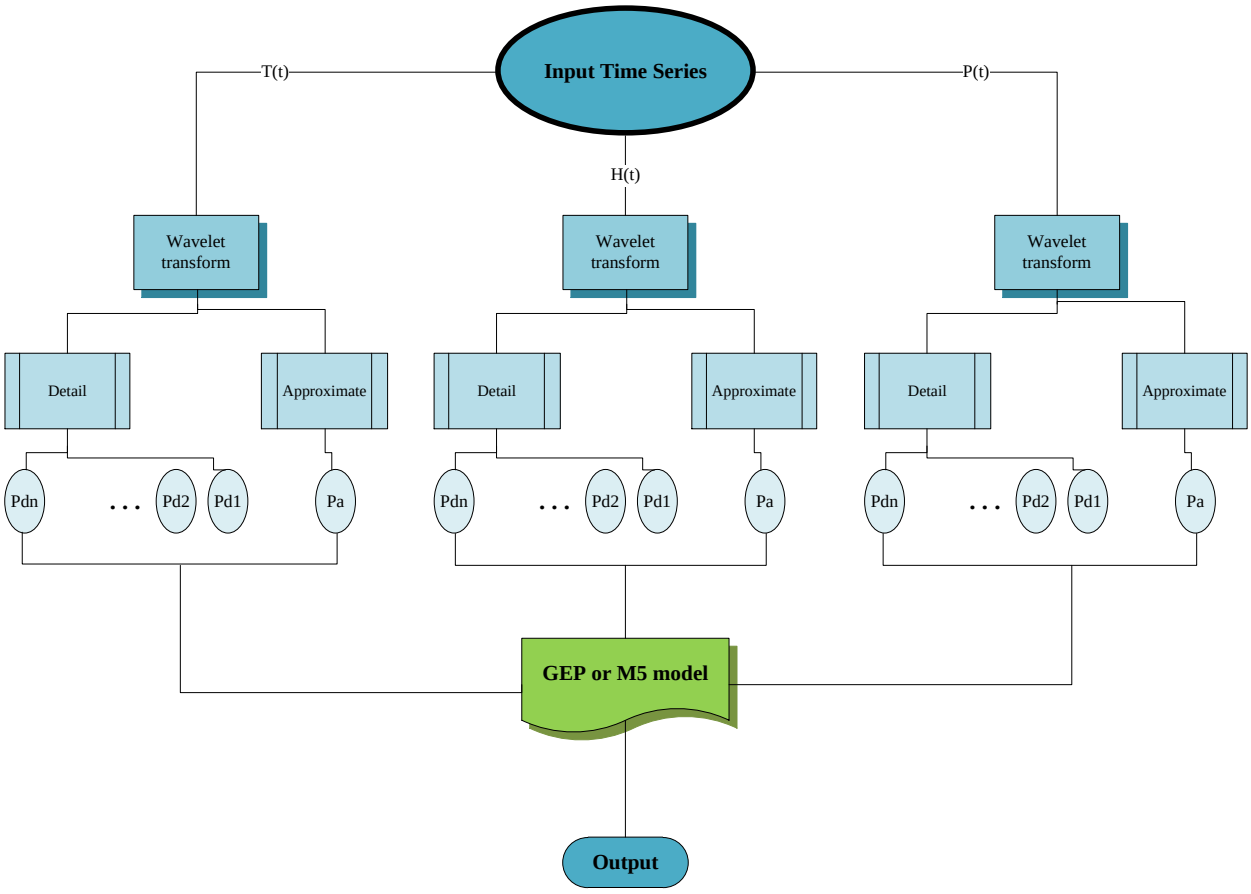


Fig. 3.2 Schematic diagram of WGEP and WM5 models

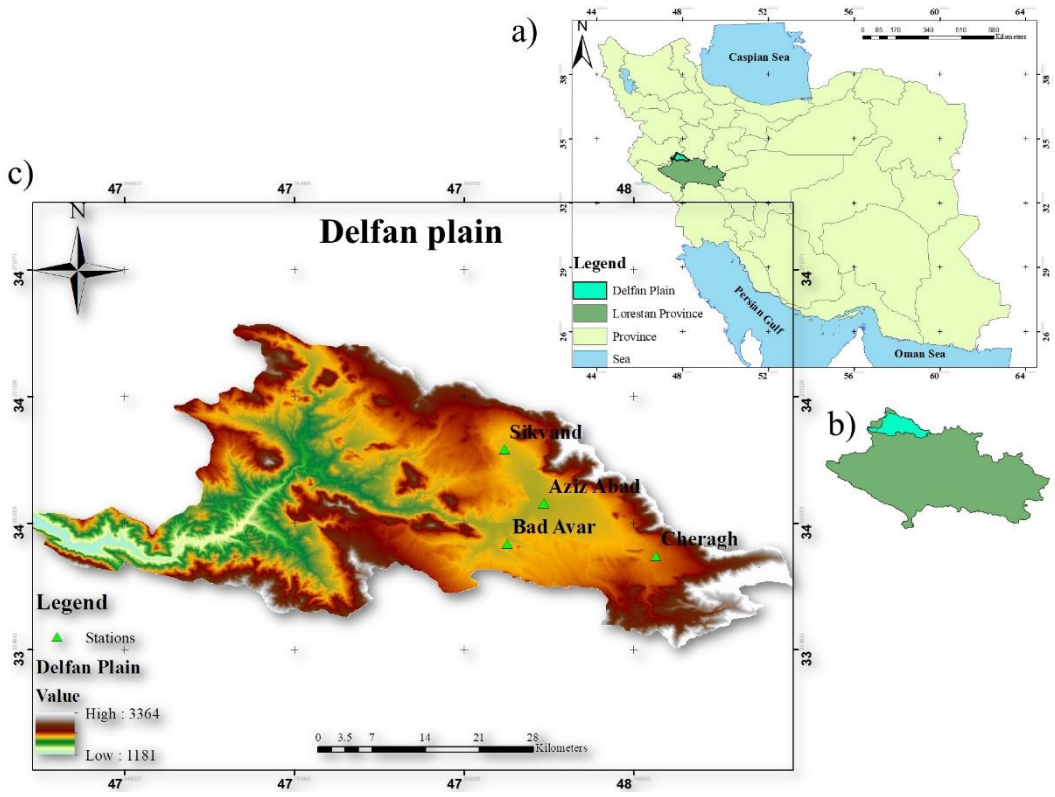


Fig. 3.3 Location of the three observation wells and one meteorological station in a) Iran, b) Lorestan province, c) Delfan Plain

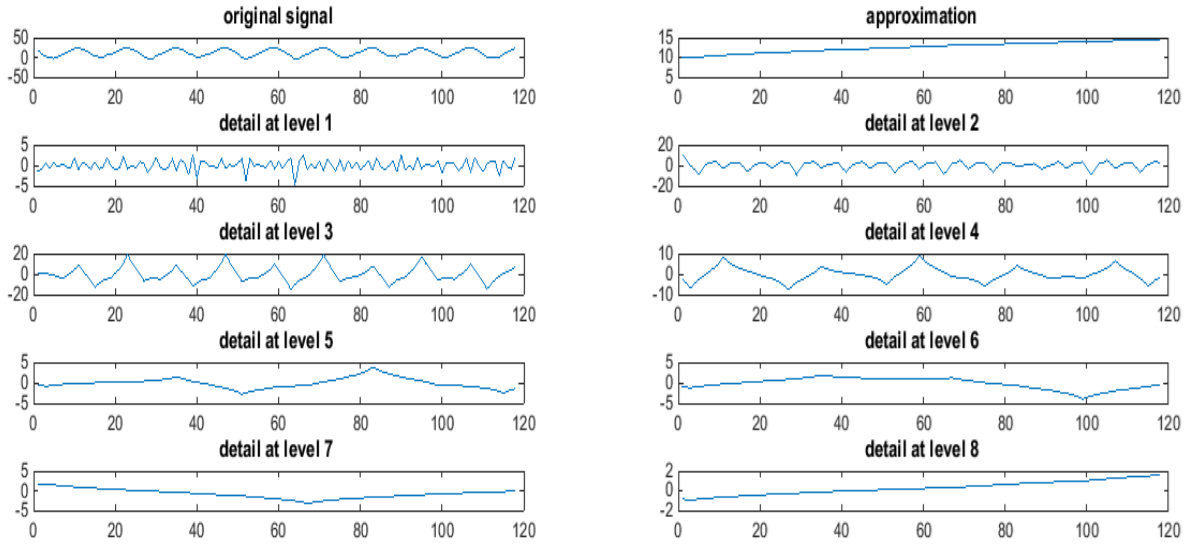
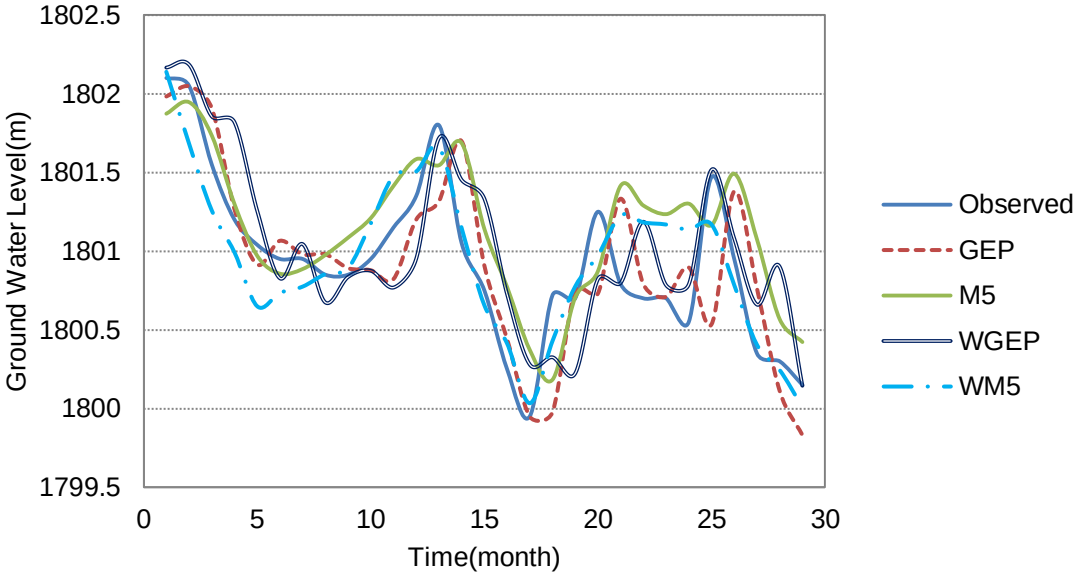
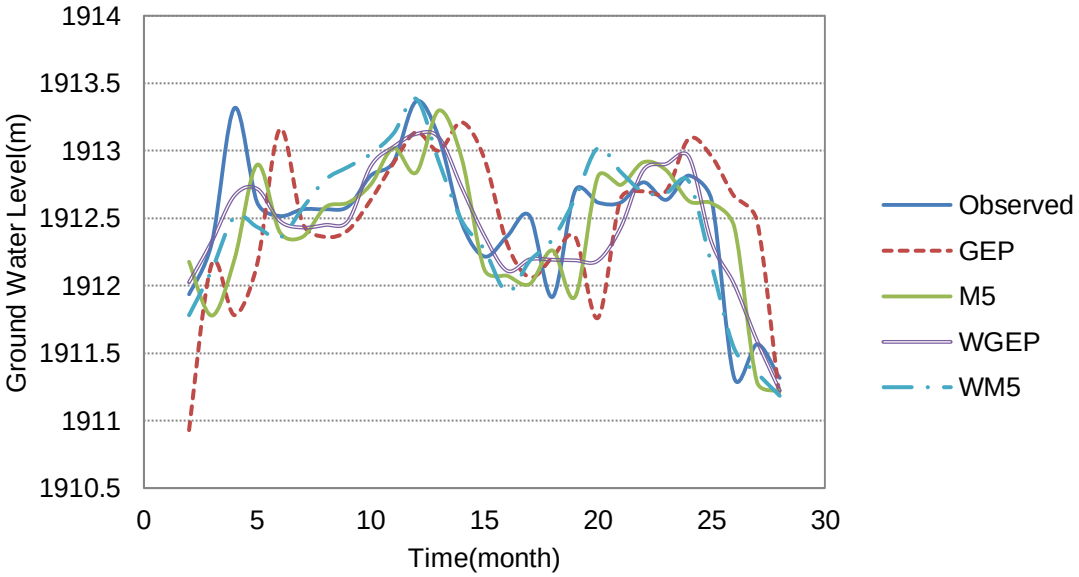


