Centre Eau Terre Environnement

MODÉLISATION DE LA TEMPÉRATURE DE L'EAU EN RIVIÈRE : RÉGRESSION PAR DÉCOMPOSITION MODALE EMPIRIQUE ET COMPARAISON AVEC D'AUTRES APPROCHES

Par

Ramzi Abaza

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

Jury d'évaluation

Président du jury et examinateur interne	Salaheddine El Adlouni Professeur Associé, INRS-ETE Université de Moncton
Examinateur externe	Ousmane Seidou Professeur, Université d'Ottawa
Directeur de recherche	Fateh Chebana Professeur, INRS-ETE
Codirecteur de recherche	André St-Hilaire Professeur, INRS-ETE
Codirecteur de recherche	Pierre Masselot London School of Hygiene & Tropical Medicine

REMERCIEMENTS

Je tiens à remercier, dans un premier temps, mon directeur de recherche Professeur Fateh Chebana pour l'aide très compétente, pour sa confiance, sa patience et ses conseils judicieux tout au long de ma maîtrise. Merci d'avoir cru en moi.

Je remercie également mon co-directeur, Professeur André St-Hilaire qui a apporté une valeur ajoutée considérable à ce travail de recherche par ses révisions et ses commentaires avisés. Merci, André, de ta disponibilité et de ton soutien matériel et moral. Mes remerciements vont aussi à mon co-directeur Pierre Masselot pour sa précieuse aide et ses commentaires pertinents. Merci pour tout.

Je tiens à remercier aussi la coordinatrice de RivTemp Claudine Boyer de m'avoir facilité l'accès aux données à l'étude.

Je tiens à exprimer ma gratitude aux membres du jury pour avoir accepté de juger mon travail de recherche. Merci aux professeurs Salaheddine El Adlouni et Ousmane Seidou.

Je voudrais exprimer ma reconnaissance chaleureuse à ma chère femme Rabiaa Ben Aicha qui est aussi ma collègue à l'INRS. Merci pour ta patience, pour tes révisions inestimables et pour ton soutien inconditionnel. Merci à mes p'tits cœurs Baraa, Jana et Layan qui savent bien me dessiner des grands sourires dans mes moments difficiles :)

Je remercie toutes les personnes formidables que j'ai rencontré à l'INRS ou ailleurs pour leur soutien et leurs encouragements.

RÉSUMÉ

La température de l'eau a une influence importante sur l'écosystème aquatique, notamment sur la qualité de l'eau ainsi que sur le métabolisme et la distribution des espèces aquatiques. Il est donc essentiel de développer des outils fiables pour prédire la température de l'eau. L'objectif, dans ce travail de recherche, est d'introduire la régression de décomposition modale empirique (EMD-R) pour la prévision de la température quotidienne de l'eau en utilisant la température de l'air comme prédicteur. L'EMD-R est ainsi comparée à deux modèles statistiques classiques : le modèle additif généralisé (GAM) et la régression sigmoïde. Cette comparaison est effectuée sur les données de deux rivières aux États-Unis et de deux rivières au Canada. Ces trois modèles sont évalués à l'aide de quatre critères de performance, à savoir l'erreur quadratique moyenne (RMSE), le coefficient de détermination (R^2), la validation croisée généralisée (GCV) et le biais. Pour les quatre cas étudiés, le modèle EMD-R est généralement celui qui fournit les performances les plus élevées par rapport aux autres modèles considérés.

Mots-clés : Température de l’eau, température de l’air, décomposition en mode empirique, régression, LASSO, prédiction

ABSTRACT

Water temperature has a significant influence on the aquatic ecosystem, including impacts on water quality as well as on the metabolism and distribution of aquatic species. It is therefore essential to develop reliable tools to predict water temperature. The objective in this research is to introduce empirical mode decomposition regression (EMD-R) for the prediction of daily water temperature using air temperature as a predictor. EMD-R is herein compared to two classical statistical models: Generalized Additive Model (GAM) and Sigmoid regression. This comparison is performed on data from two rivers in the United States and two rivers in Canada during the ice-free period. These three models are evaluated using four performance criteria, namely the Root Mean Square Error (RMSE), the coefficient of determination (R^2), the Generalized Cross Validation (GCV) and the Bias. For the four studied cases, the EMD-R model is generally the one providing the highest performances compared to the other considered models.

Keywords : Water temperature, Air temperature, Empirical Mode Decomposition, Regression, LASSO, Prediction

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

EMD-R	Régression par décomposition modale empirique (Empirical Mode Decomposition-Regression)
EEMD	Ensemble de décomposition modale empirique (Ensemble of Empirical Mode Decomposition)
IMF	Fonctions en mode intrinsèque (Intrinsic Mode Functions)
MAG (GAM)	Modèle additif généralisé (Generalized Additive Model)
RMSE	Racine de l'erreur quadratique moyenne (Root Mean Square Error)
B (Bias)	Biais
GCV	Validation croisée généralisée (Generalized Cross Validation)
LASSO	Least Absolute Shrinkage and Selection Operator
MSE	L'erreur quadratique Moyenne (Mean Square Error)

1 INTRODUCTION

1.1 Mise en contexte

La température de l'eau est une variable clé dans les études des écosystèmes aquatiques de rivières (Cluis, 1972, Caissie. *et al.*, 2001, Beaufort *et al.*, 2016, Zhu *et al.*, 2019). Elle est souvent utilisée pour expliquer ou même pour prédire la qualité de l'eau ainsi que la qualité de l'habitat faunique et floristique (Gu *et al.*, 2002, Bélanger *et al.*, 2005). La température de l'eau, une caractéristique importante dans les paramètres environnementaux, permet de renseigner directement ou indirectement sur les propriétés physiques, chimiques et biologiques d'un écosystème aquatique (Benyahya *et al.*, 2007a, Laanaya *et al.*, 2017, Sandersfeld *et al.*, 2017, Li *et al.*, 2018). En effet, la température de l'eau régule la concentration optimale d'oxygène dissous dans l'eau et affecte des caractéristiques telles que la densité, la tension superficielle, la viscosité, la pression de vapeur et la solubilité des gaz (Marceau *et al.*, 1986, Ficklin *et al.*, 2013).

En outre, la température de l'eau en rivière peut influer directement les populations des poissons, d'invertébrés, de mollusques et de plantes qui ne peuvent s'adapter que pour une plage de température spécifique et tolérer des variations limitées de la température de l'eau (Ahmadi-Nedushan *et al.*, 2007, Benyahya *et al.*, 2007a). Un changement de régime thermique a des impacts importants sur le type d'habitat de poissons, sa répartition dans la rivière, sa production et peut mettre en péril un nombre important de poissons (Lessard *et al.*, 2003, Mohseni *et al.*, 2003, Caissie, 2006, Zhu *et al.*, 2019).

D'autre part, le réchauffement thermique peut favoriser certaines menaces à la santé des écosystèmes telles que la prolifération des fleurs d'eau d'algues bleu-vert et le développement des microorganismes pathogènes, ce qui pourrait avoir, par conséquent, plusieurs impacts environnementaux. Au-delà de l'impact écologique, la température de l'eau possède indirectement une incidence sociale et économique, en modulant entre autre la qualité de l'eau potable (Caissie, 2006).

1.2 Les facteurs environnementaux influençant la température de l'eau

La dynamique de la température des rivières est influencée par de nombreux facteurs naturels que ce soient météorologiques, géophysiques ou morphologiques (Caissie, 2006, Guillemette *et al.*, 2009). Les facteurs météorologiques sont parmi les facteurs les plus prépondérants associés

aux échanges thermiques qui se font à travers l'interface air-eau. Différents processus physiques y sont impliqués tels que l'évaporation, l'intensité et la durée de la radiation solaire, etc (Olden *et al.*, 2010). La température de l'eau est régie par la température et le volume de l'eau de surface et souterraine, par la température de l'air, la pression dans l'air, les précipitations et la vitesse du vent à la surface (Bélanger *et al.*, 2005, Caissie *et al.*, 2005). En effet, plus l'eau est soumise à la radiation solaire et plus le débit de l'eau est faible, plus la température de l'eau augmente (Caissie *et al.*, 2005, Caissie, 2006).

Les facteurs topographiques et morphologiques (e.g. profondeur de la rivière, pente, degré de turbulence, dimensions de surfaces libres, géologie, substrat, végétation riveraine et ombrage) sont également importants vu qu'ils influencent les conditions régissant en partie les flux de chaleur à la surface, mais aussi avec le lit du cours d'eau. Par exemple, à débit comparable, une rivière large et peu profonde se réchauffe plus vite qu'une rivière étroite et profonde. Notons que les variations diurnes peuvent être de 1°C à plus de 15°C (Johnson *et al.*, 2004).

1.3 Les facteurs anthropiques influençant la température de l'eau

Les facteurs anthropiques tels que le changement climatique, la déforestation et la pollution thermique sont des plus considérables perturbateurs qui entraînent de façon plus ou moins rapide des modifications des régimes thermiques naturels dans les rivières (Johnson *et al.*, 2000, Caissie. *et al.*, 2001). Les activités humaines sont sources directes ou indirectes d'une pression grandissante sur les différents services écosystémiques (Poole *et al.*, 2001). Parmi ces activités, on cite l'utilisation des terres agricoles sur un bassin versant, l'urbanisation, les rejets d'eaux usées, les stations thermiques et nucléaires et les barrages hydroélectriques. En ce qui concerne la pollution thermique, elle se manifeste principalement par les effluents thermiques des centrales nucléaires et énergétiques ayant servi de liquide de refroidissement amenant, ainsi, une perturbation du régime thermique du milieu aquatique (Prats *et al.*, 2012). Les effluents chimiques, le ruissellement de l'eau de pluie réchauffée sur le sol urbain et l'extraction d'eau potable pour des besoins domestiques ou en agriculture ne sont que quelques exemples d'activités ayant une incidence considérable sur la dynamique thermique dans les eaux douces (Council, 2004).