Fig. 3.4 Temperature time series sub-signals with the wavelet function of coif1, level 8

a)


1802.5
1802
1801.5
1801
1800.5
1800
1799.5
Ground Water Level(m)
0
5
10
15
20
25
30
Time(month)
Observed
GEP
M5
WGEP
WM5

b)


1914
1913.5
1913
1912.5
1912
1911.5
1911
1910.5
Ground Water Level(m)
0
5
10
15
20
25
30
Time(month)
Observed
GEP
M5
WGEP
WM5

c)

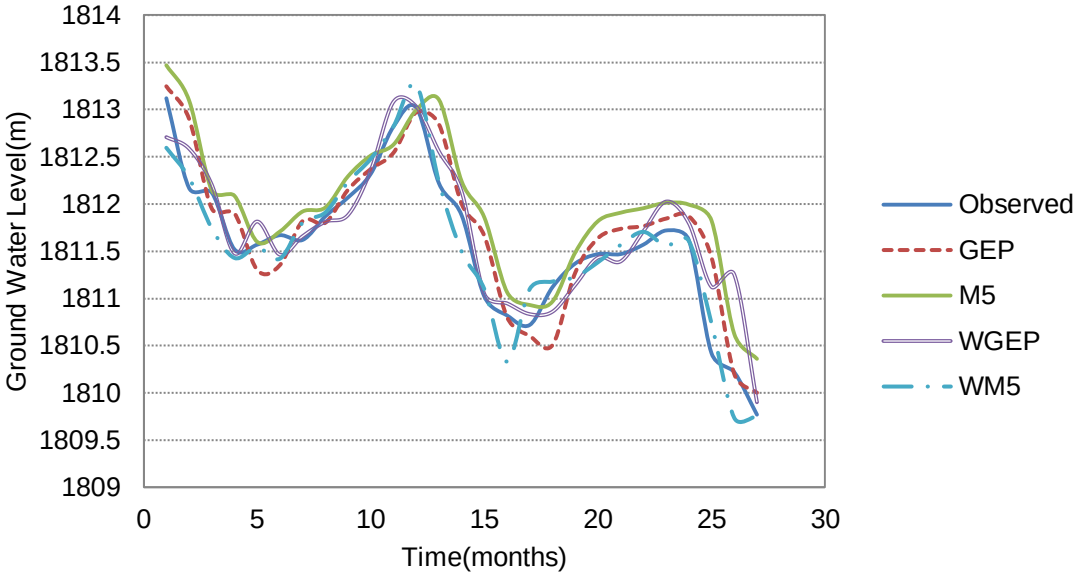


Fig. 3.5 Simulated and observed values for a) well 1, b) well 2 and c) well 3

Table 3.1. Features of the observation wells and meteorological station

Name of stations	Id	Variables	Average	Maximum	Minimum
Aziz Abad	1	Groundwater level (m)	1801.2	1802.5	1798.43
Cheragh	2	Groundwater level (m)	1912.4	1914.3	1909
Sikvand	3	Groundwater level (m)	1813.37	1815.73	1809.77
Bad Avar	-	Precipitation (mm)	48.2	291	0.0
Bad Avar	-	Temperature (°C)	11.7	25.4	-4.8

Table 3.2 Combinations of variables with time lags

Combination	Input	Combination	Input
1	H_t	9	H_{t-2}, H_{t-1}, H_t, P_{t-2}, P_{t-1}, P_t
2	H_{t-1}, H_t	10	H_t, T_t
3	H_{t-2}, H_{t-1}, H_t	11	H_{t-1}, H_t, T_{t-1}, T_t
4	P_t,	12	H_{t-2}, H_{t-1}, H_t, T_{t-2}, T_{t-1}, T_t
5	P_{t-1}, P_t	13	H_t, P_t, T_t
6	P_{t-2}, Pt_{-1}, P_t	14	H_{t-1}, H_t, P_{t-1}, P_t, T_{t-1}, T_t
7	H_t, P_t	15	H_{t-2}, H_{t-1}, H_t, P_{t-2}, Pt_{-1}, P_t, T_{t-2}, T_{t-1}, T_t
8	H_{t-1}, H_t, P_{t-1}, P_t		

Table 3.3. Evaluation criteria of calibration for GEP

Combination	Well 1					Well 2					Well 3				
	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
1	0.48	0.52	0.03	0.00	0.00	0.87	0.49	0.03	0.00	0.00	0.78	0.58	0.03	0.00	0.00
2	1.00	0.01	0.00	-0.01	0.00	0.88	0.47	0.03	0.01	0.00	0.79	0.55	0.03	0.01	0.00
3	0.47	0.54	0.03	-0.01	0.00	0.87	0.49	0.03	0.00	0.00	0.79	0.56	0.03	0.00	0.00
4	0.00	0.76	0.04	0.00	0.00	0.00	1.34	0.07	0.00	0.00	0.05	1.17	0.06	0.01	0.00
5	0.01	43.45	2.41	29.46	1.64	0.87	75.57	4.00	-40.20	-2.10	0.00	1.46	1.46	0.06	0.00
6	0.11	0.04	0.04	-0.09	-0.01	0.00	1.34	0.07	-0.01	0.00	0.04	17.31	0.95	-0.37	-0.32
7	0.57	0.50	0.03	0.00	0.00	0.87	0.48	0.03	0.00	0.00	0.80	0.55	0.03	-0.02	0.00
8	0.57	0.47	0.03	-0.04	0.00	0.89	0.45	0.02	0.06	0.00	0.81	0.53	0.03	-0.01	0.00
9	0.52	0.51	0.03	0.00	0.00	0.88	0.47	0.02	0.01	0.00	0.83	0.50	0.03	0.05	0.00
10	0.53	0.50	0.03	-0.05	0.00	0.90	0.44	0.02	-0.01	0.00	0.81	0.54	0.01	0.00	0.00
11	0.63	0.43	0.02	-0.08	-0.01	0.90	0.43	0.02	0.01	0.00	0.80	0.54	0.03	-0.02	0.00
12	0.60	0.05	0.03	0.02	0.00	0.89	0.45	0.02	-0.06	0.00	0.83	0.50	0.03	-0.01	0.00
13	0.56	0.48	0.03	0.01	0.00	0.89	0.44	0.02	0.01	0.00	0.81	0.53	0.03	-0.01	0.00
14	0.56	0.48	0.03	-0.01	0.00	0.90	0.44	0.02	-0.01	0.00	0.81	0.54	0.03	0.00	0.00
15	0.64	0.41	0.02	0.05	0.00	0.92	0.39	0.02	-0.02	0.00	0.81	0.53	0.03	0.00	0.00

Table 3.4. Evaluation criteria of validation for GEP

Combination	Well 1					Well 2					Well 3				
	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
1	0.56	0.38	0.02	-0.09	-0.01	0.33	0.52	0.03	0.04	0.00	0.71	0.46	0.03	-0.04	0.00
2	0.56	0.38	0.02	-0.08	0.00	0.33	0.54	0.03	0.02	0.00	0.63	0.51	0.03	-0.05	0.00
3	0.58	0.39	0.02	-0.05	0.00	0.31	0.54	0.03	0.02	0.00	0.64	0.51	0.03	-0.05	0.00
4	0.21	0.57	0.03	-0.29	-0.02	0.31	0.54	0.03	0.02	0.00	0.01	2.32	0.13	-2.16	-0.12
5	0.11	41.68	2.31	33.86	1.88	0.00	74.85	3.91	-46.80	-2.44	0.20	2.09	0.12	-1.86	-0.10
6	0.09	0.65	0.04	-0.39	-0.01	0.00	0.55	0.03	0.04	0.00	0.19	16.32	0.90	-6.85	-0.38
7	0.60	0.42	0.02	-0.05	0.00	0.32	0.54	0.03	0.06	0.00	0.77	0.43	0.02	-0.03	0.00
8	0.32	0.69	0.04	-0.14	-0.01	0.33	0.54	0.03	0.08	0.00	0.70	0.47	0.03	-0.08	0.00
9	0.58	0.40	0.02	-0.05	0.00	0.33	0.53	0.03	0.04	0.00	0.63	0.53	0.03	-0.02	0.00
10	0.60	0.38	0.02	-0.10	-0.01	0.37	0.53	0.03	0.06	0.00	0.77	0.41	0.01	-0.05	0.00
11	0.62	0.38	0.02	-0.11	-0.01	0.35	0.55	0.03	0.05	0.00	0.66	0.49	0.03	-0.07	0.00
12	0.27	0.70	0.04	0.07	0.00	0.33	0.52	0.03	0.01	0.00	0.69	0.48	0.03	-0.07	0.00
13	0.62	0.40	0.04	0.06	0.00	0.40	0.51	0.03	0.06	0.00	0.81	0.38	0.02	-0.06	0.00
14	0.61	0.40	0.02	-0.06	0.00	0.36	0.54	0.03	0.05	0.00	0.77	0.43	0.02	-0.04	0.00
15	0.63	0.39	0.02	0.01	0.00	0.38	0.58	0.03	0.07	0.00	0.53	0.68	0.04	-0.07	0.00

Table 3.5. Evaluation criteria of calibration for M5

Combination	Well 1					Well 2					Well 3				
	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
1	0.47	0.48	0.03	0.05	0.00	0.87	0.48	0.03	-0.01	0.00	0.78	0.56	0.03	0.04	0.00
2	0.50	0.50	0.03	-0.16	-0.01	0.87	0.48	0.03	0.05	0.00	0.83	0.50	0.03	-0.03	0.00
3	0.67	0.39	0.02	-0.04	0.00	0.88	0.49	0.03	-0.14	-0.01	0.83	0.50	0.03	0.06	0.00
4	0.25	0.58	0.03	0.02	0.00	0.06	1.30	0.07	0.00	0.00	0.00	1.18	--	--	--
5	0.24	0.58	0.03	0.00	0.00	0.11	1.26	0.07	0.00	0.00	0.04	1.17	0.06	0.00	0.00
6	0.31	0.56	0.03	0.00	0.00	0.17	1.22	0.06	0.00	0.00	0.09	1.14	0.06	0.00	0.00
7	0.53	0.46	0.03	-0.03	0.00	0.87	0.49	0.03	-0.09	-0.01	0.80	0.54	0.03	0.08	0.00
8	0.76	0.32	0.02	0.01	0.00	0.88	0.47	0.02	0.01	0.00	0.84	0.48	0.03	0.01	0.00
9	0.77	0.33	0.02	0.01	0.00	0.88	0.46	0.02	0.01	0.00	0.87	0.47	0.03	0.15	0.01
10	0.59	0.43	0.02	-0.03	0.00	0.89	0.45	0.02	0.09	0.01	0.80	0.54	0.03	-0.06	0.00
11	0.70	0.37	0.02	-0.07	0.00	0.90	0.44	0.02	0.08	0.00	0.82	0.50	0.03	-0.03	0.00
12	0.80	0.31	0.02	-0.07	0.00	0.89	0.43	0.02	0.00	0.00	0.87	0.43	0.02	0.02	0.00
13	0.59	0.43	0.02	-0.03	0.00	0.89	0.45	0.02	0.09	0.01	0.80	0.54	0.03	0.08	0.00
14	0.70	0.37	0.02	-0.07	0.00	0.93	0.37	0.02	0.10	0.01	0.84	0.48	0.03	0.01	0.00
15	0.80	0.31	0.02	-0.07	0.00	0.90	0.42	0.02	0.01	0.00	0.89	0.42	0.02	-0.10	-0.01

Table 3.6. Evaluation criteria of validation for M5

Combination	Well 1					Well 2					Well 3				
	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
1	0.59	0.37	0.02	-0.07	0.00	0.34	0.50	0.03	0.01	0.00	0.72	0.52	0.03	-0.24	-0.01
2	0.57	0.47	0.03	-0.28	-0.02	0.24	0.51	0.03	0.09	0.01	0.34	0.72	0.04	-0.23	-0.01
3	0.37	0.70	0.04	0.06	0.00	0.30	0.49	0.03	-0.15	-0.01	0.40	0.67	0.04	-0.29	-0.02
4	0.26	0.52	0.03	-0.20	-0.01	0.01	0.61	0.03	0.05	0.00	0.00	--	--	--	--
5	0.18	0.56	0.03	-0.23	-0.01	0.01	0.61	0.03	0.14	0.01	0.17	2.35	0.13	-2.22	-0.12
6	0.12	0.59	0.03	-0.27	-0.02	0.01	0.67	0.03	0.18	0.01	0.06	2.40	0.13	-2.28	-0.13
7	0.55	0.41	0.02	-0.14	-0.01	0.33	0.52	0.03	-0.06	0.00	0.75	0.46	0.03	-0.16	-0.01
8	0.36	0.78	0.04	0.13	0.01	0.25	0.52	0.03	0.07	0.00	0.67	0.54	0.03	-0.19	-0.01
9	0.36	0.79	0.04	0.11	0.01	0.32	0.45	0.02	0.00	0.00	0.50	0.62	0.03	-0.05	0.00
10	0.62	0.39	0.02	-0.17	-0.01	0.38	0.53	0.03	0.13	0.01	0.79	0.49	0.03	-0.29	-0.02
11	0.31	0.66	0.04	-0.14	-0.01	0.31	0.53	0.03	0.14	0.01	0.70	0.54	0.03	-0.28	-0.02
12	0.32	0.62	0.03	-0.18	-0.01	0.35	0.46	0.02	0.03	0.00	0.59	0.65	0.04	-0.40	-0.02
13	0.62	0.39	0.02	-0.17	-0.01	0.38	0.53	0.03	0.13	0.01	0.75	0.46	0.03	-0.16	-0.01
14	0.31	0.66	0.04	-0.14	-0.01	0.27	0.77	0.04	0.23	0.01	0.67	0.54	0.03	-0.19	-0.01
15	0.32	0.62	0.03	-0.18	-0.01	0.41	0.43	0.02	0.05	0.00	0.55	0.68	0.04	-0.43	-0.02

Table 3.7. Evaluation criteria of calibration for WGEP

Wavelet function	Level	Well 1					Well 2					Well 3				
		R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
Coif1	**1**	0.62	0.43	0.02	-0.01	0.00	0.94	0.34	0.02	-0.01	0.00	0.90	0.38	0.02	0.00	0.00
	2	0.68	0.38	0.02	0.02	0.00	0.94	0.33	0.02	-0.03	0.00	0.92	0.33	0.02	-0.03	0.00
	3	0.66	0.39	0.02	-0.02	0.00	0.94	0.32	0.02	0.01	0.00	0.90	0.37	0.02	-0.01	0.00
	4	0.71	0.36	0.02	-0.03	0.00	0.93	0.35	0.02	-0.03	0.00	0.87	0.45	0.02	-0.08	0.00
Db2	**1**	0.80	0.30	0.02	0.00	0.00	0.93	0.36	0.02	0.00	0.00	0.92	0.33	0.02	0.01	0.00
	2	0.73	0.35	0.02	-0.01	0.00	0.94	0.34	0.02	-0.02	0.00	0.91	0.35	0.02	0.04	0.00
	3	0.74	0.34	0.02	0.00	0.00	0.92	0.38	0.02	0.00	0.00	0.92	0.34	0.02	0.00	0.00
	4	0.84	0.28	0.02	-0.02	0.00	0.94	0.33	0.02	0.05	0.00	0.90	0.38	0.02	-0.03	0.00
Db4	**1**	0.74	0.37	0.02	-0.03	0.00	0.93	0.35	0.02	-0.02	0.00	0.91	0.37	0.02	-0.11	-0.01
	2	0.71	0.36	0.02	0.00	0.00	0.94	0.33	0.02	0.00	0.00	0.92	0.35	0.02	0.05	0.00
	3	0.74	0.34	0.02	0.01	0.00	0.95	0.31	0.02	-0.01	0.00	0.93	0.31	0.02	-0.01	0.00
	4	0.70	0.37	0.02	0.01	0.00	0.93	0.35	0.02	-0.03	0.00	0.90	0.37	0.02	-0.01	0.00
Haar	**1**	0.66	0.41	0.02	0.04	0.00	0.90	0.42	0.02	-0.01	0.00	0.90	0.38	0.02	0.01	0.00
	2	0.58	0.44	0.02	0.00	0.00	0.93	0.36	0.02	0.00	0.00	0.90	0.38	0.02	-0.01	0.00
	3	0.40	0.52	0.03	0.06	0.00	0.93	0.36	0.02	-0.05	0.00	0.92	0.35	0.02	0.09	0.00
	4	0.77	0.33	0.02	-0.02	0.00	0.93	0.36	0.02	0.00	0.00	0.89	0.40	0.02	0.01	0.00
Sym3	**1**	0.80	0.31	0.02	-0.06	0.00	0.94	0.35	0.02	0.03	0.00	0.92	0.34	0.02	-0.03	0.00
	2	0.76	0.33	0.02	-0.01	0.00	0.93	0.36	0.02	0.01	0.00	0.92	0.34	0.02	0.00	0.00
	3	0.75	0.34	0.02	-0.04	0.00	0.92	0.39	0.02	-0.10	-0.01	0.92	0.33	0.02	0.01	0.00
	4	0.75	0.33	0.02	-0.02	0.00	0.90	0.42	0.02	-0.05	0.00	0.92	0.34	0.02	-0.01	0.00

Table 3.8. Evaluation criteria of validation for WGEP

Wavelet function	Level	Well 1					Well 2					Well 3				
		R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
Coif1	**1**	0.68	0.34	0.02	-0.09	0.00	0.70	0.34	0.02	-0.01	0.00	0.69	0.52	0.03	-0.09	0.00
	2	0.63	0.33	0.02	-0.06	0.00	0.64	0.36	0.02	-0.02	0.00	0.79	0.45	0.02	-0.13	-0.01
	3	0.62	0.37	0.02	-0.11	-0.01	0.55	0.42	0.02	-0.02	0.00	0.79	0.45	0.02	-0.14	-0.01
	4	0.50	0.38	0.02	-0.05	0.00	0.59	0.40	0.02	0.03	0.00	0.73	0.52	0.03	-0.15	-0.01
Db2	**1**	0.70	0.34	0.02	-0.06	0.00	0.62	0.38	0.02	0.03	0.00	0.80	0.42	0.02	-0.03	0.00
	2	0.69	0.32	0.02	-0.08	0.00	0.69	0.33	0.02	0.00	0.00	0.79	0.42	0.02	-0.03	0.00
	3	0.41	0.48	0.03	-0.11	-0.01	0.76	0.30	0.02	0.05	0.00	0.82	0.39	0.02	0.00	0.00
	4	0.70	0.32	0.02	-0.07	0.00	0.54	0.42	0.02	0.12	0.01	0.68	0.55	0.03	-0.15	-0.01
Db4	**1**	0.69	0.35	0.02	-0.14	-0.01	0.68	0.36	0.02	-0.01	0.00	0.82	0.43	0.02	-0.17	-0.01
	2	0.70	0.32	0.02	-0.08	0.00	0.63	0.37	0.02	-0.01	0.00	0.82	0.38	0.02	0.00	0.00
	3	0.67	0.32	0.02	-0.05	0.00	0.48	0.46	0.02	-0.05	0.00	0.90	0.31	0.02	-0.09	-0.01
	4	0.65	0.33	0.02	-0.05	0.00	0.43	0.50	0.03	-0.03	0.00	0.30	0.85	0.05	-0.29	-0.02
Haar	**1**	0.60	0.39	0.02	-0.01	0.00	0.51	0.44	0.02	0.03	0.00	0.77	0.45	0.02	-0.04	0.00
	2	0.60	0.36	0.02	-0.08	0.00	0.52	0.43	0.02	0.04	0.00	0.80	0.41	0.02	-0.02	0.00
	3	0.64	0.34	0.02	0.02	0.00	0.54	0.42	0.02	0.01	0.00	0.82	0.39	0.02	0.04	0.00
	4	0.60	0.35	0.02	-0.06	0.00	0.53	0.42	0.02	0.01	0.00	0.74	0.47	0.03	-0.07	0.00
Sym3	**1**	0.53	0.45	0.02	-0.02	0.00	0.63	0.43	0.02	0.12	0.01	0.80	0.42	0.02	-0.08	0.00
	2	0.60	0.34	0.02	-0.05	0.00	0.59	0.43	0.02	0.04	0.00	0.77	0.45	0.02	-0.05	0.00
	3	0.59	0.44	0.02	-0.17	-0.01	0.64	0.37	0.02	-0.07	0.00	0.71	0.50	0.03	-0.08	0.00
	4	0.52	0.39	0.02	-0.14	-0.01	0.58	0.40	0.02	0.00	0.00	0.68	0.53	0.03	-0.10	-0.01

Table 3.9. Evaluation criteria of calibration for WM5

Wavelet function	Level	Well 1					Well 2					Well 3				
		R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
Coif1	1	0.82	0.29	0.02	0.08	0.00	0.93	0.37	0.02	0.09	0.00	0.90	0.39	0.02	0.07	0.00
	2	0.82	0.28	0.02	-0.02	0.00	0.94	0.33	0.02	0.09	0.00	0.91	0.36	0.02	-0.07	0.00
	3	0.71	0.36	0.02	0.06	0.00	0.95	0.30	0.02	0.03	0.00	0.91	0.37	0.02	-0.06	0.00
	4	0.81	0.29	0.02	-0.06	0.00	0.95	0.31	0.02	0.03	0.00	0.91	0.36	0.02	-0.04	0.00
Db2	1	0.57	0.57	0.03	0.08	0.00	0.92	0.37	0.02	-0.02	0.00	0.92	0.34	0.02	-0.03	0.00
	2	0.70	0.38	0.02	0.09	0.01	0.94	0.33	0.02	-0.09	-0.01	0.93	0.32	0.02	0.04	0.00
	3	0.82	0.28	0.02	0.02	0.00	0.94	0.33	0.02	-0.08	0.00	0.93	0.31	0.02	0.00	0.00
	4	0.86	0.25	0.01	-0.04	0.00	0.95	0.31	0.02	-0.02	0.00	0.93	0.32	0.02	0.02	0.00
Db4	1	0.71	0.36	0.02	-0.04	0.00	0.95	0.30	0.02	0.04	0.00	0.93	0.32	0.02	0.06	0.00
	2	0.79	0.31	0.02	-0.05	0.00	0.94	0.34	0.02	-0.02	0.00	0.94	0.30	0.02	0.01	0.00
	3	0.85	0.26	0.01	0.00	0.00	0.94	0.33	0.02	0.03	0.00	0.93	0.32	0.02	0.07	0.00
	4	0.82	0.29	0.02	0.05	0.00	0.94	0.32	0.02	0.01	0.00	0.93	0.32	0.02	0.01	0.00
Haar	1	0.68	0.38	0.02	-0.02	0.00	0.92	0.38	0.02	0.01	0.00	0.89	0.39	0.02	0.02	0.00
	2	0.78	0.31	0.02	0.06	0.00	0.92	0.37	0.02	0.06	0.00	0.89	0.40	0.02	0.06	0.00
	3	0.77	0.32	0.02	0.05	0.00	0.93	0.37	0.02	0.08	0.00	0.90	0.25	0.01	-0.03	0.00
	4	0.60	0.42	0.02	0.01	0.00	0.92	0.37	0.02	0.00	0.00	0.84	0.48	0.03	0.09	0.00
Sym3	1	0.86	0.25	0.01	0.01	0.01	0.93	0.37	0.02	-0.09	-0.01	0.93	0.33	0.02	0.09	0.00
	2	0.87	0.24	0.01	-0.05	0.00	0.94	0.35	0.02	-0.08	0.00	0.93	0.31	0.02	0.01	0.00
	3	0.85	0.26	0.01	0.00	0.00	0.95	0.29	0.01	-0.01	0.00	0.93	0.32	0.02	0.09	0.00
	4	0.84	0.27	0.01	-0.01	0.00	0.95	0.30	0.02	-0.02	0.00	0.91	0.37	0.02	0.07	0.00