Changement climatique

On désigne souvent le « changement climatique anthropique » par la forme abrégée « changement climatique ». Dans ce contexte et selon divers scénarios, d'ici la fin du 21ème siècle, la température de l'air augmentera dans un intervalle de 1.1°C et 6.4°C (Meehl *et al.*, 2007). Une

hausse des températures de l'air conduit habituellement à une augmentation de la température de l'eau selon un patron assez similaire. Ainsi, plusieurs études signalent qu'il y aura des répercussions sur la thermie des rivières (Webb, 1996, Mohseni *et al.*, 1999, Poirel *et al.*, 2010, Van Vliet *et al.*, 2011). En guise d'exemple, Van Vliet *et al.* (2011) prédisent une augmentation de la température moyenne de l'eau de 3,8°C suite à une hausse de la température de l'air de 6°C. Par ailleurs, l'étude de Mohseni *et al.* (2003) montre que le réchauffement climatique aura une incidence directe sur les communautés piscicoles conduisant à une baisse de 36% au niveau de la quantité d'habitats thermiques favorables pour les poissons d'eau froide, au profit des poissons d'eau chaude qui augmenteront de 31% alors qu'une baisse de 15% pour les poissons d'eau tiède est prévue.

Déforestation

La déforestation a été identifiée comme source importante de perturbation du régime thermique d'une rivière et de nombreuses études ont été menées dans ce contexte (Caissie, 2006). Or, il est largement admis que l'exploitation forestière sans protection de la bande riveraine, réduit l'ombrage sur les rivières, ce qui occasionne une amplification de l'action du rayonnement solaire incident et par conséquent une hausse de la température de l'eau et un changement de l'habitat aquatique (Beschta *et al.*, 1987, St-Hilaire *et al.*, 2000). En effet, l'ombrage permet l'interception d'une partie du rayonnement solaire (Larson *et al.*, 1996) par le biais de la végétation riveraine, ce qui empêche l'insolation directe de l'eau en écoulement libre. De façon analogue, les travaux de Greenberg *et al.* (2012) montrent que la réduction de l'incidence du rayonnement solaire provenant à la rivière est possible via les canopées modérant ainsi une quantité des flux de chaleur. Il est à noter que dans le sud du Québec, de 30000 à 40000 km linéaires de cours d'eau auraient été aménagés par le ministère de l'agriculture du Québec (Grégoire *et al.*, 2007). Il est utile de mentionner que Johnson *et al.* (2000) rapportent une augmentation de 7 °C de la température quotidienne moyenne maximale de l'eau après la récolte forestière et estiment une reprise progressive du régime thermique naturel d'avant récolte dans une période de 15 ans. La même étude montre que le régime thermique saisonnier est altéré par la récolte forestière, ce qui est manifesté par l'augmentation hâtive des températures des eaux en rivières en début de l'été comparé à des secteurs non affectés par la récolte.

Barrages

De nombreuses études ont été menées en vue de comprendre et mettre en contexte l'impact des retenues sur le régime thermique en rivières (Poirel *et al.*, 2010). Un effet de lissage des cycles journaliers et/ou annuels en aval a été mis en évidence (Liu *et al.*, 2005). Cet effet dépend entre

autres de la profondeur de la prise d'eau, du mode de fonctionnement, du type d'ouvrage, du positionnement dans le bassin versant et de la taille du réservoir (Webb, 1996, Webb *et al.*, 1997, Bartholow *et al.*, 2004, Olden *et al.*, 2010). Par exemple, des petits barrages peuvent libérer de l'eau chaude directement des réservoirs et entraîner ainsi une augmentation des températures en aval (Olden *et al.*, 2010). Les mêmes résultats ont été retrouvés dans les travaux de Singer *et al.* (2011) qui confirment que l'installation d'un petit barrage en Alabama a occasionné un réchauffement thermique de la rivière à l'étude.

Par contre, comme le relate les travaux de Olden *et al.* (2010), les plus grands barrages gèrent volontairement des régimes thermiques en libérant de manière sélective de l'eau froide des réservoirs profonds pour maintenir les habitats thermiques adéquats pour les espèces d'eau froide. Ce même constat a été mis en évidence par le travail de thèse de Maheu (2015) pour la période d'été et d'automne. Outre les impacts sur le régime thermique des rivières, les barrages et retenues peuvent modifier les communautés aquatiques des rivières régulées (Bunn *et al.*, 2002, Poff *et al.*, 2010).

Il est à noter que tous ces facteurs d'intérêt précités s'ajoutent à un milieu qui possède une hétérogénéité temporelle complexe. Une rivière connait en général des variations à grandes échelles (e.g. annuelle, saisonnière) mais aussi à petites échelles (e.g. journalières, horaires) (Caissie, 2006).

1.4 Outils de modélisation de la température de l'eau en rivière

Vue l'importance de la température de l'eau en rivières soulignée ci-dessus, la modélisation et la prévision de cette variable est d'une grande importance. Il existe une large littérature traitant la prévision de la température de l'eau d'une rivière avec une grande variété de méthodes et modèles. Deux principales approches de modélisation de la température de l'eau ont été utilisées dans le passé à savoir les approches déterministes et statistiques (St-Hilaire *et al.*, 2000, Caissie. *et al.*, 2001, Benyahya *et al.*, 2007a, Zhu *et al.*, 2019).

1.4.1 Modèles déterministes

Les modèles déterministes ou physiques de la température de l'eau en rivière nécessitent un grand nombre d'intrants puisqu'ils tiennent compte de la majorité des paramètres hydrologiques et météorologiques qui sont parfois difficilement mesurables (Mohseni *et al.*, 1999, Caissie. *et al.*, 2001, Bélanger *et al.*, 2005, Benyahya *et al.*, 2007a). Ces modèles se basent sur les bilans hydrologiques et thermiques de la température de l'eau en rivière et nécessitent souvent un grand

temps de préparation et de calcul en raison de leur complexité (Bélanger *et al.*, 2005, Ahmadi-Nedushan *et al.*, 2007, Zhu *et al.*, 2018). Il peut être avantageux d'élaborer des modèles de prévisions plus simples.

1.4.2 Modèles statistiques

L'approche statistique est basée sur la structure temporelle (ou parfois spatiale) de la relation entre la variable réponse et les variables explicatives. Cette approche nécessite habituellement un nombre des paramètres beaucoup moins nombreux que l'approche déterministe (St-Hilaire *et al.*, 2000, Bélanger *et al.*, 2005, Laanaya, 2015).

Deux grandes catégories de modèles statistiques de prévision sont les plus utilisées : les modèles paramétriques et les modèles non paramétriques (Ahmadi-Nedushan *et al.*, 2007, Benyahya *et al.*, 2007a).

1.4.2.1 Modèles paramétriques

Les modèles paramétriques se basent généralement sur une relation statistique spécifiée. Ces modèles ont été largement utilisés avec succès pour prédire la température de l'eau en fonction d'une ou de plusieurs variables indépendantes (Mohseni *et al.*, 1998a, Neumann *et al.*, 2003). Ce type de modèles qui est relativement simple inclut la régression linéaire (Morrill *et al.*, 2005, Krider *et al.*, 2013, Zhu *et al.*, 2018), certains modèles de régression non linéaire (Mohseni *et al.*, 1998a, Mohseni *et al.*, 1999, Neumann *et al.*, 2003, Zhu *et al.*, 2019) ainsi que les modèles stochastiques (Caissie *et al.*, 1998, Bélanger *et al.*, 2005, Benyahya *et al.*, 2007a). Selon les travaux effectués, les modèles paramétriques sont moins adaptés à des échelles de temps journalières, vue l'autocorrélation forte de la température de l'eau (Caissie, 2006, Laanaya *et al.*, 2017). Plus particulièrement, on note que les modèles les plus référencés dans la littérature à savoir les modèles de régression linéaire sont souvent basés uniquement sur la relation entre la température de l'air et celle de l'eau. Ils sont connus par leur efficacité dans la prédiction sur des échelles de temps plus longues (e.g. hebdomadaires, mensuelles, annuelles) (Erickson *et al.*, 2000, Johnson *et al.*, 2000, Caissie, 2006, Benyahya *et al.*, 2007b, Laanaya *et al.*, 2017). Toutefois, ces modèles se voient moins appropriés en cas de non-linéarité vérifiée des données et ici il y a souvent recours aux modèles de régression non linéaire tel que le modèle logistique (sigmoïde) ajusté par Mohseni *et al.* (1998a).