Table 3.10. Evaluation criteria of validation for WM5

Wavelet function	Level	Well 1					Well 2					Well 3				
		R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS	R^2	RMSE (M)	rRMSE	BIAS (M)	rBIAS
Coif1	1	0.65	0.83	0.05	0.36	0.02	0.72	0.35	0.02	0.14	0.01	0.87	0.32	0.02	0.06	0.00
	2	0.60	0.47	0.03	-0.03	0.00	0.68	0.36	0.02	0.15	0.01	0.83	0.34	0.02	-0.01	0.00
	3	0.77	0.26	0.01	0.00	0.00	0.66	0.36	0.02	0.10	0.01	0.83	0.33	0.02	-0.03	0.00
	4	0.43	0.55	0.03	-0.32	-0.02	0.62	0.37	0.02	0.07	0.00	0.82	0.35	0.02	-0.06	0.00
Db2	1	0.63	1.10	0.06	0.53	0.03	0.53	0.41	0.02	0.00	0.00	0.91	0.26	0.01	0.01	0.00
	2	0.47	0.54	0.03	0.12	0.01	0.64	0.35	0.02	-0.03	0.00	0.89	0.31	0.02	0.16	0.01
	3	0.60	0.36	0.02	-0.08	0.00	0.62	0.36	0.02	0.00	0.00	0.85	0.34	0.02	0.07	0.00
	4	0.60	0.39	0.02	-0.18	-0.01	0.73	0.31	0.02	0.05	0.00	0.84	0.37	0.02	0.15	0.01
Db4	1	0.64	0.35	0.02	-0.11	-0.01	0.63	0.47	0.02	0.12	0.01	0.87	0.33	0.02	0.09	0.01
	2	0.60	0.44	0.02	0.00	0.00	0.52	0.43	0.02	0.01	0.00	0.85	0.34	0.02	0.07	0.00
	3	0.42	0.48	0.03	-0.12	-0.01	0.66	0.37	0.02	0.10	0.00	0.85	0.34	0.02	0.06	0.00
	4	0.30	0.76	0.04	-0.03	0.00	0.45	0.49	0.03	0.12	0.01	0.85	0.39	0.02	-0.03	0.00
Haar	1	0.77	0.27	0.02	-0.08	-0.01	0.70	0.33	0.02	0.07	0.00	0.91	0.25	0.01	0.06	0.00
	2	0.55	0.40	0.02	-0.12	-0.01	0.68	0.35	0.02	0.11	0.01	0.90	0.27	0.02	0.12	0.01
	3	0.58	0.36	0.02	-0.04	0.00	0.67	0.36	0.02	0.14	0.01	0.89	0.41	0.02	-0.09	-0.01
	4	0.46	0.50	0.03	-0.27	-0.02	0.63	0.39	0.02	0.14	0.01	0.71	0.60	0.03	0.42	0.02
Sym3	1	0.38	0.81	0.04	0.09	0.00	0.49	0.44	0.02	-0.06	0.00	0.88	0.33	0.02	0.12	0.01
	2	0.74	0.31	0.02	-0.13	0.47	0.44	0.02	-0.04	0.00	0.47	0.87	0.31	0.02	0.06	0.00
	3	0.67	0.34	0.02	-0.12	-0.01	0.61	0.37	0.02	0.08	0.00	0.86	0.34	0.02	0.16	0.01
	4	0.30	0.49	0.03	-0.13	-0.01	0.45	0.44	0.02	0.03	0.00	0.71	0.58	0.03	0.37	0.02

Table 11 Comparison of different models

Observation Wells	Model Type	R2		RMSE	
		Calibration	Validation	Calibration	Validation
1	GEP	0.64	0.63	0.41	0.39
	M5	0.59	0.62	0.43	0.39
	WGEP	0.71	0.70	0.36	0.32
	WM5	0.71	0.77	0.36	0.26
2	GEP	0.89	0.40	0.44	0.51
	M5	0.90	0.41	0.37	0.43
	WGEP	0.94	0.76	0.34	0.30
	WM5	0.95	0.73	0.31	0.31
3	GEP	0.81	0.81	0.53	0.38
	M5	0.80	0.79	0.54	0.49
	WGEP	0.93	0.90	0.31	0.31
	WM5	0.92	0.91	0.34	0.26

CHAPITRE III: GROUNDWATER LEVEL SIMULATION USING GENE EXPRESSION PROGRAMMING AND M5 MODEL TREE COMBINED WITH ENSEMBLE EMPIRICAL MODE DECOMPOSITION

Groundwater Level Simulation Using Gene Expression Programming and M5 Model Tree Combined with Ensemble Empirical Mode Decomposition

Ramin Bahmani[1], Taha B. M. J. Ouarda[1]

[1] Canada Research Chair in Statistical Hydro-climatology, INRS-ETE, Québec (Québec), Canada.

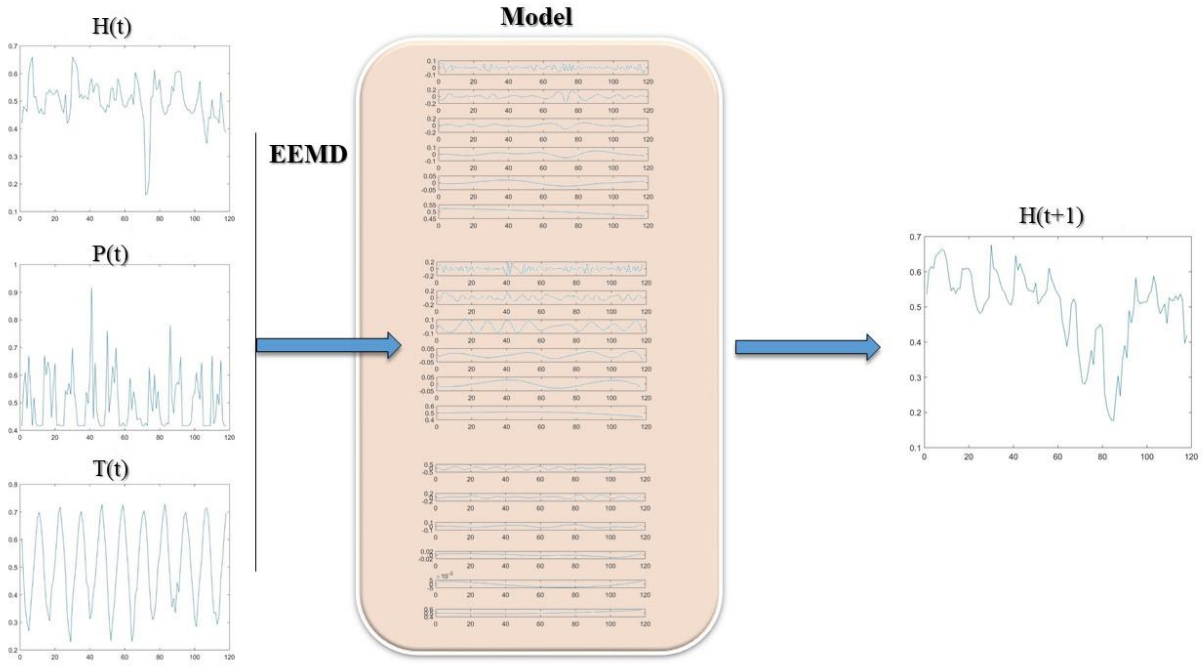
Journal : To be submitted to Environmental Modelling & Software

Abstract

It is necessary to simulate groundwater level as one of the important resources to supply water for human activities. Hence, in the current study, Gene Expression Programming (GEP) and M5 model tree (M5) are combined with Ensemble Empirical Mode Decomposition (EEMD) method to produce hybrid models for groundwater level simulation. Moreover, the performance of the hybrid models is compared with the outputs of GEP and M5 and the combined models with Wavelet transform (WT). The results indicate that the hybrid models are more accurate than simple GEP and M5 and WT is better than EEMD to produce hybrid models. GEP combined with WT is shown to be the most accurate model for groundwater water simulation, while M5 combined with EEMD is not recommended for the simulation due to weak performance.

Keywords: Ensemble Empirical Mode Decomposition; Groundwater level simulation; Gene Expression Programming; Hybrid Model; M5 model tree; Signal pre-processing methods.

Graphical abstract



Abbreviations

AI	Artificial Intelligence
ANFIS	Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
BIAS	Mean Bias
EEMD	Ensemble Empirical Mode Decomposition
EMD	Empirical Mode Decomposition
IMF	Intrinsic Mode Function
GA	Genetic Algorithm
GEP	Gene Expression Programming
GP	Genetic Programming
M5	M5 model tree
rBIAS	Relative Mean Bias
RMSE	Root Mean Squared Error
rRMSE	Relative Root Mean Squared Error
SVM	Support Vector Machine
WT	Wavelet transform

1. Introduction

One of the important resources to supply water for human activities such as farming or daily usage is groundwater (Nourani and Mousavi, 2016). Hence, it is necessary to simulate the groundwater level to investigate its availability for human activities. Due to the complexity of hydrological systems (Ouali et al., 2017), Artificial Intelligence (AI) models with high ability of complex phenomena modeling are used for groundwater level simulation (Sivapragasam et al., 2015). For instance, Daliakopoulos et al. (2005) and Taormina et al. (2012) utilized Artificial Neural Network (ANN) for groundwater level simulation and prediction. Shiri et al. (2013) employed ANN, Adaptive Neuro-Fuzzy Inference System (ANFIS), Gene Expression Programming (GEP), and Support Vector Machine (SVM) for groundwater simulation.

Despite the numerous applications of AI-based models, due to the high fluctuations of hydrologic time series, most of the models are not capable of simulating highly non-stationary time series. Thus, a number of studies, with the purpose of improving the models, have integrated methods to simulate highly non-stationary time series (Lee and Ouarda, 2012). The methods such as Wavelet transform (WT), Empirical Mode Decomposition (EMD) and ensemble EMD (EEMD) have been used to pre-process a time series and generate hybrid models (Kisi and Shiri, 2011; Nourani et al., 2009; Solgi et al., 2017; Wang et al., 2015b). For example, Karthikeyan and Nagesh Kumar (2013) improved the accuracy of auto-regressive models by using WT and EMD for pre-processing of non-stationary time series.

EEMD is a direct, intuitive, and self-adaptive method and is suitable for analyzing non-stationary hydrologic time series (Wang et al., 2015b). Research indicates that using EEMD improves the ability of models for simulation (Lee and Ouarda, 2011; Lee and Ouarda, 2012; Masselot et al., 2018). Huang et al. (2014) combined EEMD with SVM for monthly streamflow simulations and presented the high accuracy of the hybrid model. In another study, Jiao et al. (2016) illustrated the high performance of the hybrid SVM, combined with EEMD, for simulating hydrological data.

Since no study regarding the integration of EEMD with GEP and M5 model tree (M5) to simulate groundwater levels has been published, the objective of the present paper is to simulate groundwater levels by GEP and M5 combined with EEMD. To do so, EEMD is applied to decompose monthly groundwater levels, precipitation, and temperature time series into sub-series. Then, the sub-series are used as the inputs of GEP and M5 models to produce hybrid models to simulate groundwater level in one month ahead. Finally, the improved models are evaluated using different evaluation criteria.

2. Methods

In the current study, the hybrid models of EEMD and GEP (EEMD-GEP) plus EEMD and M5 (EEMD-M5) are employed for groundwater simulation. As there are many articles and books to introduce the EEMD, GEP and M5, the present paper does not delve into the method and models and their concepts are briefly presented.

2.1 Gene Expression Programming (GEP)

Ferreira (2001) introduced GEP based on the evolutionary algorithms for modeling. The evolutionary algorithms mimic the mechanism of living organisms. GEP exploits the advantages of Genetic Algorithm (GA) and Genetic Programming (GP) to overcome their disadvantages. In the GEP model, the genotype of chromosomes has a linear structure similar to GA, while the phenotype of chromosomes has a tree structure with different shapes and sizes like GP (Ferreira, 2001). The genotype is all or a part of the genetic compositions of a cell and the phenotype is observable characteristics of an organism.

2.2 M5 Model Tree (M5)

Quinlan (1992) presented the M5 according to decision tree to form a relationship between dependent and independent variables. M5 is the combination of linear regression and decision tree. Moreover, it can be applied to analyze quantitative and qualitative variables (Quinlan, 1992).

M5 utilizes a top-down approach to reach from a node at the top to a leaf at the down. Building M5 has two steps. The first step is dividing a node, called parent node, into nodes

with smaller standard deviation using the splitting criterion (Kisi, 2015). If the splitting criterion is not satisfied, the node is converted to a leaf. The second step is pruning the grown tree to avoid overfitting by replacing leaves by linear regression functions (Kisi, 2015).

For more details about M5, refer to Quinlan (1986) and Quinlan (1992). For more information about GEP, refer to Ferreira (2001), Ferreira (2002), and Ferreira (2006).

2.3 Empirical Mode Decomposition (EMD)

For the first time, Huang et al. (1998) introduced EMD as an empirical method to decompose a signal to its frequency components. The decomposed components are known as intrinsic mode functions (IMFs) if they meet the following conditions (Lee and Ouarda, 2012):

A. The number of local extrema should be equal to the number of zero crossing in case of any difference, just one is allowed.

B. The mean of upper envelop, calculated by local maxima, and lower envelop, calculated by local minima, should be zero.

The following steps demonstrate the decomposition of a time series (x(t)) through the sifting process by the EMD algorithm (Huang et al., 1998):

1. Determine the local maxima and minima of x(t).

2. Fit an upper envelope on the local maxima by a cubic spline.

3. Fit a lower envelope on the local minima by a cubic spline.

4. Calculate the mean (m(t)) of upper and lower envelop which is the low frequency of the signal. Then, subtract the m(t) from the x(t). $h_1(t) = x(t) - m_1(t)$ where $h_1(t)$ can be the first IMF.

5. Determine a criterion to stop the process. The criterion is called D_k and defined as $D_k=\frac{\sum_{t=0}^{T}|h_1^{k-1}(t)-h_1^k(t)|^2}{\sum_{t=0}^{T}|h_1^{k-1}(t)|^2}$ where k refers to the additional repetition number for the sifting process. So, k indicates the current $h_1(t)$ and k-1 indicates the previous $h_1(t)$. A typical value for D_k is between 0.2 to 0.3 (Huang et al., 1998).

6. If D_k is larger than the selected typical value, the x(t) is replaced by the signal calculated from the step 4. Then, the process is repeated from step 1 again.

7. If D_k is smaller than the selected typical value, $C_1 = h_1^k$ is known as the first IMF. Here, the subscript 1 refers to the first IMF and increases for each iteration of steps 1 to 5. The superscript k shows the number of sifting process.

8. Define the residual as $r_1 = x(t) - C_1^k$. If the residual satisfies the condition A, it is considered as x(t) and the steps 1 to 4 is retrieved to produce the next IMF. In the case of not satisfying the condition, r is considered as the residual and the algorithm is stopped.

The described steps result in a signal is the sum of produced IMFs and the residual. Fig. 1 illustrates the flowchart of the EMD algorithm, the steps 1 to 8, adapted from Premanode et al. (2013) and Qin et al. (2015).

The EMD method may cause different IMFs with the same scale residing components or IMFs with widely disparate scale components, which is known as mode mixing problem (Lee and Ouarda, 2010; Wang et al., 2015b). To overcome the problem, Wu and Huang (2009) improved EMD, called ensemble EMD (EEMD). The EEMD method adds finite white noises into the targeted time series and into IMFs of each repetition of steps 1 to 3 (Wang et al., 2015a). Wu and Huang (2009) explained the necessity of prescribing the number of ensembles and the noise amplitude as the parameters of EEMD. They recommended controlling the added noise amplitude effect with the following equations (Wu and Huang, 2009).

$$\varepsilon_n = \frac{\varepsilon}{\sqrt{N}} \tag{1a}$$

Or

$$ln\varepsilon_n + \frac{\varepsilon}{2} lnN = 0 \tag{1b}$$

where ε is the added noise amplitude, N is the number of ensemble members, and ε_n is the standard deviation of the error. It is worthy of note that the added noise may not change the extrema since their amplitudes are too small (Wu and Huang, 2009).

In the present paper, the number of ensemble members is set to 100. The amplitude is set to 0.1, 0.2, and 0.3 with the purpose of finding the best one for each time series. Also, the EMD method is investigated by considering the amplitude and ensemble number equal to zero and one, respectively.

2.4 Evaluation Criteria

To evaluate the performance of different models, the following criteria are used in the present paper (Chebana et al., 2014):

a- The determination coefficient or R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(x_{i_{obs}} - x_{i_{sim}}\right)^2}{\sum_{i=1}^{n}\left(x_{i_{obs}} - \bar{x}\right)^2} \quad (2)$$

b- The root mean square error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_{i_{obs}} - x_{i_{sim}})^2}{n}} \quad (3)$$

c- The relative root mean square error:

$$rRMSE = 100\sqrt{\frac{\sum_{i=1}^{n}(\frac{x_{i_{obs}} - x_{i_{sim}}}{x_{i_{obs}}})^2}{n}} \quad (4)$$

d- The mean bias:

$$BIAS = \frac{\sum_{i=1}^{n}(x_{i_{obs}} - x_{i_{sim}})}{n} \quad (5)$$

e- The relative mean bias

$$BIAS = 100 * (\frac{\sum_{i=1}^{n}\left(\frac{x_{i_{obs}} - x_{i_{sim}}}{x_{i_{obs}}}\right)}{n}) \quad (6)$$

where $x_{i_{obs}}$ and $x_{i_{sim}}$ are the observed and simulated values, $\bar{x}$ is the average of observed values, n is the number of observations, and i is 1, 2, …, n.

3. Data used

Monthly groundwater levels (m) of three observation wells as well as monthly precipitation (mm) and temperature (˚C) values of one meteorological station of the Delfan plain, located in Iran are considered as the inputs of the hybrid models to simulate groundwater level one month ahead. The plain is in the Zagros mountain chain between 47°, 21' and 48°, 21' E to 33°, 48' and 34°, 22' N and is 1700 to 2100 m above sea level. Table 1 shows the geographical coordinates of the stations. The observed values are divided into two parts. The first part, about 75% of observations, is employed for calibration. The second part, about 25% of observations, is used for validation.

4. Results and Discussion

4.1 Results of hybrid EEMD-GEP

The applied evaluation criteria for EEMD-GEP, in order to calibrate and validate the model, are presented in Table 2. In Table 2, ε=0 stands for the EMD method and ε= 0.1, 0.2, and 0.3 refer to the amplitude of the added noises in the EEMD method. According to Table 2, for Well 1, the EEMD-GEP model indicates the best accuracy for ε=0.3. In addition, the hybrid model has a good accuracy with ε=0 and even its R^2 is better than the R^2 of the hybrid models of ε= 0.1 and 0.2. For Well 2, for ε= 0.2 and 0.3, the performance of the hybrid EEMD-GEP is almost the same. Since ε equals to 0.2 has smaller RMSE, rRMSE, BIAS, and rBIAS in comparison with ε= 0.3, it is designated as the best ensemble for the hybrid model. For Well 3, the best performance is for ε= 0.2 because it has good performance in both of the validation and calibration. For well 3, likewise well 1, the performance of the hybrid model with ε= 0 is better than ε= 0.1 and 0.3 for validation.

4.2 Results of hybrid EEMD-M5

The evaluation criteria for EEMD-M5 are presented in Table 3. Form Table 3, it can be concluded that EEMD-M5 is incapable of simulation. For Well 1, although EEMD-M5 with ε=0.3 has the highest potentiality of calibration, regarding validation, its R^2 equals to zero. Thus, ε=0.3 does not yield a model that can be used for simulation. Regarding validation, it can be seen that the utilized parameter of ε=0 provided better performance of the hybrid model in comparison with ε=0.1 and 0.3. For Well 2, EEMD-M5 showed high accuracy in calibration. On the other hand, in the case of validation, the performance of the hybrid model, maybe due to overfitting, is too weak and EEMD-M5 fails for groundwater level simulation. The same conclusion holds true for Well 3.

4.3 Comparison of different models

In the following section, initially, the performance of EEMD-GEP and EEMD-M5 are compared with the observed values. Then, the results of the present paper are compared with the study of Bahmani et al. (Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform, submitted to Journal of Hydrology, 2018) which used the same variables to simulate groundwater levels. That study applied GEP and M5 for the simulation, but instead of EEMD, wavelet transform (WT) was used to produce hybrid models called Wavelet Gene Expression Programming (WGEP) and Wavelet M5 Model tree (WM5).

To clarify the comparison of simulated value by EEMD-GEP and EEMD-M5 with the observed values, Fig. 2 is presented. Based on Fig. 2, for well 1, hybrid EEMD-GEP values are close to observed values, while there is a big difference between EEMD-M5 and the observed values. In the case of Well 2 and 3, likewise Well 1, EEMD-GEP indicates close simulated and observed values, but EEMD-M5 is not able to simulate the groundwater levels properly.

The M5 model uses Greedy search to divide observed values into perfectly split subsets (Blum and Langley 1997). The Greedy search is an algorithm that hopes to find a global optimum by utilizing a heuristic analysis from locally optimal choices (Cormen et al. 2009). If there is a big difference between the calibration and validation values, M5 is not able to

simulate the values in validation, while GEP randomly produces chromosomes to find the best solution. Hence, by running the GEP model several times, recognizing well-simulated values is possible.

The comparison of EEMD-GEP and EEMD-M5 with the study of Bahmani et al. (2018)(Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform, submitted to Journal of Hydrology) is done through R^2s and RMSEs and is presented in Table 4. For Well 1, the hybrid models have better performance than the simple models. The simulation capability of EEMD-GEP and WM5 are higher than the other models, for calibration and validation, respectively. Regarding Well 2, WGEP shows good results for both calibration and validation, hence it can be chosen as the best model for the simulation. For Well 3, WGEP and WM5 have close performances and perform better than EEMD-GEP and EEMD-M5.

From the comparison of the present paper and the study of Bahmani et al. (Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform, submitted to Journal of Hydrology, 2018), it can be understood that the hybrid models generated by WT have better performance than the hybrid models generated by EEMD method. The comparison also confirms the conclusion from the study of Karthikeyan and Nagesh Kumar (2013) which found that the wavelet-based method is better than the EMD based method for simulations in the field of hydrology.

5. Conclusions

The results of integrated GEP and M5 with EEMD to simulate groundwater levels are presented in the current paper. Moreover, the performance of the best models is compared with the study of Bahmani et al. (Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform, submitted to Journal of Hydrology, 2018) to find the best model for groundwater simulation.

The outcomes of the presented paper demonstrate that the performance of EEMD-GEP is highly better than the EEMD-M5. The weak performance of EEMD-M5 may be caused by overfitting

Furthermore, the results of EEMD-GEP and EEMD-M5 present the importance of ε in the EEMD method and its impact on the accuracy of the hybrid models. No specified rule for ε determination has been presented and selecting ε depends on the features of a time series such as mean and extrema values. Therefore, it is recommended to find an optimal value for ε, regarding the hydrological time series.

As the last point, to pre-process a time series and simulate groundwater level, WT and GEP, respectively, are recommended. Hence, utilizing WGEP for groundwater simulation is suggested.

For the future studies, it is highly recommended to investigate the effect of GEP parameters on the simulation to reach the better performance. In addition, to improve the EEMD methods, new techniques such as Complementary Ensemble Mode Decomposition and Partly Ensemble Mode Decomposition have been developed. It is suggested to investigate the ability of the new techniques to produce hybrid models for groundwater level simulation.

6. Acknowledgements

The presents work was partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

7. References

[1] Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. Artificial Intelligence. 97:245-271. doi:https://doi.org/10.1016/S0004-3702(97)00063-5

[2] Chebana F, Charron C, Ouarda TBMJ, Martel B (2014) Regional Frequency Analysis at Ungauged Sites with the Generalized Additive Model. Journal of Hydrometeorology. 15:2418-2428 doi:10.1175/jhm-d-14-0060.1

[3] Cormen TH, Leiserson CE, Rivest RL, Stein C (2009) Introduction to Algorithms. Third Edition edn. The MIT Press, Cambridge, Massachusetts,

[4] Daliakopoulos IN, Coulibaly P, Tsanis IK (2005) Groundwater level forecasting using artificial neural networks. Journal of Hydrology. 309:229-240. doi:https://doi.org/10.1016/j.jhydrol.2004.12.001

[5] Ferreira C (2001) Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. vol 13. Complex Systems Publications, Angra do Heroísmo, Portugal

[6] Ferreira C (2002) Gene Expression Programming in Problem Solving. In: Roy R, Köppen M, Ovaska S, Furuhashi T, Hoffmann F (eds) Soft Computing and Industry: Recent Applications. Springer London, London, pp 635-653. doi:10.1007/978-1-4471-0123-9_54

[7] Ferreira C (2006) Gene Expression Programming Mathematical Modeling by an Artificial Intelligence. Studies in Computational Intelligence, vol 21. Springer-Verlag Berlin Heidelberg, Germany. doi:10.1007/3-540-32849-1

[8] Huang NE et al. (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis Proceedings of the Royal Society of London Series A: Mathematical, Physical and Engineering Sciences 454:903-995 doi:10.1098/rspa.1998.0193

[9] Huang S, Chang J, Huang Q, Chen Y (2014) Monthly streamflow prediction using modified EMD-based support vector machine. Journal of Hydrology. 511:764-775. doi:https://doi.org/10.1016/j.jhydrol.2014.01.062

[10] Jiao G, Guo T, Ding Y (2016) A New Hybrid Forecasting Approach Applied to Hydrological Data: A Case Study on Precipitation in Northwestern China. Water. 8:367

[11] Karthikeyan L, Nagesh Kumar D (2013) Predictability of nonstationary time series using wavelet and EMD based ARMA models. Journal of Hydrology. 502:103-119. doi:https://doi.org/10.1016/j.jhydrol.2013.08.030

[12] Kisi O (2015) Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. Journal of Hydrology. 528:312-320. doi:https://doi.org/10.1016/j.jhydrol.2015.06.052

[13] Kisi O, Shiri J (2011) Precipitation Forecasting Using Wavelet-Genetic Programming and Wavelet-Neuro-Fuzzy Conjunction Models. Water Resources Management. 25:3135-3152. doi:10.1007/s11269-011-9849-3