1.4.2.2 Modèles non paramétriques

Quant aux modèles statistiques non paramétriques utilisés pour la prévision de la température de l'eau, ils ont l'avantage d'être souples et de ne pas imposer de forme a priori (Benyahya *et al.*, 2007a). Ce type de modèles inclue entre autres les K plus proches voisins (St-Hilaire *et al.*, 2000, Benyahya *et al.*, 2007b) ,les réseaux de neurones artificiels (RNA) (Bélanger *et al.*, 2005, Karacor *et al.*, 2007, Jeong *et al.*, 2013, Hadzima-Nyarko *et al.*, 2014, Piotrowski *et al.*, 2015) et les modèles additifs généralisés (Laanaya *et al.*, 2017). Malgré la grande diversité de modèles utilisés dans le domaine, la majorité présente des limites (Tableau 1.1). Les modèles statistiques les plus fréquemment utilisés sont basés sur l'hypothèse de la stationnarité de la série chronologique de la température de l'eau. Cependant, une caractéristique importante souvent observée dans les séries temporelles hydro-climatiques à petite échelle de temps est la présence de non-stationnarité à l'échelle temporelle saisonnière ou à long terme (Langan *et al.*, 2001, Benyahya *et al.*, 2007a).

Notons que les modèles non-paramétriques qui se montrent capables de traiter les problèmes inhérents à la série de la variable expliquée (e.g. RNA, K plus proches voisin) offrent une description peu claire de la relation entre les données d'entrées et de sortie (Benyahya *et al.*, 2007a). Le risque de sur-ajustement étant plus élevé qu'avec d'autres modèles, il est possible que leur capacité d'extrapolation soit limitée.

Problèmes inhérents de la série temporelle	Régression linéaire	Régression non linéaire	Modèles stochastiques (ARMA)	Réseaux des neurones artificiels	Modèles linéaires généralisés	Modèles additifs généralisés	K plus proches voisins
Non Linéarité	-	+	-	+	-	+	+
Non stationnarité	-	+	-	+	-	+	+
Saisonnalité	-	-	-	+	-	-	+
Non Normalité	-	+	-	+	+	+	+

(+) : la méthode traite le problème désigné

(-) : la méthode ne traite pas le problème désigné

Tableau 1.1 : Limites des modèles statistiques

La présente étude s'inscrit, ainsi, dans un cadre de régression visant la modélisation de la relation entre les températures journalières de l'eau et de l'air tout en traitant les défis susmentionnées associés aux propriétés statistiques des séries temporelles qui peuvent souvent exister. À cet égard, il convient de souligner que la transformation des données comme étape préliminaire à l'analyse de régression est souvent nécessaire dans les études se dotant de séries chronologiques problématiques. En outre, le besoin de prendre en compte à la fois la variation temporelle et fréquentielle complexe des séries de données a suscité l'émergence des méthodes d'analyse temps-fréquence ou temps échelles. Ces approches, dites aussi de décomposition, ont été suggérées par de nombreux chercheurs (Thioune, 2015a). En tenant compte de l'aspect de données de séries chronologiques, des nombreux chercheurs ont proposés différentes approches de décomposition spectrale dans divers domaines d'application parmi lesquelles l'analyse de Fourier (Dominici *et al.*, 2003), la transformée en ondelettes (Küçük *et al.*, 2006, KisÌ§i, 2009) et la décomposition modale empirique (Huang *et al.*, 1998b, Qin *et al.*, 2016, Masselot *et al.*, 2018).

Cependant, la transformée de Fourier est naturellement limitée car elle n'est pas capable de prendre en compte la localisation dans le temps des échelles de variation (Thioune, 2015a), ce qui en fait une méthode peu adéquate pour les séries non stationnaires (Sifuzzaman *et al.*, 2009). Si les ondelettes ne souffrent pas de ces problèmes, une connaissance à priori sur le signal à décomposer est nécessaire pour un choix d'ondelette adéquat à chaque type de signal (Thioune, 2015a).

Récemment, une méthode de décomposition, appelée décomposition en mode empirique (EMD), a généré des progrès dans l'analyse temps-fréquence des séries temporelles et s'est avérée robuste et efficace dans l'analyse de données non linéaires, non stationnaires et bruitées (Huang *et al.*, 1998b). Cette approche présente l'avantage, par rapport à l'analyse par ondelettes, de ne pas nécessiter de connaissances a priori, ce qui la rend entièrement adaptative et souvent très efficace (Huang *et al.*, 2008). Cette propriété a permis à la méthode de s'introduire avec succès dans de nombreuses applications (Huang *et al.*, 1998b, Rilling, 2007b, Lee *et al.*, 2012). Elle a été appliquée, entre autres, en sismologie (Loh *et al.*, 2001), en océanographie (Huang *et al.*, 1999), en biologie (Lio, 2003) et en hydrologie (Lee *et al.*, 2010, Durocher *et al.*, 2016). Or, d'après la littérature, l'approche par EMD n'a pas été utilisée auparavant pour la modélisation de la température de l'eau.

La décomposition par EMD des séries temporelles considérées produit des composantes oscillatoires plus un résidu (Huang *et al.*, 1998b, Loh *et al.*, 2001, Rilling *et al.*, 2003). Ces

composantes, appelées fonctions de mode intrinsèque (IMF), contiennent les informations d'une seule fréquence et peuvent être liées à la même gamme de fréquences d'autres variables étudiées (Boudraa *et al.*, 2007, Fan *et al.*, 2017, Chu *et al.*, 2018). Ceci permet d'identifier les relations temps-fréquence entre les variables étudiées. Les IMFs sont intégrées ultérieurement dans la régression sous la forme d'un ensemble de variables explicatives potentielles (Yang *et al.*, 2011b, Qin *et al.*, 2016). Comme les IMFs sont souvent nombreuses et qu'elles ne sont pas toutes pertinentes, le défi consiste à éliminer celles qui sont insignifiantes et à ne retenir que les variables explicatives les plus significatives. Bien que le modèle de régression linéaire classique soit largement utilisé, il manque souvent de précision et le grand nombre de composantes rend l'interprétation plus difficile (Tibshirani, 1996). Récemment, une méthode de régression populaire, à savoir le LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996), a été appliquée avec succès dans un contexte de régression linéaire à haute dimension, permettant de réduire le nombre de régresseurs étudiés et de développer des modèles parcimonieux (Bernard *et al.*, 2018). Cette méthode permet, en effet, d'améliorer la qualité de la prédiction et de fournir des interprétations plus précises (Tibshirani, 1996). En particulier, dans deux études récentes basées sur le modèle EMD-R, la régression LASSO a été utilisée pour sélectionner les IMFs les plus significatives (Qin *et al.*, 2016, Masselot *et al.*, 2018).

1.5 Objectif de l’étude

Le présent projet de recherche vise à modéliser la température de l’eau journalière en rivière en fonction de la température de l’air dans un cadre de régression. Dans cette étude, on se propose de se servir de la méthode de régression par décomposition modale empirique (EMD-R) et étudier son potentiel comme outil de prévision de la température de l'eau en rivières dans une optique d’aide à la décision et à la gestion des ressources hydriques. L’étude est appliquée à plusieurs rivières d'Amérique du Nord. L'EMD-R sera comparée, en termes de performance prédictive, à deux autres modèles de régression souvent utilisés dans la modélisation de la température de l'eau, à savoir les modèles GAM et le modèle de régression logistique (sigmoïde).

2 SYNTHÈSE DES RESULTATS

Les détails complets ainsi que tous les tableaux, les figures et les équations sont présentées dans la section 4 (Article).

2.1 Méthodologie

Notre approche de modélisation non paramétrique est l'EMD-R (Régression par Décomposition Modale Empirique). Cette technique permet de décrire plus précisément la relation entre une variable réponse et des variables explicatives en choisissant les variables les plus importantes (Qin *et al.*, 2016, Masselot *et al.*, 2018). Deux modèles statistiques qui décrivent la relation entre la température de l'eau et la température de l'air seront comparés au modèle EMD-R proposé à savoir le modèle additif généralisé et le modèle logistique (Sigmoïde) (Wehrly *et al.*, 2009, Laanaya *et al.*, 2017).

2.1.1 Régression par décomposition modale empirique (EMD-R) :

L'EMD-R se décompose de deux étapes principales (Figure 4.2). Premièrement, la décomposition modale empirique (EMD) est appliquée à la variable explicative, pour obtenir des composantes oscillatoires appelées fonctions en mode intrinsèque (IMFs) et une partie résiduelle non oscillatoire. Deuxièmement, les IMFs produits sont considérés comme de nouvelles variables explicatives dans l'opérateur de sélection et de réduction la plus faible en valeur absolue (LASSO) pour sélectionner les IMFs qui sont les meilleurs prédicteurs (Tibshirani, 1996, Qin *et al.*, 2016).

EMD et le processus de tamisage : Comme l'exprime son expression (Équation 2.1), l'algorithme EMD suppose que la série temporelle originale $x(t)$ peut être décomposée en plusieurs sous-séries ($IMF_k(\mathrm{t})$), en plus d'un résidu non oscillant $r_K(t)$ (Huang *et al.*, 1998b).

$$\mathrm{x(t)} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{IMF_k(t)} + \mathrm{r_K(t)}, \quad \mathrm{t} = 1,2,\ldots \mathrm{T} \qquad (2.1)$$

Les IMFs doivent satisfaire deux conditions principales : (i) avoir une moyenne locale nulle à tout moment t ; (ii) le nombre d'extrema et le nombre de passages par zéro doivent être égaux ou différer au maximum de un (Huang *et al.*, 1998b, Boudraa *et al.*, 2007, Huang *et al.*, 2008, Lee *et al.*, 2012). Ces deux conditions permettent d'obtenir des IMFs qui oscillent symétriquement autour de zéro. Les IMFs sont obtenus de manière itérative en utilisant l'approche suivante (Huang *et al.*, 1998b) :

a) Identifier les maxima et minima locaux de $x(t)$ et les interpoler respectivement pour générer les enveloppes supérieure et inférieure $x_{\max}(t)$ et $x_{min}(t)$.
b) Calculer la moyenne locale $m(t) = (x_{\max}(t) + x_{min}(t))/2$
c) Retrancher $m(t)$ à $x(t)$ pour obtenir le prototype $h(t) = x(t) - m(t)$.

Si $h(t)$ remplit les deux conditions mentionnées ci-dessus, alors $h(t)$ est soustrait à la série et devient la composante $IMF_1(t)$. Sinon, il faut répéter les étapes précédentes sur $h(t)$ jusqu'à ce qu'il remplisse les conditions d'un IMF. $h(t)$ est alors le premier IMF ($IMF_1(t)$).

d) Répétez les étapes a à c sur le résidu $r_1(t) = x(t) - IMF_1(t)$ jusqu'à ce que le résidu obtenu contienne au maximum un extrême. Le résidu final est alors considéré comme une estimation de la tendance à basse fréquence de la série chronologique.

Un problème reconnu de l'algorithme EMD est le mélange de modes. Il se présente sous la forme de fréquences différentes dans un même IMF ou lorsqu'une fréquence est partagée entre deux IMFs. Ce problème est, souvent, résolu par l'ensemble EMD (EEMD) qui consiste à ajouter du bruit blanc à $x(t)$. Cette opération est répétée N fois pour obtenir la moyenne de tous les ensembles des IMFs calculés (Zhang *et al.*, 2010, Wang *et al.*, 2018). L'algorithme de l'EEMD est décrit par Wu *et al.* (2009) comme suit :

a) Ajouter une série de bruit blanc aux données ciblées;
b) Décomposer les données obtenues de l'étape a) en IMFs;
c) Répéter les étapes a) et b) N fois mais en choisissant, à chaque fois, des séries de bruit blanc différentes; et enfin
d) Se procurer les moyennes (ensemble) des IMFs correspondants des décompositions comme résultat final.

Bien que l'EEMD résout le problème du mélange de modes, il est très important de choisir le nombre approprié de répétitions et l'écart-type du bruit blanc ajouté. Le choix de ces paramètres affecte la qualité de la décomposition et ses résultats (Zhang *et al.*, 2010).

Régression LASSO : Proposée par Tibshirani (1996), la régression LASSO est une méthode de pénalisation en régression. Pour un modèle de régression, le principe de base du LASSO est d'estimer les coefficients de régression β en minimisant l'expression des moindres carrés pénalisés (Équation 2.2)**.** Plus le λ est élevé (coefficient de pénalisation), plus la régularisation est forte. Ce paramètre de régularisation contrôle directement le nombre de variables explicatives qui restent dans le modèle final. La valeur la plus adéquate de λ correspondant au minimum de

l'erreur quadratique moyenne (MSE) est généralement estimée par validation croisée. Le principal avantage de l'utilisation du LASSO par rapport aux autres méthodes de régression est qu'elle permet de sélectionner des variables en annulant certains coefficients de régression (Tibshirani, 2011, Qin *et al.*, 2016, Chu *et al.*, 2018). Les prédicteurs sélectionnés par LASSO formeront le modèle de prédiction final de l'EMD-R.

$$\hat{\beta} = \arg_{\beta}\min\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j X_j\right) + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \qquad (2.2)$$

Ou y est la variable réponse, $X_j\ (j = 1 \ldots p)$ sont les variables explicatives et λ est le coefficient de pénalité.

2.1.2 Modèle additif généralisé (GAM)

Le MAG est un modèle linéaire généralisé avec un prédicteur impliquant une somme de fonctions lisses de covariables. Cette approche, qui a été définit par Hastie *et al.* (1986), possède un ensemble des avantages. Le modèle MAG ne nécessite pas la contrainte de forte linéarité, au contraire, il donne des bons résultats dans le cadre non linéaire. Le principe de base de la régression MAG est de modéliser la variable réponse à partir de la somme des fonctions de variables explicatives (Équation 2.3).

L'application du modèle MAG se base sur l'estimation des fonctions de lissage $f_i(x_i)$ qui sont des splines cubiques appelées aussi splines pénalisées (Wood, 2017). Mathématiquement parlant, ces splines sont définies comme la solution au problème d'optimisation suivant : parmi les deux fonctions différentiables en continu, ne conservant que celles qui minimisent la somme des carrés pénalisés, ce que l'on appelle la somme résiduelle pénalisée des carrés (Laanaya *et al.*, 2017):

$$argmin_{f_i \epsilon C^2}(\|y - f_i(x_i)\|^2 + \sum_i \lambda_i \int f_i^n(x)^2 dx) \qquad (2.3)$$

avec λ_p qui représente les pénalités sur la rugosité de la fonction ajustée associé à chaque variable explicative. Ce paramètre contrôle le niveau de lissage de chaque fonction f_i.

2.1.3 Modèle logistique (Sigmoïde)

Le modèle logistique de régression est une fonction non linéaire qui permet d'estimer et de modéliser la température de l'eau dans la rivière (variable réponse) en fonction de la température moyenne de l'air. L'ajustement du modèle nécessite l'estimation de trois paramètres à savoir α qui est le coefficient d'estimation de la température de l'eau maximale; β qui est la valeur de la

température de l'air au point d'inflexion et γ représente la plus forte pente de la fonction logistique (Équation 2.4) (Mohseni *et al.*, 1998a, St-Hilaire *et al.*, 2012).

$$y = \frac{\alpha}{1+e^{\gamma(\beta-x)}} \qquad (2.4)$$

Ou y et x représentent respectivement les températures de l'eau et de l'air.

2.1.4 Critères de performances

Dans cette étude, on utilise quatre critères de performance pour évaluer la robustesse des différentes approches à savoir le coefficient de détermination R-carré (R^2), la racine de l'erreur quadratique moyenne (RMSE), le biais (B) et validation croisée généralisée (GCV) (voir les équations (4.5), (4.6), (4.7) et (4.8) (Ahmadi-Nedushan *et al.*, 2007, Benyahya, 2007, Laanaya *et al.*, 2017, Zhu *et al.*, 2018).

2.2 Données et sites d'étude

Dans cette étude, deux rivières au Canada et deux autres aux États-Unis sont prises en compte, avec des séries chronologiques de température de l'eau relativement longues (> 14 ans) et des superficies de bassin versant très différentes. Des séries chronologiques quotidiennes de la température moyenne de l'eau et de la température moyenne de l'air ont été enregistrées dans des stations hydrométriques et météorologiques pour différentes périodes d'une station à l'autre. En outre, comme il manque de nombreuses valeurs avant Mai et après Octobre, la période d'étude hors hiver avec des données continues sera différente d'une rivière à l'autre.

Pour le fleuve Missouri et le fleuve Potomac, les données saisonnières de température journalière moyenne ont été obtenues auprès de l'United States Geological Survey (USGS 2017), et pour les deux rivières du Canada, la rivière Trinité et le ruisseau Catamaran, les données ont été extraites de la base de données Rivtemp (http://rivtemp.ca/rivtemp-data/) (voir Figure 4.1).

Le ruisseau Catamaran est situé au centre du Nouveau-Brunswick (Canada) et est un affluent de la petite rivière Miramichi sud-ouest, avec une latitude de 46 52,7' N et une longitude de 66 06,0' O. Ce ruisseau a une aire de drainage de $51\ km^2$. Les données sur la température de l'eau et de l'air ont été enregistrées de mai 1993 à septembre 2010.

La rivière Trinité est située près de la municipalité de Baie-Trinité, à 95 km à l'est de Baie-Comeau dans la province de Québec (Canada). La superficie drainée est $562\ km^2$. Les séries chronologiques de température sont disponibles de mai 1985 à octobre 2017. Le Missouri est le principal affluent du fleuve Mississippi, qui s'écoule sur plus de 3600 km de trois Forks au

Montana à St. Louis, Missouri (Zhu et al., 2018). Les séries chronologiques disponibles et complètes des températures de l'eau et de l'air vont de mai 2001 à juillet 2015. Le Potomac est un fleuve de l'est des États-Unis, sa longueur est de 655 km. Il prend sa source à une altitude de 933 m au sud-ouest de l'État du Maryland. Son bassin versant est $38{,}018 km^2$ et son débit annuel moyen est estimé à $306\, m^3/s$. Les données sont disponibles pour l'étude de juin 2001 à septembre 2015. Les détails concernant les différentes stations météorologiques et hydrométriques des rivières étudiées sont présentés dans le Tableau 4.1.

Malgré qu'on utilise une seule variable explicative (température de l'air), la décomposition de cette dernière par EMD résulte en un ensemble de composantes (IMFs). Ces dernières vont être introduites comme nouvelles variables explicatives dans le modèle EMD-R. Par conséquent, le LASSO est utilisé afin de sélectionner les IMFs les plus pertinents.