[14] Lee T, Ouarda TBMJ (2010) Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. Journal of Geophysical Research: Atmospheres. 115:D13107. doi:10.1029/2009JD012801

[15] Lee T, Ouarda TBMJ (2011) Prediction of climate nonstationary oscillation processes with empirical mode decomposition. Journal of Geophysical Research: Atmospheres. 116:D06107. doi:10.1029/2010JD015142

[16] Lee T, Ouarda TBMJ (2012) Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition. Water Resources Research. 48:W02514. doi:10.1029/2011WR010660

[17] Masselot P, Chebana F, Bélanger D, St-Hilaire A, Abdous B, Gosselin P, Ouarda TBMJ (2018) EMD-regression for modelling multi-scale relationships, and application to weather-related cardiovascular mortality. Science of The Total Environment. 612:1018-1029. doi:https://doi.org/10.1016/j.scitotenv.2017.08.276

[18] Nourani V, Komasi M, Mano A (2009) A Multivariate ANN-Wavelet Approach for Rainfall–Runoff Modeling. Water Resources Management. 23:2877. doi:10.1007/s11269-009-9414-5

[19] Nourani V, Mousavi S (2016) Spatiotemporal groundwater level modeling using hybrid artificial intelligence-meshless method. Journal of Hydrology. 536:10-25. doi:https://doi.org/10.1016/j.jhydrol.2016.02.030

[20] Ouali D, Chebana F, Ouarda TBMJ (2017) Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. Journal of Advances in Modeling Earth Systems. 9:1292-1306.. doi:10.1002/2016MS000830

[21] Premanode B, Vongprasert J, Toumazou C (2013) Noise Reduction for Nonlinear Nonstationary Time Series Data using Averaging Intrinsic. Mode Function Algorithms. 6:407

[22] Qin S, Wang Q, Kang J (2015) Output-Only Modal Analysis Based on Improved Empirical Mode Decomposition Method. Advances in Materials Science and Engineering. 2015:12. doi:10.1155/2015/945862

[23] Quinlan JR (1986) Induction of decision trees. Machine Learning. 1:81-106. doi:10.1007/bf00116251

[24] Quinlan JR (1992) Learning with Continuous Classes Proceedings of Australian Joint Conference on Artificial Intelligence, Hobart, Australia, World Scientific, Singapore:343-348

[25] Shiri J, Kisi O, Yoon H, Lee K-K, Hossein Nazemi A (2013) Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques. Computers & Geosciences. 56:32-44. doi:https://doi.org/10.1016/j.cageo.2013.01.007

[26] Sivapragasam C, Kannabiran K, Karthik G, Raja S (2015) Assessing Suitability of GP Modeling for Groundwater Level. Aquatic Procedia. 4:693-699. doi:https://doi.org/10.1016/j.aqpro.2015.02.089

[27] Solgi A, Pourhaghi A, Bahmani R, Zarei H (2017) Pre-processing data using wavelet transform and PCA based on support vector regression and gene expression programming for river flow simulation. Journal of Earth System Science. 126:65. doi:10.1007/s12040-017-0850-y

[28] Taormina R, Chau K-w, Sethi R (2012) Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. Engineering Applications of Artificial Intelligence. 25:1670-1676. doi:https://doi.org/10.1016/j.engappai.2012.02.009

[29] Wang W-c, Chau K-w, Qiu L, Chen Y-b (2015a) Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. Environmental Research. 139:46-54. doi:https://doi.org/10.1016/j.envres.2015.02.002

[30] Wang W-c, Chau K-w, Xu D-m, Chen X-Y (2015b) Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition. Water Resources Management. 29:2655-2675. doi:10.1007/s11269-015-0962-6

[31] Wu Z, Huang NE (2009) ENSEMBLE EMPIRICAL MODE DECOMPOSITION: A NOISE-ASSISTED DATA ANALYSIS METHOD. Advances in Adaptive Data Analysis. 01:1-41. doi:10.1142/s1793536909000047

8. Figure Captions

9. Table Captions

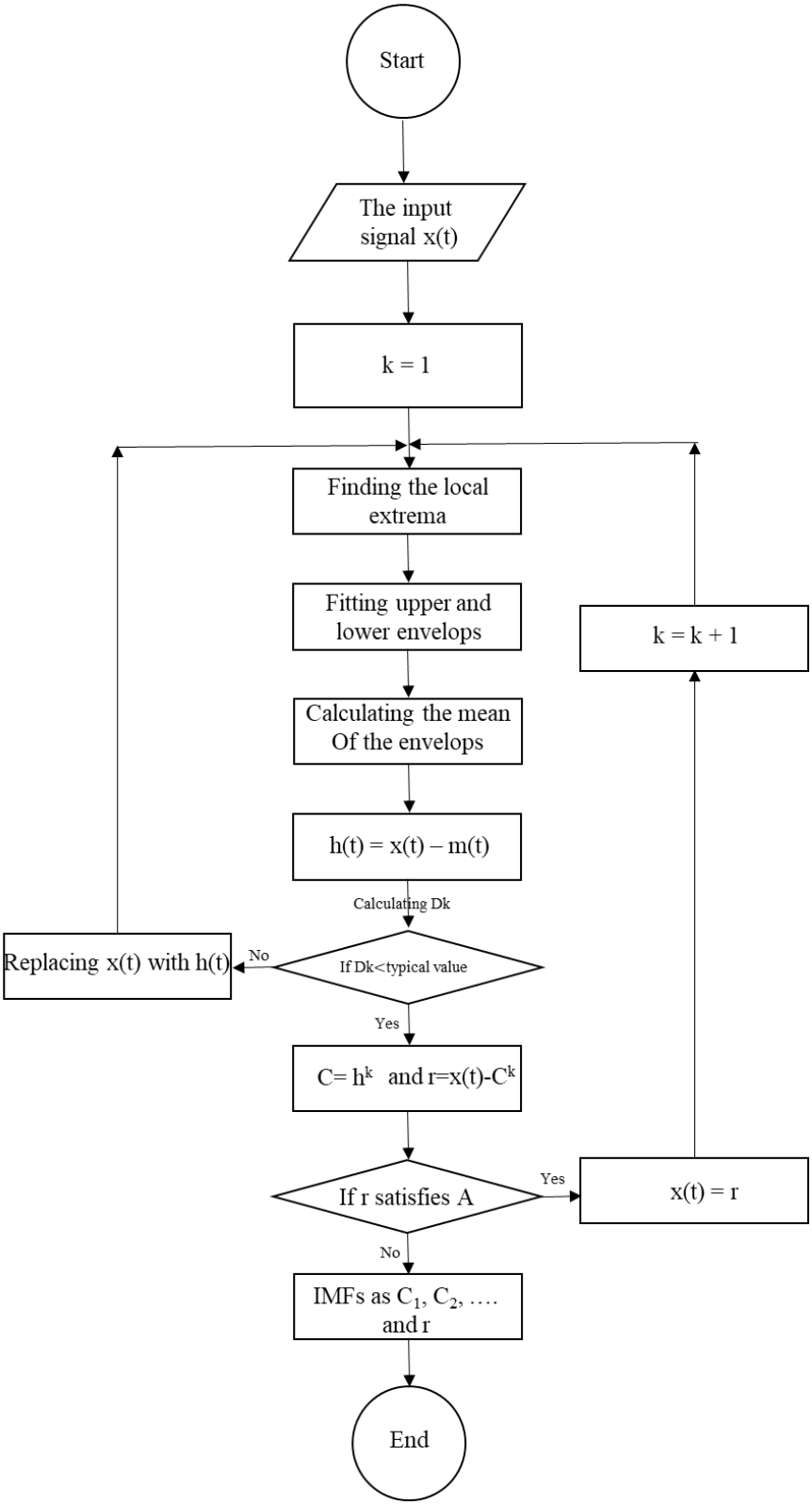


Fig. 4.1 The flowchart of the EMD algorithm

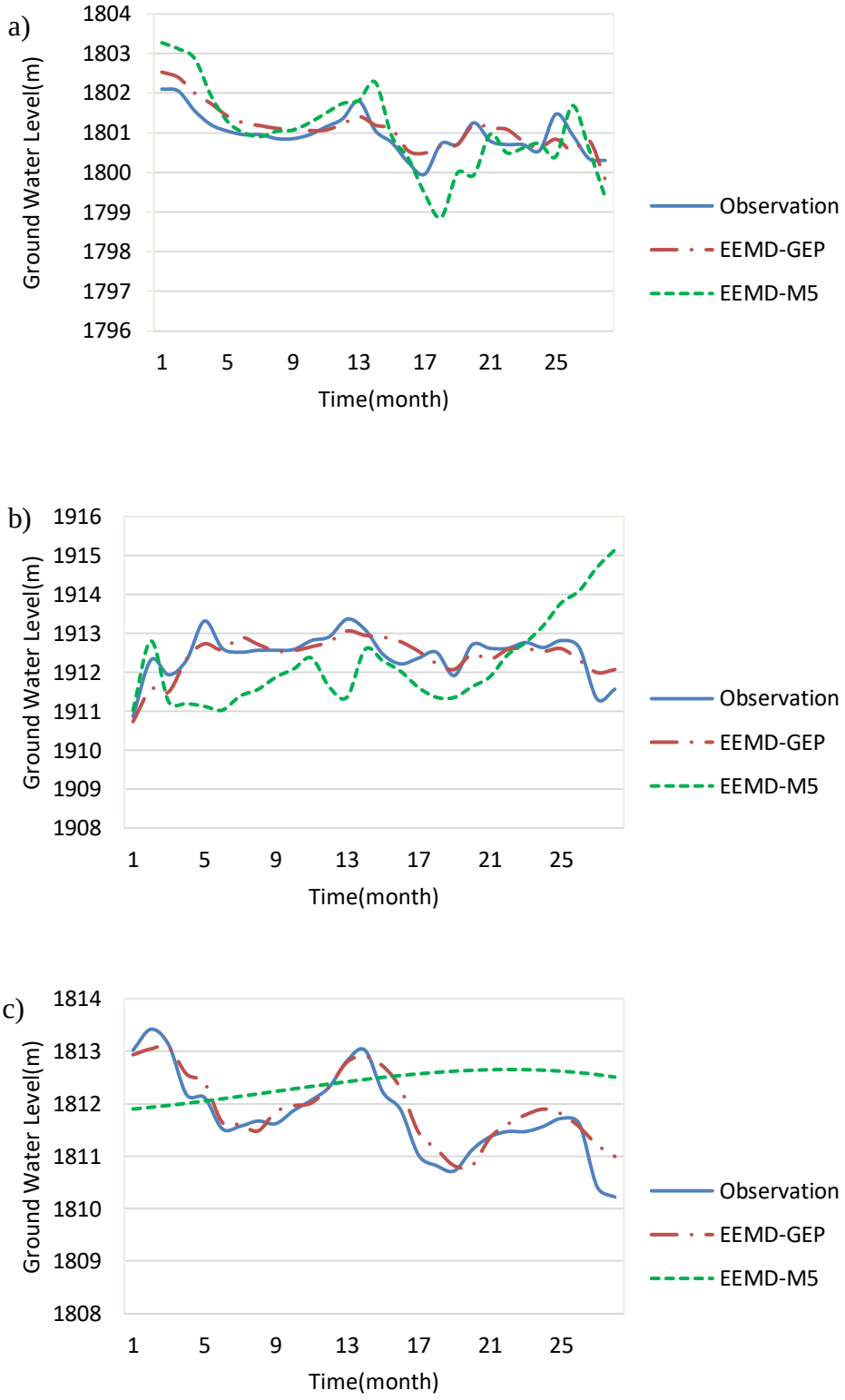


Fig. 4.2 Simulated and observed values of a) well 1, b) well 2 and c) well 3

Table 4.1. The geographical coordinates of the stations

Id	Type of station	Name of station	Longitude	Latitude	Elevation (m)
1	Observation well	Aziz Abad	48° 02' E	34° 06' N	1803
2	Observation well	Cheragh	48° 10' E	34° 03' N	1914
3	Observation well	Sikvand	47° 58' E	34° 10' N	1826
4	Meteorological	Bad Avar	47° 59' E	34° 04' N	1800