2.3 Principaux résultats

Une présentation graphique des séries temporelles de la température de l'eau et celle de l'air de la rivière pour les quatre cas d'étude nous donne une idée préliminaire sur leurs variations journalières importantes. Selon les Figure 4.3, Figure 5.1 et Figure 5.2, on remarque que les amplitudes de variations de la température de l'air sont plus importantes que celles de la température de l'eau. D'un autre côté, les séries hydro-climatiques contiennent souvent une composante saisonnière qui est remarquable par analyse visuelle. Les Figure 4.3, Figure 5.1 et Figure 5.2 suggèrent une relation non linéaire entre la variable expliquée qui est la température de l'eau et la variable explicative qui est la température de l'air. A cet égard, nous avons appliqué la régression par décomposition modale empirique sur la température de l'air afin de bien étudier le régime thermique de chaque rivière. Seuls les résultats de la rivière Missouri (Etats Unis) et le ruisseau Catamaran (Canada) sont détaillés dans la discussion des résultats qui suivent. Ce choix a été fait vu le grand nombre de données, de résultats et de discussions à présenter pour les trois modèles statistiques étudiés et comparés.

Résultats de l'EMD-R :

Pour nos cas d'études (Figure 4.7), nous appliquons l'EEMD, cette version est développée pour résoudre le problème du mélange des modes rencontré avec l'EMD classique (Thioune, 2015a). Dans ce travail, les paramètres de l'EEMD sont choisis en référant aux travaux précédents (Rilling *et al.*, 2003, Rehman *et al.*, 2013) à savoir un seul bruit blanc avec une variance de 10% et un nombre d'ensembles assez grand (Ne=1000).

Les résultats de la décomposition EEMD montrent une séparation claire des fréquences des IMFs. Cela indique que les résultats de l'EEMD donnent des composantes qui peuvent être interprétées. Par la suite, on additionne les IMFs qui possèdent la même fréquence en obtenant finalement 10 composantes IMFs pour le cas du Missouri et le cas du Catamaran. Selon le Tableau 4.3, on peut voir que pour les deux études de cas, les composantes IMF1 et IMF2 présentent des pics quasi réguliers d'une durée moyenne comprise entre 3 et 6 jours, avec une amplitude moyenne variant entre 2°C et 3°C pour la rivière Missouri et entre 2,5°C et 3,5°C pour le ruisseau Catamaran. Ces oscillations aléatoires de haute fréquence peuvent être liées aux périodes chaudes de la saison estivale durant lesquelles la température de l'air enregistre des valeurs élevées. Les IMF3 et IMF4 ont une période moyenne comprise entre une et trois semaines avec une amplitude proche de celle des deux premières composantes. La composante IMF5 a une période moyenne d'environ 40 jours, avec une amplitude relativement faible par rapport aux premières composantes. Les composantes IMF6 et IMF7 sont des composantes biannuelles d'une durée moyenne d'environ 6 mois et d'une amplitude plus importante que les composantes précédentes. Les causes soupçonnées de ces cycles sont attribuées aux cycles semestriels et annuels de la circulation atmosphérique. Les autres composantes représentent des variations interannuelles. Le IMF8 est quasi-biannuel, et le IMF9 a une période moyenne légèrement supérieure à trois ans. Pour les deux dernières composantes IMF10 et IMF11:17 ou IMF11:16, la période dépasse trois ans, l'amplitude moyenne étant d'environ 5°C pour le IMF10 et variant entre 1,5°C et 2,5°C pour le IMF11:16 et le IMF11:17 respectivement pour les deux études de cas.

La Figure 4.8 montre un graphique du MSE pour différentes valeurs de λ. Lorsque la valeur de λ augmente, les coefficients de régression tendent vers zéro et le MSE devient plus élevée, ce qui indique que le pouvoir prédictif du modèle est faible. Alors que, lorsque la valeur de λ diminue, les coefficients de régression n'atteignent pas zéro et le graphique semble s'aplatir. Le modèle ayant un faible MSE associé au plus petit λ (c'est-à-dire 0,079 pour la rivière Missouri et 0,097 pour le ruisseau Catamaran) est identifié dans la Figure 4.8. Dans cette figure, les points rouges représentent les MSE, les lignes verticales représentent la valeur de λ sélectionnée selon la méthode des MSE et l'axe horizontal en haut représente le nombre de IMFs restant dans le modèle pour la valeur appropriée de λ. Pour la rivière Missouri, le LASSO conserve tous les IMFs pendant la décomposition en accordant à chacun un coefficient de régression. Cependant, dans le cas du ruisseau Catamaran, le LASSO a accordé la valeur zéro à l'IMF1 et à l'IMF10, en ne retenant que 8 parmi les 10 obtenus. Nous notons que la composante IMF6+7 a enregistré le coefficient de régression le plus élevé pour les deux études de cas, ce qui montre l'effet de cette

composante sur notre modèle de régression obtenu. En revanche, les composantes IMF1 et IMF2 ont obtenu respectivement les coefficients de régression les plus faibles ; dans le cas du rivière Missouri et du ruisseau Catamaran, ces composantes ont un effet moins important que les autres IMFs.

Résultats du GAM : La Figure 4.5 montre les effets de la température de l'air sur la température de l'eau. Pour le ruisseau de Catamaran, la relation estimée entre la température de l'air et de l'eau est clairement non linéaire avec une forme en S, en particulier entre 12,5 °C et 22,5 °C (Figure 4.5b). Aux valeurs extrêmes des températures de l'air, les effets de lissage s'aplatissent. En revanche, pour la rivière Missouri, le graphique des effets de lissage montre une relation presque linéaire entre la température de l'air et celle de l'eau (Figure 4.5a). Les résultats du MAG pour les deux études de cas mentionnées dans le Tableau 4.2 montrent les effets non-linéaires de la température de l'air avec une valeur de probabilité inférieure à 0,0001. Cette dernière montre que la composante non linéaire n'est pas négligeable.

On remarque que la fonction de lissage de la température de l'air du MAG dans le cas de ruisseau Catamaran est très proche de celle du modèle sigmoïde (voir Figure 4.4 et Figure 4.5).

Résultats du Logistique (Sigmoïde) :

La Figure 4.4 montre la régression logistique ajustée entre la température de l'eau et de l'air. On remarque qu'il y a une forte dispersion entre la température moyenne quotidienne de l'eau et de l'air. L'application du modèle a donné des variances totales expliquées égales à 80,39% (la plus élevée de toutes les stations) et 55,30% (la plus faible) pour le fleuve Missouri et le ruisseau Catamaran respectivement. Les équations du modèle qui en résulte pour le fleuve Missouri et le ruisseau Catamaran sont mentionnées dans l'article (Équation 4.9 et Équation 4.10).

2.4 Comparaison des résultats

Les performances des modèles EMD-R, GAM et Logistique pour les quatre études de cas sont présentées dans le Tableau 4.4. De manière générale, l'EMD-R est plus performant que les autres modèles. Elle a enregistré le R^2 le plus élevé avec une variance expliquée entre 87,58 % pour la rivière Trinité et 91,41 % pour la rivière Missouri. En comparaison, les coefficients de détermination les plus bas et les plus élevés de la régression logistique et du MAG sont respectivement d'environ 55 % pour le ruisseau Catamaran et de 80 % pour le Missouri.

Le critère RMSE, indique une meilleure performance de l'EMD-R avec des valeurs allant de 1,01 ° C à 2,38 ° C pour les quatre études de cas. On peut noter que les valeurs RMSE obtenues pour

les modèles MAG et Logistique sont très proches mais avec un résultat légèrement meilleur pour le MAG. Pour le GCV, l'EMD-R est à nouveau le modèle le plus performant pour les quatre études de cas avec une valeur de 1,03 pour la rivière Missouri et de 5,69 pour le fleuve Potomac. Pour les autres cas de comparaison, les GCV sont très proches mais le MAG est toujours meilleur que le modèle logistique. Pour le critère de biais, c'est le MAG et le modèle logistique qui ont donné les valeurs les plus proches de zéro, mais cela se justifie par le fait que l'utilisation du LASSO biaise la régression.

3 CONCLUSION ET RECOMMANDATIONS

L'objectif principal de ce travail était de modéliser la température moyenne quotidienne de l'eau dans quatre rivières en utilisant la température moyenne de l'air. Nous proposons de comparer une nouvelle méthode, EMD-R, à d'autres méthodes couramment utilisées (GAM et Sigmoïde). Les modèles EMD-R, GAM et Logistique (Sigmoïde) sont testés en utilisant les critères de performance suivants : R-carré, RMSE, GCV et Biais. L'EMD-R a montré une performance prédictive supérieure à celle du MAG et du modèle logistique en termes de R-carré, GCV et RMSE. L'EMD-R offre la possibilité d'exploiter les composantes du signal de température de l'air à différentes fréquences, tout en conservant les avantages des approches non paramétriques (par exemple, pas de définition des fonctions a priori ou des distributions ; pas d'imposition de la stationnarité). Cette étude a été réalisée sur quatre rivières avec de données journalières ou il existe un peu de station de mesure météorologiques et hydrologiques. Enfin, on peut conclure que l'EMD-R est une méthode performante dans la gestion environnementale qui peut être une approche efficace dans la modélisation de variables hydro-climatologiques.