Table 4.2. Evaluation criteria of the simulation for EEMD-GEP

Criteria		Well 1				Well 2				Well 3			
		ε				ε				ε			
		0	**0.1**	**0.2**	**0.3**	**0**	**0.1**	**0.2**	**0.3**	**0**	**0.1**	**0.2**	**0.3**
R^2	**calibration**	0.73	0.61	0.60	0.79	0.9	0.91	0.93	0.93	0.89	0.88	0.91	0.91
	validation	0.68	0.58	0.60	0.69	0.62	0.60	0.63	0.64	0.88	0.78	0.88	0.85
RMSE(m)	**calibration**	0.34	0.42	0.42	0.31	0.37	0.40	0.36	0.35	0.39	0.41	0.35	0.36
	validation	0.35	0.33	0.34	0.34	0.97	0.35	0.33	0.40	0.54	0.40	0.32	0.33
rRMSE	**calibration**	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	validation	0.02	0.02	0.02	0.02	0.05	0.02	0.02	0.02	0.03	0.02	0.02	0.02
BIAS(m)	**calibration**	-0.01	-0.02	0.01	0.00	-0.01	0.02	0.02	0.02	0.04	-0.02	-0.05	-0.03
	validation	-0.08	-0.05	-0.09	-0.13	-0.03	-0.02	0.05	0.13	-0.19	-0.06	-0.15	-0.07
rBIAS	**calibration**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	validation	0.00	0.00	-0.01	-0.01	0.00	0.00	0.00	0.01	-0.01	0.00	-0.01	0.00

ε is the added noise amplitude

Table 4.3. Evaluation criteria of the simulation for EEMD-M5

Criteria		Well 1				Well 2				Well 3			
		ε				ε				ε			
		0	0.1	0.2	0.3	0	0.1	0.2	0.3	0	0.1	0.2	0.3
R^2	calibration	0.68	0.76	0.76	0.79	0.95	0.97	0.94	0.97	0.88	0.92	0.91	0.88
	validation	0.59	0.19	0.59	0.00	0.02	0.04	0.00	0.02	0.18	0.02	0.37	0.03
RMSE(m)	calibration	0.37	0.33	0.33	0.30	0.31	0.24	0.32	0.25	0.41	0.34	0.35	0.88
	validation	0.69	0.76	0.74	1.57	2.19	1.34	4.56	1.82	4.56	1.98	1.13	1.82
rRMSE	calibration	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.41
	validation	0.04	0.04	0.04	0.09	0.11	0.07	0.24	0.10	0.25	0.11	0.06	0.10
BIAS(m)	calibration	0.00	0.00	-0.06	0.00	-0.02	0.00	0.00	-0.03	0.00	0.01	0.01	0.02
	validation	-0.60	-0.54	-0.08	-1.00	1.69	0.26	-3.67	1.59	-3.98	-1.64	-0.60	-1.17
rBIAS	calibration	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	validation	-0.03	-0.03	0.00	-0.06	0.09	0.01	-0.19	0.08	-0.22	-0.09	-0.03	-0.06

ε is the added noise amplitude

Table 4.4. Comparison of different models

Observation Wells	Model Type	R^2		RMSE(m)	
		Calibration	Validation	Calibration	Validation
1	*GEP	0.64	0.63	0.41	0.39
	*M5	0.59	0.62	0.43	0.39
	*WGEP	0.71	0.70	0.36	0.32
	*WM5	0.71	0.77	0.36	0.26
	EEMD-GEP	0.79	0.69	0.31	0.34
	EEMD-M5	0.76	0.59	0.33	0.74
2	*GEP	0.89	0.40	0.44	0.51
	*M5	0.90	0.41	0.37	0.43
	*WGEP	0.94	0.76	0.34	0.30
	*WM5	0.95	0.73	0.31	0.31
	EEMD-GEP	0.93	0.63	0.36	0.33
	EEMD-M5	0.97	0.04	0.24	1.34
3	*GEP	0.81	0.81	0.53	0.38
	*M5	0.80	0.79	0.54	0.49
	*WGEP	0.93	0.90	0.31	0.31
	*WM5	0.92	0.91	0.34	0.26
	EEMD-GEP	0.91	0.88	0.35	0.32
	EEMD-M5	0.91	0.37	0.35	1.13

* Bahmani et al. (Groundwater level simulation using Gene Expression Programming and M5 Model tree combined with wavelet transform, submitted to Journal of Hydrology, 2018).

CHAPITRE IV: A DISCUSSION ON “THE INCORRECT USAGE OF SINGULAR SPECTRAL ANALYSIS AND DISCRETE WAVELET TRANSFORM IN HYBRID MODELS TO PREDICT HYDROLOGICAL TIME SERIES” BY DU ET AL. (2017)

A Discussion on "The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series" by Du et al. (2017)

Ramin Bahmani[1*], Taha B. M. J. Ouarda[1], Abazar Solgi[2]

[1] Canada Research Chair in Statistical Hydro-climatology, INRS-ETE, Québec (Québec), Canada.

[2] Department of Water Resources Engineering, Faculty of Water Sciences Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

* Corresponding author: ramin.bahmani@ete.inrs.ca

Journal : submitted to the Journal of Hydrology

Abstract

Du et al. (2017) addressed the incorrect usage of singular spectrum analysis (SSA) and discrete wavelet transform (DWT) to preprocess data to be used in artificial neural networks (ANN) and support vector machine (SVM) models. In the present discussion, the same data is used to verify the validity of the authors claims. To do this, ANN and DWT-ANN are considered based on the experiment design of the original paper. The only difference between the analysis carried out in the present discussion and in the original article is the use of time lags that are applied to the inputs of the models. The results show that if the time lags 1 to 7 are used as inputs to the models, using the whole time series to be transformed by the DWT does not lead to incorrect results.

Keywords: Artificial Neural Network, Hybrid Model, Wavelet transform, Signal spectrum analysis, Time series, Estimation error.

1. Discussion and Results

Artificial Neural Networks (ANN) represent a set of models that are commonly used for modeling non-linear time series in environmental and water sciences. The ANN model can be easily adapted to various types of data formats with no assumption about the relation between the independent and dependent variables (Chokmani et al., 2008). However, the model may not be so successful for modeling time series with high non-stationary fluctuations (Nourani et al., 2009a). Therefore, wavelet transform (WT) is recommended to decompose the main time series to sub-signals. The sub-signals are used as inputs to the ANN to improve its accuracy (Nourani et al., 2009a; Solgi et al., 2017a).

WT was introduced by Grossmann and Morlet (1984) and was adopted by several researchers in a large number of fields (Oh et al., 2017; Ouachani et al., 2013; Partal and Kişi, 2007; Shoaib et al., 2015). WT represents a convenient combination of different theories in mathematics, engineering, and physical sciences and is mathematically easy to use (Young, 1993). Mallat (1999) defined the WT of a continuous time series as:

$$T(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} g^* \left(\frac{t-b}{a}\right) x(t).dt \quad (1)$$

where $g^*(t)$ is the wavelet function or mother wavelet with the complex conjugate, a is a dilation factor, b is a temporal translation, and $x(t)$ is the continuous time signal.

In practical applications, there is no continuous time signal in hydrology and data is recorded as discrete time signals (Nourani et al., 2014). Therefore, the discretization of equation (1) can be carried out using a logarithmically uniform spacing, leading to what is called “discrete wavelet transform” (DWT). The discretization process allows hydrologists to decompose a time signal to sub-signals (Nourani et al., 2014).

Du et al. (2017) reported that the use of DWT may not be correct in hydrology because sub-signals include future information and concluded that the ANN model combined with DWT leads to incorrect high prediction performances. The authors designed an experiment to verify their claims. They supposed they have a monthly rainfall time series

of length N and argued that if they use all the data (N) for decomposing, the results should be similar to the case where only the first M (M<N) observations of the time series are considered. Then, the authors used db7 as a mother wavelet for decomposing data and generating sub-signals. Next, the sub-signals were used as inputs to the ANN and the model was called DWT-ANN.

The Du et al. (2017) assumption ignores the fact that the level of decomposition is related to the length of the time series. Also, Du et al. (2017) did not explain how it is possible to use the same level of decomposition for sets of data with different lengths. Nourani et al. (2009b) suggested a formula for the identification of the level of decomposition as a function of the time series length.

The objective of the present work is to discuss the results of the original article by Du et al. (2017) for modeling rainfall time series using ANN and DWT-ANN. The only difference between the analysis carried out in the present discussion and in the original article is the use of time lags that are applied to the inputs of ANN. The time lag means the use of the values of rainfall (mm) corresponding to previous months. For example, time lag 1 refers to the rainfall value corresponding to the previous month, time lag 2 refers to the rainfall value corresponding to two months ago, and time lag n refers to the rainfall value corresponding to n months ago. The advantage of using time lags is to improve the model performance since there is some useful information for modeling in previous months (Kisi and Shiri, 2011; Shiri and Kişi, 2011; Shiri et al., 2013; Solgi et al., 2017b).

To prepare inputs for the DWT-ANN, similarly to the original paper, the whole monthly rainfall time series (mm) is considered for decomposition using the same mother wavelet, db7. Then, the generated sub-signals are divided to four data sets and modeled to check if the conclusions of the original paper are valid. For modeling, similarly to the original paper, the first data set (1 to 750) is used for training, the second data set (751 to 900) for validation, the third data set (901 to 1000) for the inner test, and the fourth data set (1001 to 1100) for the outer test.

The Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Nash-Sutcliffe coefficient (NS) are used to compare the results of the original paper and the present discussion. They are defined as (Du et al., 2017):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_{i_{obs}} - x_{i_{est}})^2}{N}} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{N}|x_{i_{est}} - x_{i_{obs}}|}{N} \quad (3)$$

$$NS = 1 - \frac{\sum_{i=1}^{N}(x_{i_{est}} - x_{i_{obs}})^2}{\sum_{i=1}^{N}(x_{i_{obs}} - \bar{x}_{obs})^2} \quad (4)$$

Where N is the length of the data set, $x_{i_{obs}}$ is the monthly-observed rainfall, $x_{i_{est}}$ is the monthly-simulated rainfall, and $\bar{x}_{obs}$ is the average of monthly-observed rainfalls.

We recall from Table 2 of the original paper that, for the simple ANN, the inner and outer tests led to similar values of the RMSE, MAE, and NS, whereas for DWT-ANN a large difference of values was observed. The RMSE of DWT-ANN increased from 42.3 mm, for the inner test, to 664.3 mm, for the outer test. In other words, the RMSE of the outer test was 15.7 times larger than for the inner test. Similarly, for MAE, the outer test value was 15.4 times larger than for the inner test.

The simulation results associated to the present discussion are presented in Table 1. They show that, when time lags of 1 to 7 are used as inputs to the ANN and DWT-ANN, the models lead to the highest performances in terms of RMSE, MAE, and NS values. The comparison of the inner and outer test results associated to the present discussion show no large differences between RMSE, MAE, and NS values. The outer test RMSE is 1.3 times larger than the inner test, and the outer test MAE is 1.2 times larger than the inner test. RMSE, MAE, and NS values of the outer test for the present discussion are not as good as the inner test. However, this can be the effect of data characteristics, such as high standard deviation, skewness, and mean (Du et al., 2017).

In the original paper, the authors explained that, because of the overfitting of the inner test, the RMSE and MAE of the DWT-ANN were low and the RMSE and MAE of the outer test were considerably high. However, for the present discussion results, no big differences are observed between the RMSE and MAE of DWT-ANN. Therefore, it cannot be concluded that the overfitting is the result of using DWT.

Simulated monthly rainfall values (mm) for the inner and outer tests are presented in Fig. 1 and 2 respectively. These figures indicate that the simulated values by ANN and DWT-ANN are close for both the inner and outer tests. The only relatively small difference between the simulated and observed values is noted for the peaks.

Fig. 3 presents the scatter plot of observed and simulated values. Fig. 3 shows a high correlation between the simulated and observed values obtained by the present simulations. In the original paper, for DWT-ANN, R^2 values decreased from 0.998 for the inner test to 0.578 for the outer test. On the other hand, for the present discussion, R^2 values do not show a large difference between the inner and outer tests. Finally, the R^2 corresponding to the outer test improved from 0.578 in the original paper to 0.916 for the present discussion.

2. Conclusions

In the present discussion, the claim of Du et al. (2017) could not be verified for DWT-ANN. The use of time lag 1 to 7 showed that the DWT-ANN does not lead to incorrect simulation if the inputs of the model are selected correctly and the model is trained properly. In addition, it was not proven that the overfitting was the results of using DWT.

In the present discussion, the effect of using time lags in inputs was evaluated. Another important factor, which may affect modeling with DWT, is the selection of a mother wavelet. The effect of the mother wavelet selection on modeling results should be evaluated in future research efforts.