Il serait important d'étudier la température de l'eau sur des rivières qui possèdent plusieurs stations de mesure. Aussi, Il serait intéressant que les travaux futurs incluent l'étude du potentiel de l'EMD-R à des petit pas de temps (e.g. horaire) vu la grande variabilité de la température de l'eau et par conséquent son influence potentielle sur la tolérance des espèces aquatiques et plus spécifiquement les poissons (par exemple le saumon). D'un autre côté avec plus de deux variables (e.g. débit, humidité, vent, etc), l'étude de la température de l'eau sera plus robuste et efficace, où chaque variable peut être intégrée avec une structure complexe qui nécessite des méthodes plus sophistiquées et plus avancées. Ces méthodes permettent de décrire les relations réelles entre les différentes variables, qui sont souvent non linéaires.

4 ARTICLE

Empirical mode decomposition regression to predict river water temperature

By

Ramzi Abaza[a*]

Fateh Chebana[a]

André St-Hilaire[ab,]

Pierre Masselot[a]

[a] Institut National de la Recherche Scientifique : Centre Eau Terre Environnement
490 de la couronne, Québec, G1K9A9, Canada

[b] Canadian River Institue, University of New Brunswick, Fredericton, Canada

Manuscript submitted

05 November 2020

[a*] Corresponding author : email : ramzi.abaza@ete.inrs.ca

Abstract

Water temperature has a significant influence on the aquatic ecosystem, including impacts on water quality as well as on the metabolism and distribution of aquatic species. It is therefore essential to develop reliable tools to predict water temperature. The objective, in this research work, is to introduce empirical mode decomposition regression (EMD-R) for the prediction of daily water temperature using air temperature as a predictor. EMD-R is hereby compared to two classical statistical models: Generalized Additive Model (GAM) and the logistic or sigmoid regression. This comparison is performed on data from two rivers in the United States and two rivers in Canada during the ice-free period. These three models are evaluated using four performance criteria, namely the Root Mean Square Error (RMSE), the coefficient of determination (R^2), the Generalized Cross Validation (GCV) and the Bias. For the four studied cases, the EMD-R model is generally the one providing the best performance compared to the other statistical tested models. For the Missouri River and Catamaran Brook case studies, EMD-R gives respectively a RMSE of 1.01°C and 1.57°C versus values of 1.71°C and 3.20°C for the GAM, the most competitive model. The same superior performance is shown through the GCV and R^2 criteria.

Keywords: Water temperature, Air temperature, Empirical Mode Decomposition, Regression, LASSO, Prediction

Résumé

La température de l'eau a une influence importante sur l'écosystème aquatique, notamment sur la qualité de l'eau ainsi que sur le métabolisme et la distribution des espèces aquatiques. Il est donc essentiel de développer des outils fiables pour prédire la température de l'eau. L'objectif, dans ce travail de recherche, est d'introduire la régression de décomposition en mode empirique (EMD-R) pour la prévision de la température quotidienne de l'eau en utilisant la température de l'air comme prédicteur. L'EMD-R est ainsi comparée à deux modèles statistiques classiques : le modèle additif généralisé (GAM) et la régression logistique ou sigmoïde. Cette comparaison est effectuée sur les données de deux rivières aux États-Unis et de deux rivières au Canada pendant la période sans glace. Ces trois modèles sont évalués à l'aide de quatre critères de performance, à savoir la racine carrée de l'erreur quadratique moyenne (RMSE), le coefficient de détermination (R^2), la validation croisée généralisée (GCV) et le biais. Pour les quatre cas étudiés, le modèle EMD-R est généralement celui qui offre les meilleures performances par rapport aux autres modèles testés statistiquement. Pour les études de cas du fleuve Missouri et du ruisseau Catamaran, EMD-R donne respectivement une RMSE de 1,01°C et 1,57°C contre des valeurs de 1,71°C et 3,20°C pour le GAM, le modèle le plus compétitif. La même performance supérieure est démontrée par les critères GCV et R^2.

Mots-clés : Température de l'eau, Température de l'air, Décomposition en mode empirique, Régression, LASSO, Prédiction

4.1 Introduction

River water temperature is a very important variable in aquatic ecosystem studies (Cluis, 1972, Caissie. *et al.*, 2001, Zhu *et al.*, 2019) as an indicator of the health of aquatic systems and water quality (Gu *et al.*, 2002, Bélanger *et al.*, 2005). Indeed, water temperature influences various other water quality variables such as dissolved oxygen concentration (Ficklin *et al.*, 2013) and the metabolic activity of aquatic organisms (Allen *et al.*, 2005, Demars *et al.*, 2011, Sandersfeld *et al.*, 2017). In addition, river water temperature plays a key role for stenotherm fish that can only adapt to a specific temperature range (Edwards *et al.*, 1979, Bovee, 1982). A change in thermal regime in rivers has significant impacts on fish, distribution and habitat quality. Thus, a shift in the thermal regime may be putting a significant number of fish at risk (Isaak *et al.*, 2012, Hedger *et al.*, 2013).

River thermal regime is governed by anthropogenic impacts such as impoundment, agriculture, deforestation and direct sources of thermal pollution (Ahmadi-Nedushan *et al.*, 2007, Dupuis *et al.*, 2009). Geophysical variables such as river depth, groundwater input and turbulence are also among the main drivers of river water temperature variability (Crisp *et al.*, 1982, Caissie. *et al.*, 2001). Moreover, heat balance of a river is greatly influenced by meteorological variables, including solar radiation, air temperature, relative humidity, wind speed, etc. (Bélanger *et al.*, 2005, Zhu *et al.*, 2018).

For this study, air temperature was used as the only independent, since it is deemed to have the most significant impact on water temperature variation in rivers and is readily available (Cluis, 1972, Erickson *et al.*, 2000, Caissie, 2006, Benyahya *et al.*, 2007a).

There is a large body of literature on predicting river water temperature, that describe different approaches. They are generally classified in three groups, namely the deterministic, stochastic and regression approaches (Caissie, 2006, Benyahya *et al.*, 2010).

Deterministic models typically require a large number of input variables and are bases on the calculation of a thermal budget to predict river water temperature. Hence they are sometimes deemed, relatively complex and time consuming (Caissie. *et al.*, 2001, Bélanger *et al.*, 2005, Zhu *et al.*, 2018).

Unlike the deterministic approaches, statistical models, including regression and stochastic models are often more straightforward in application than deterministic approaches to predict water temperature, using a fewer number of input variables (Bélanger *et al.*, 2005, Benyahya *et al.*, 2007a) and are often based solely on the air-water temperature relation. Simple and multiple

linear regression models are the most referenced models in the literature that predict water temperature more efficiently for longer time scales (e.g. weekly, monthly and annually) than for a daily scale (Erickson *et al.*, 2000, Caissie, 2006, Benyahya *et al.*, 2007a, Laanaya *et al.*, 2017). These models are, however, less suitable when a nonlinearity of the relationship can be verified in the data (Erickson *et al.*, 2000, Ahmadi-Nedushan *et al.*, 2007). In a similar case, the non-linear regression model the most commonly used to determine the air to water relation is the logistic regression function (also called sigmoid function) (Mohseni *et al.*, 1998a, Caissie. *et al.*, 2001, Caissie, 2006). As for liner regressions, this model is underperforming in some cases when using daily data due to the autocorrelation within the water temperature time series(Mohseni *et al.*, 1998a, Caissie. *et al.*, 2001, Webb *et al.*, 2003). Hence, stochastic models are often preferred in predicting water temperature on a daily or a sub-daily basis; however, they are not appropriate when residuals are non-stationary(Cluis, 1972, Caissie *et al.*, 1998, Benyahya *et al.*, 2007a). In a non-linear regression context, the Generalized Additive Model (GAM), which is a non parametric regression model, has shown great flexibility in modelling stream temperature (Wehrly *et al.*, 2009, Laanaya *et al.*, 2017) while outperforming the logistic regression for mean daily air temperatures (Laanaya *et al.*, 2017).

The majority of statistical models, including those aforementioned, have limitations. The most frequently used statistical models are based on the assumption of stationarity of the water temperature time series. However, an important characteristic often observed in small-scale hydro-climatic time series is the presence of non-stationarity at the seasonal or long-term temporal scales.

Non-parametric models (e.g. Artificial Neural Networks (ANN), K-Nearest Neighbours (KNN)), which are able to deal with the problems associated with the endogenous variable series , although they are intuitive, offer an unclear description of the relationship between input and output data which makes them poor extrapolators (Benyahya *et al.*, 2007a, Benyahya *et al.*, 2010).

Therefore, the present study aims to model daily water temperature to air temperature relationship in a regression framework, while tackling the aforementioned challenges associated with the statistical properties of time series (e.g. seasonality, normality of residuals, stationarity, etc.). It should be noted that data transformation as a preliminary step to regression analysis is often required in studies with problematic time series. In this regard, the need to improve representation of time series often showing complex fluctuations over time, like water temperature with multiple periodicities, has led to the emergence of time-frequency or time-scale analysis methods. These,

also called decomposition approaches, have been suggested by many researchers. They include Fourier analysis (Dominici *et al.*, 2003), wavelet transform (Küçük *et al.*, 2006, KisÌ§i, 2009, Qin *et al.*, 2016) and empirical mode decomposition (EMD) (Huang *et al.*, 1998b, Qin *et al.*, 2016).

Fourier analysis is naturally limited because it cannot adjust to shifts in frequencies or periods (Thioune, 2015a). In addition, this technique also requires linearity (Huang *et al.*, 1998b) and have been found to be ineffective in a non-stationary context (Sifuzzaman *et al.*, 2009). Although the wavelet method does not suffer from these problems, an a priori knowledge of the data to be decomposed is needed (Thioune, 2015a).

Recently, a new decomposition method, called empirical mode decomposition (EMD) has generated advances in the time-frequency analysis of time series/signals and has proven to be robust and effective in analyses of nonlinear, non stationary and noisy data (Huang *et al.*, 1998b). This approach has the advantage over wavelet analysis that it does not require a priori knowledge, which means that it is entirely adaptive and often highly efficient (Huang *et al.*, 2008). The usefulness of the method has made it widely and successfully introduced in many applications (Huang *et al.*, 1998a, Rilling, 2007a, Lee *et al.*, 2012). Although EMD is a relatively recent method, it has been applied in oceanography (Huang *et al.*, 1999), seismology (Loh *et al.*, 2001), biology (Lio, 2003), hydrology (Lee *et al.*, 2010, Durocher *et al.*, 2016). However, to our knowledge, the EMD have not been previously used for water temperature modelling.

The decomposition, by EMD method, of the considered time series yield oscillatory components plus a residue (Huang *et al.*, 1998b, Loh *et al.*, 2001, Rilling *et al.*, 2003). These components are called Intrinsic Mode Functions (IMFs), which contain the information of a single frequency, and can be related to the same frequency range of other studied variables (Boudraa *et al.*, 2007, Fan *et al.*, 2017, Chu *et al.*, 2018). This allows for the identification of the time-frequency relationships between the variables studied. The IMFs will be integrated subsequently in the regression as a set of potential explanatory variables (Yang *et al.*, 2011a, Qin *et al.*, 2016). As IMFs are often numerous and not all of them are relevant, the challenge is to eliminate insignificant ones and retain only the most significant explanatory variables. Although, conventional linear regression model is widely used, it often lacks precision and the large number of components makes interpretation more difficult (Tibshirani, 1996). Recently, a popular regression method namely Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) was successfully applied in high-dimensional linear regression context allowing to reduce the number of studied regressors and develop parsimonious models (Bernard *et al.*, 2018). This method allows to improve the quality of prediction and to produce more precise interpretations (Tibshirani, 1996),

That is why we will adopt it in the present paper. In particular, in two recent based-EMD studies, LASSO regression was used to select the most significant IMFs (Qin *et al.*, 2016, Masselot *et al.*, 2018).

The objective of the present study is to investigate the potential of EMD regression as a tool for predicting river water temperature by applying it to several rivers in North America. The EMD-R will be compared to two other regression models often used in water temperature modelling namely GAM and Logistic regression models. The remainder of this article is organized in the following manner. In section 2, we present the study areas and the methods. In section 3, the EMD-R, GAM and the Logistic Model are applied to model water temperature using air temperatures as a predictor in four different locations. Finally, section 4 provides a discussion and conclusion.

4.2 Materials and Methods

4.2.1 Study Area

In this study, two rivers in Canada and two others in the United States are considered, with relatively long (> 14 years) water temperature time series and very different drainage areas. Daily time series of mean water temperature and mean air temperature were recorded at hydrometric and meteorological stations for different periods from one station to another. In addition, since there are many missing values before May and after October, the off-winter study period with continuous data will be different from one river to another.

For the Missouri River and Potomac River, seasonally mean daily temperature data were obtained from the United States Geological Survey (USGS 2017), and for the two rivers in Canada, Trinity River and Catamaran Brook, the data were retrieved from the Rivtemp database (http://rivtemp.ca/rivtemp-data/) (Figure 4.1).

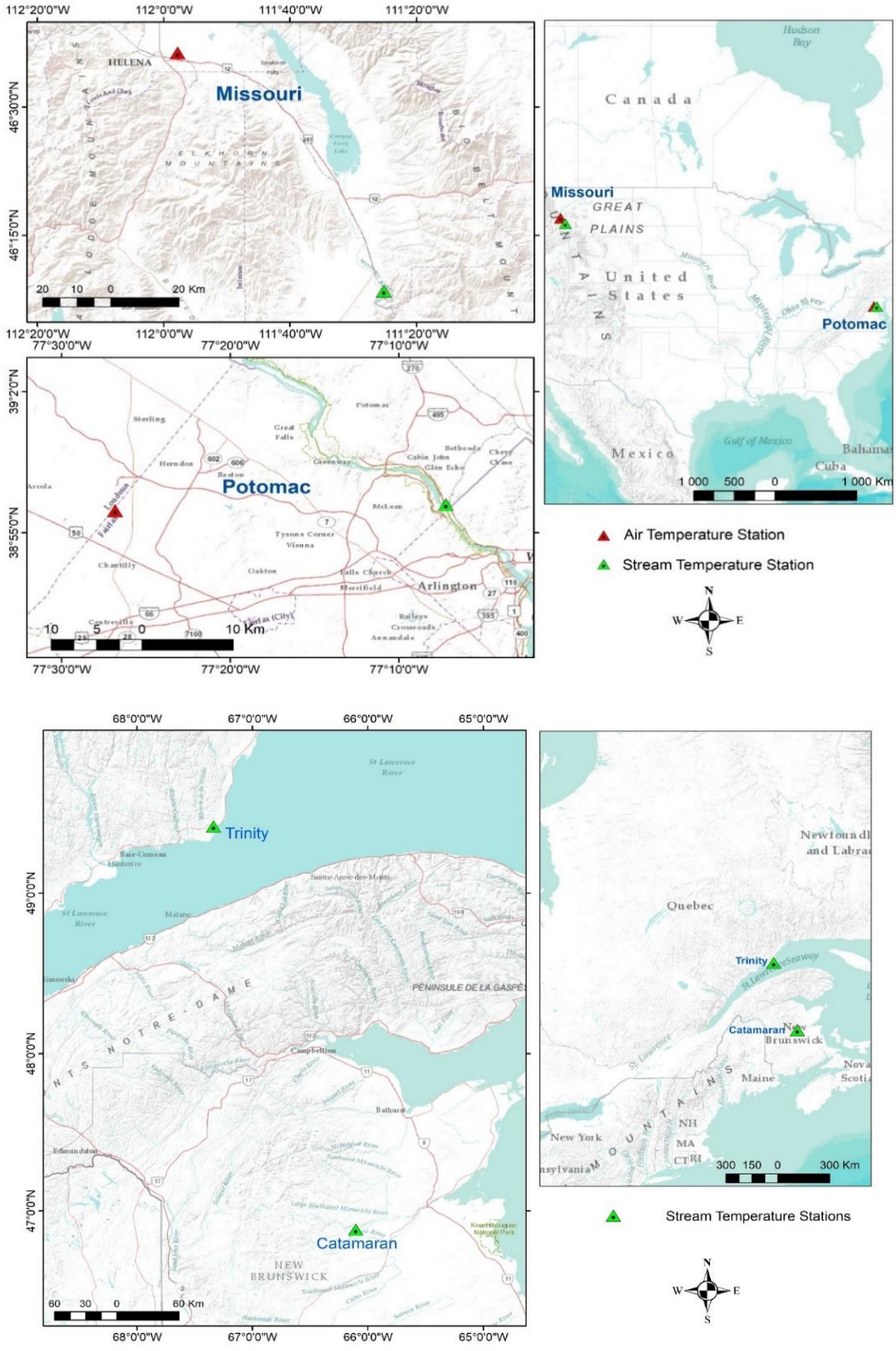


Figure 4.1: Geographic locations of hydrometric and meteorological stations

Catamaran Brook is located in central New Brunswick (Canada) and is a tributary of the Little Southwest Miramichi River, with a latitude of 46 52.7' N and a longitude of 66 06.0' W. This stream has a drainage area of $51\ km^2$. According to fishing surveys, Atlantic salmon is the most common species in Catamaran Brook (Cunjak *et al.*, 1990). Water and air temperature data were recorded from May 1993 to September 2010.

The Trinity River is located near the municipality of Baie-Trinity, 95 km east of Baie-Comeau in the province of Quebec (Canada). The drainage area is $562\ km^2$. Temperature time series are available from May 1985 to October 2017. The Missouri River is the main tributary of the Mississippi River, which flows more than 3600 km from Three Forks at Montana to St.Louis, Missouri (Zhu *et al.*, 2018). The available and complete water and air temperatures time series are from May 2001 to July 2015. The Potomac River is a river in the eastern United States, its length is $655km$. It originates at an altitude of $933m$ southwest of the State of Maryland. Its catchment area is $38{,}018km^2$ and its average annual flow is estimated at $306\ m^3/s$. The data are available for the study from June 2001 to September 2015.

Details about the different meteorological and hydrometric stations of the studied rivers are shown in Table 1.

Tableau 4.1 Detailed information about the four cases studied

River name	Missouri	Potomac
Watershed area	1376180 km²	38018 km²
Length	3600 km	655 km
Name of the weather station	Helena Airport Asos, MT USW 00024144	Washington Dulles international airoport, VA US USW 00093738
Latitude	46.6056	38.9349
Longitude	-111.9636	-77.4473
Name of the hydrometric station	Missouri River at Toston ,MT USGS 06054500	Potomac River Near Wash, DC USGS 01646500
Latitude	46.14	38.94

Longitude	-111.42	-77.12
Water and air temperature period	**01/05/2001-17/07/2015**	**07/06/2001-30/09/2015**

River name	Catamaran	Trinity
Watershed area	51 km^2	562 km^2
Length	20,5 km	75 km
Name of station	Catamaran Brk	Trinité
Latitude	46,878268	49,410555
Longitude	-66,105565	-67,336944
Water and air temperature period	**1993/05/01-2010/09/27**	**1985/05/17-2017/10/16**

4.2.2 Methods

Two statistical models that describe the relationship between water temperature and air temperature will be compared to the proposed EMD-R model. These models are the Generalized Additive Models and a Logistic function.