3. References

[1] Chokmani K, Ouarda TBMJ, Hamilton S, Ghedira MH, Gingras H (2008) Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques. Journal of Hydrology. 349:383-396 doi:https://doi.org/10.1016/j.jhydrol.2007.11.024

[2] Du K, Zhao Y, Lei J (2017) The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. Journal of Hydrology. 552:44-51. doi:http://dx.doi.org/10.1016/j.jhydrol.2017.06.019

[3] Grossmann A, Morlet J (1984) Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape SIAM. Journal on Mathematical Analysis. 15:723-736. doi:10.1137/0515056

[4] Kisi O, Shiri J (2011) Precipitation Forecasting Using Wavelet-Genetic Programming and Wavelet-Neuro-Fuzzy Conjunction Models. Water Resources Management. 25:3135-3152. doi:10.1007/s11269-011-9849-3

[5] Mallat S (1999) A Wavelet Tour of Signal Processing (Second Edition). In. San Diego. eBook ISBN: 9780080520834

[6] Nourani V, Alami MT, Aminfar MH (2009a) A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation. Engineering Applications of Artificial Intelligence. 22:466-472. doi:https://doi.org/10.1016/j.engappai.2008.09.003

[7] Nourani V, Hosseini Baghanam A, Adamowski J, Kisi O (2014) Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. Journal of Hydrology. 514:358-377. doi:https://doi.org/10.1016/j.jhydrol.2014.03.057

[8] Nourani V, Komasi M, Mano A (2009b) A Multivariate ANN-Wavelet Approach for Rainfall–Runoff Modeling. Water Resources Management. 23:2877 doi:10.1007/s11269-009-9414-5

[9] Oh Y-Y, Yun S-T, Yu S, Hamm S-Y (2017) The combined use of dynamic factor analysis and wavelet analysis to evaluate latent factors controlling complex groundwater level fluctuations in a riverside alluvial aquifer. Journal of Hydrology. 555:938-955. doi:https://doi.org/10.1016/j.jhydrol.2017.10.070

[10] Ouachani R, Zoubeida B, Taha O (2013) Power of teleconnection patterns on precipitation and streamflow variability of upper Medjerda Basin International. Journal of Climatology. 33:58-76. doi:doi:10.1002/joc.3407

[11] Partal T, Kişi Ö (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting Journal of Hydrology 342:199-212 doi:https://doi.org/10.1016/j.jhydrol.2007.05.026

[12] Shiri J, Kişi Ö (2011) Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Computers & Geosciences. 37:1692-1701. doi:http://dx.doi.org/10.1016/j.cageo.2010.11.010

[13] Shiri J, Kisi O, Yoon H, Lee K-K, Hossein Nazemi A (2013) Predicting groundwater level fluctuations with meteorological effect implications—A comparative study among soft computing techniques. Computers & Geosciences. 56:32-44 doi:https://doi.org/10.1016/j.cageo.2013.01.007

[14] Shoaib M, Shamseldin AY, Melville BW, Khan MM (2015) Runoff forecasting using hybrid Wavelet Gene Expression Programming (WGEP) approach. Journal of Hydrology. 527:326-344. doi:http://dx.doi.org/10.1016/j.jhydrol.2015.04.072

[15] Solgi A, Nourani V, Bagherian Marzouni M (2017a) Evaluation of nonlinear models for precipitation forecasting. Hydrological Sciences Journal. 62:2695-2704. doi:10.1080/02626667.2017.1392529

[16] Solgi A, Pourhaghi A, Bahmani R, Zarei H (2017b) Pre-processing data using wavelet transform and PCA based on support vector regression and gene expression programming for river flow simulation. Journal of Earth System Science. 126:65. doi:10.1007/s12040-017-0850-y

[17] Young RK (1993) Wavelet Theory and Its Applications. The Springer International Series in Engineering and Computer Science. vol 189. Springer US. doi:10.1007/978-1-4615-3584-3

Table 2.1. RMSE, MAE, and NS of the original article and the present discussion simulations

	Original paper (Du et al., 2017)			Present discussion (use of the time lag 1 to 7)		
	RMSE (mm)	**MAE (mm)**	**NS**	**RMSE (mm)**	**MAE (mm)**	**NS**
ANN, inner test	237.1	182.2	0.945	230.4	173.9	0.944
ANN, outer test	279.0	187.2	0.926	276.4	190.1	0.927
DWT-ANN, inner test	42.3	32.5	0.998	236.3	173.7	0.941
DWT-ANN, outer test	664.3	500	0.578	303.9	203.7	0.912

4. Figure captions

Fig. 2.1 Simulated and observed rainfall for inner test

Fig. 2.2 Simulated and observed rainfall for the outer test

Fig. 2.3 Scatter plots of simulated and observed rainfall

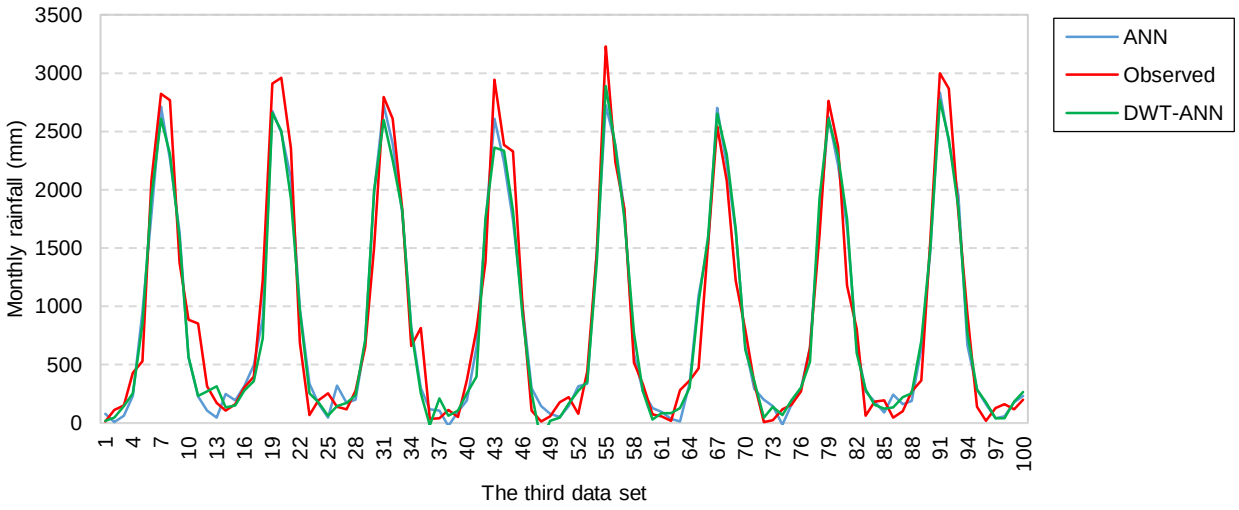


Fig. 2.1 Simulated and observed rainfall for inner test.

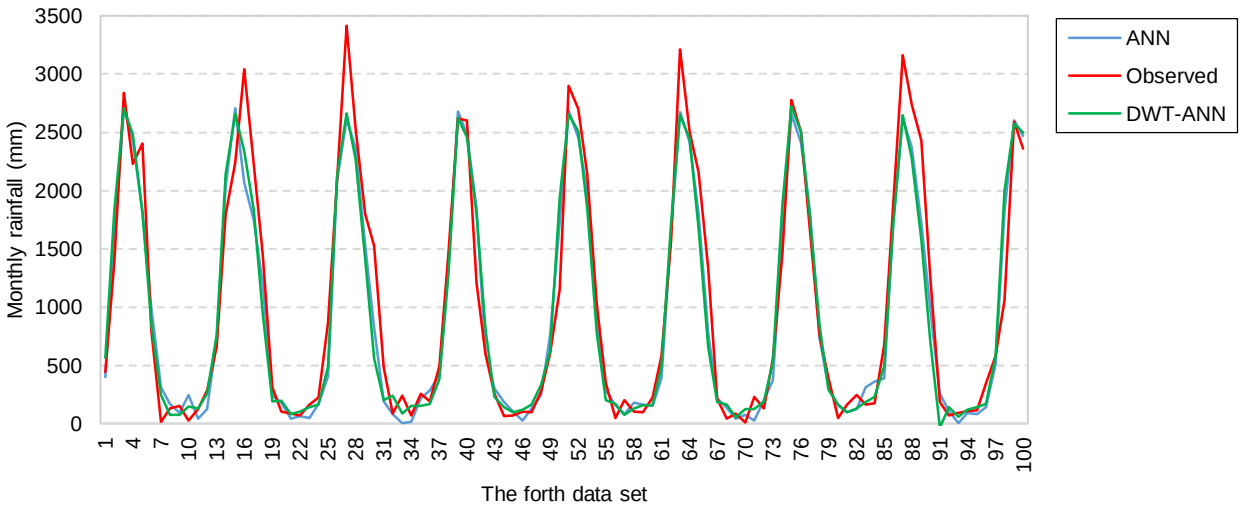


Fig. 2.2 Simulated and observed rainfall for the outer test.

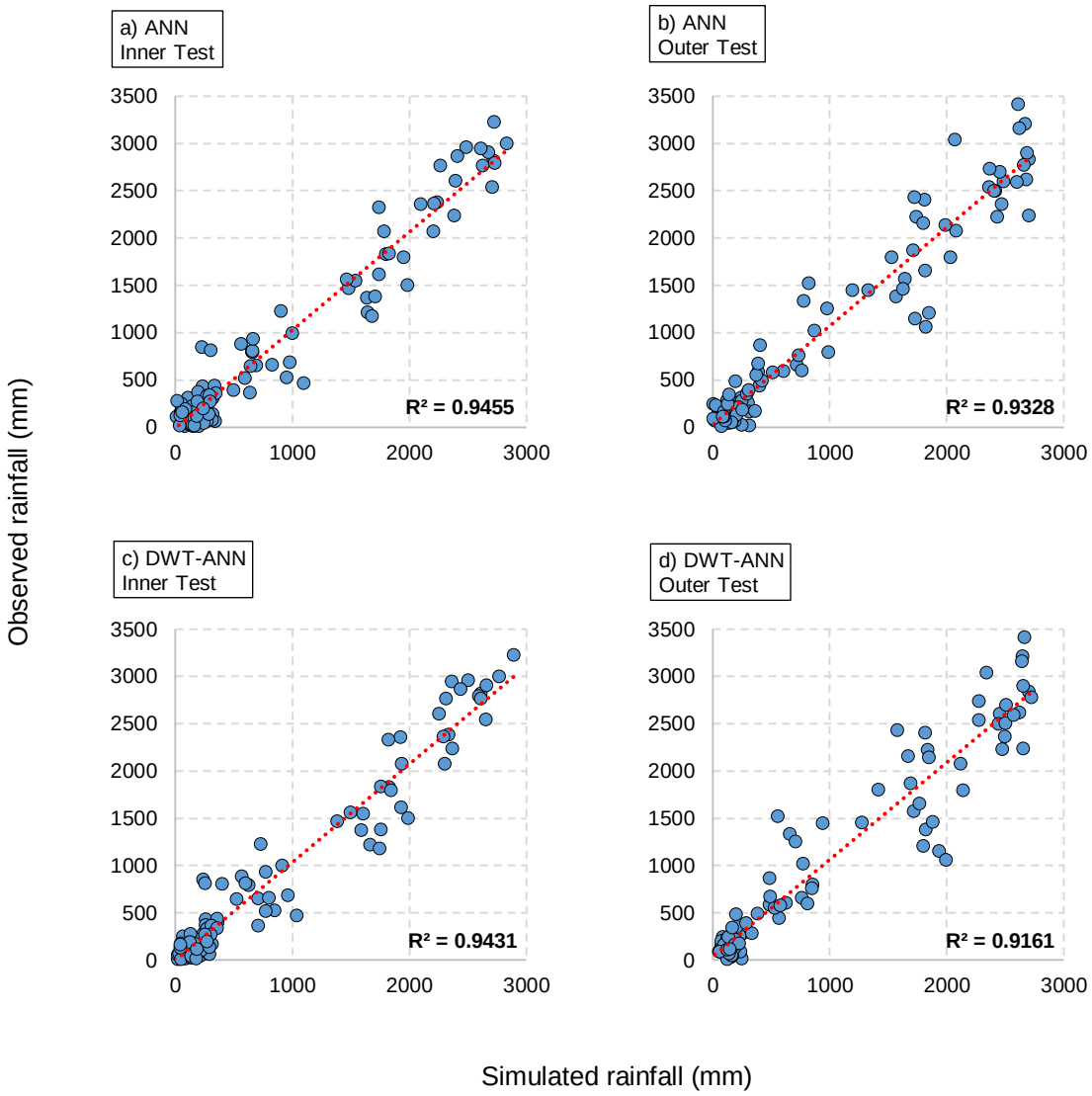


Fig. 3 Scatter plots of simulated and observed rainfall.