4.2.2.1 Empirical mode decomposition regression (EMD-R)

The EMD-R consists of two main steps, as shown in Figure 4.2. First, the Empirical Mode Decomposition (EMD) is applied to the explanatory variable, to obtain oscillatory components (the so-called IMFs) and a non-oscillatory residual part. Second, IMFs are considered as new explanatory variables in the Least Absolute Shrinkage and Selection Operator (LASSO) to select the IMFs that are the best predictors (Tibshirani, 1996, Qin *et al.*, 2016).

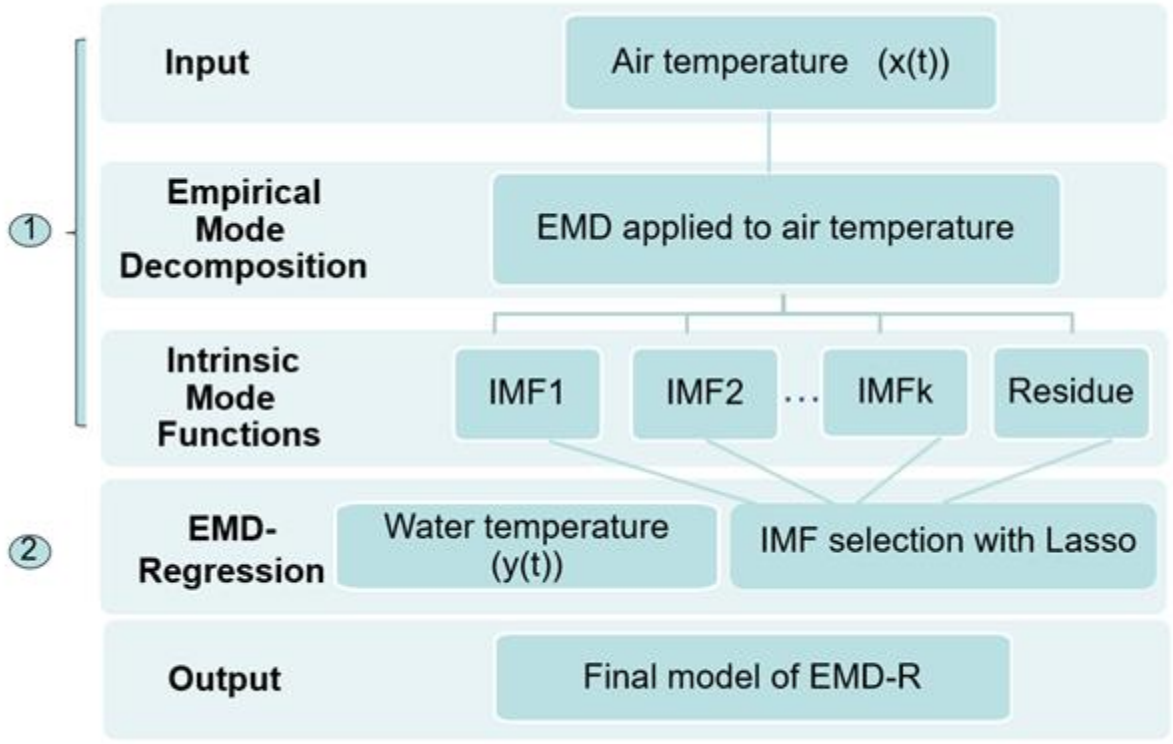


Figure 4.2 : Illustration of the EMD-R method

EMD and the Sifting Process

As expressed by Equation (4.1) below, the EMD algorithm assumes that the original time series $x(t)$ can be decomposed into several sub-series ($IMF_k(\mathrm{t})$), in addition to a non-oscillating residue $r_K(t)$ (Huang *et al.*, 1998b):

$$x(t) = \sum_{k=1}^{K} IMF_k(t) + r_K(t),\ t = 1,2,\dots T \qquad (4.1)$$

The IMFs should satisfy two main conditions: (i) have a null local average at any time point t; (ii) the number of extrema and the number of zero-crossings must either be equal or differ at most by one (Huang *et al.*, 1998a, Boudraa *et al.*, 2007, Huang *et al.*, 2008, Lee *et al.*, 2012). The IMFs are iteratively obtained using the following approach (Huang *et al.*, 1998a):

a) Identify local maxima and minima of $x(t)$ and respectively interpolate them to generate upper and lower envelopes $x_{\max}(t)$ and $x_{min}(t)$.
b) Calculate the local average $m(t) = (x_{\max}(t) + x_{min}(t))/2$
c) Retrieve $m(t)$ from $x(t)$ to obtain the prototype $h(t) = x(t) - m(t)$.

If $h(t)$ fulfills the two abovementioned conditions of IMF, then $h(t)$ is $IMF_1(t)$. If not, iterate steps a to c on $h(t)$ until it satisfies the conditions of an IMF. $h(t)$ is then the first IMF $IMF_1(t)$.

d) Repeat the previous sifting procedure on the residue $r_1(t) = x(t) - IMF_1(t)$ until the obtained residue contains at most one extremum. The final residue is then considered as an estimate of the time series' trend.

A recognized shortcoming of the EMD algorithm is mode-mixing. It appears when there are many different frequencies in the same IMF or when a frequency is shared between two IMFs. This issue is addressed through the Ensemble EMD (EEMD) which consists in adding white noise to $x(t)$ in order to populate its frequencies before decomposition. This is repeated a large number of times to obtain the average of all the computed noisy IMF sets (Zhang *et al.*, 2010, Wang *et al.*, 2018).

Although the EEMD solves the mode mixing problem, it is very important to choose the appropriate number of repetitions and the standard deviation of added white noise. The choice of these parameters affects the quality of decomposition and its results (Zhang *et al.*, 2010).

LASSO Regression

Proposed by Tibshirani (1996), the LASSO Regression is a shrinkage estimation method. For a regression model, the basic LASSO principle is to estimate the regression coefficients β by minimizing the expression of the following penalized least squares:

$$\hat{\beta} = arg_{\beta} min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j X_j \right) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad 4.2)$$

where, y is the response variable, X_j $(j = 1, \ldots, p)$ is the explanatory variables and λ is the penalty coefficient. The higher this coefficient, the stronger the regularization. This regulation parameter directly controls the number of explanatory variables left in the final model. The value of λ is usually estimated by cross-validation.

The main advantage of using LASSO over other regression methods is that it allows for a selection of variables to be made by cancelling some regression coefficients (Tibshirani, 2011, Qin *et al.*, 2016, Chu *et al.*, 2018). The predictors selected by LASSO will form the final prediction model of EMD-R.

4.2.2.2 Generalized additive model (GAM)

The generalized additive model (GAM) is a nonlinear model with an additive predictor structure. This approach, which was defined by Hastie and Tibshirani (1986), allows for a wide flexibility in representing nonlinear associations while retaining interpretative power through its additive structure (Chebana *et al.*, 2014, Iddrisu *et al.*, 2017, Wood, 2017, Rahman *et al.*, 2018). GAM can be expressed through the equation:

$$\boldsymbol{g}\big(\boldsymbol{E}(\boldsymbol{y})\big) = \boldsymbol{f_1}(\boldsymbol{x_1}) + \boldsymbol{f_2}(\boldsymbol{x_2}) + \cdots + \boldsymbol{f_p}\big(\boldsymbol{x_p}\big) + \boldsymbol{\varepsilon} \qquad (4.3)$$

where, y is the response variable, x_i ($i = 1, \dots, p$) are explanatory variables, g is the link function allowing for extension of the Gaussian distribution to the exponential family, $E(y)$ is the expected value of the response variable , f_i is the associated smooth nonlinear function and ε is the error assumed to be normally distributed with variance σ_ε

The GAM application is based on the estimation of the smoothing functions $f_i(x_i)$. The method is implemented in the mgcv package for the R software (Wood, 2006, Wood, 2017).

4.2.2.3 Logistic model (Sigmoid)

The logistic regression model is a non-linear function often used to model river water temperature as a function of air temperature. This regression function is expressed using three parameters as follows,

$$y = \frac{\alpha}{1+e^{\gamma(\beta-x)}} \qquad (4.4)$$

where y and x represent the water and air temperatures respectively, α is the maximum water temperature estimation coefficient; β is the value of the air temperature at the inflection point and γ represents the steepest slope of the logistics function (Equation 4.4). These parameters are estimated by minimizing the sum of quadratic errors (Omid Mohseni *et al.*, 1998b, Salter *et al.*, 2000).

4.2.3 Model Evaluation

In this study, four performance criteria are used to assess the predictive power of the different approaches, namely the coefficient of determination (R^2) (Zhu *et al.*, 2018), the root of the mean square error (RMSE) (Ahmadi-Nedushan *et al.*, 2007), the bias (B) (St-Hilaire *et al.*, 2012) and

the generalized cross-validation (GCV) (Tibshirani, 1996, Laanaya *et al.*, 2017). These criteria are given respectively by the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \qquad (4.5)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{P_i} - \mathrm{O_i})^2} \qquad (4.6)$$

$$B = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i) \qquad (4.7)$$

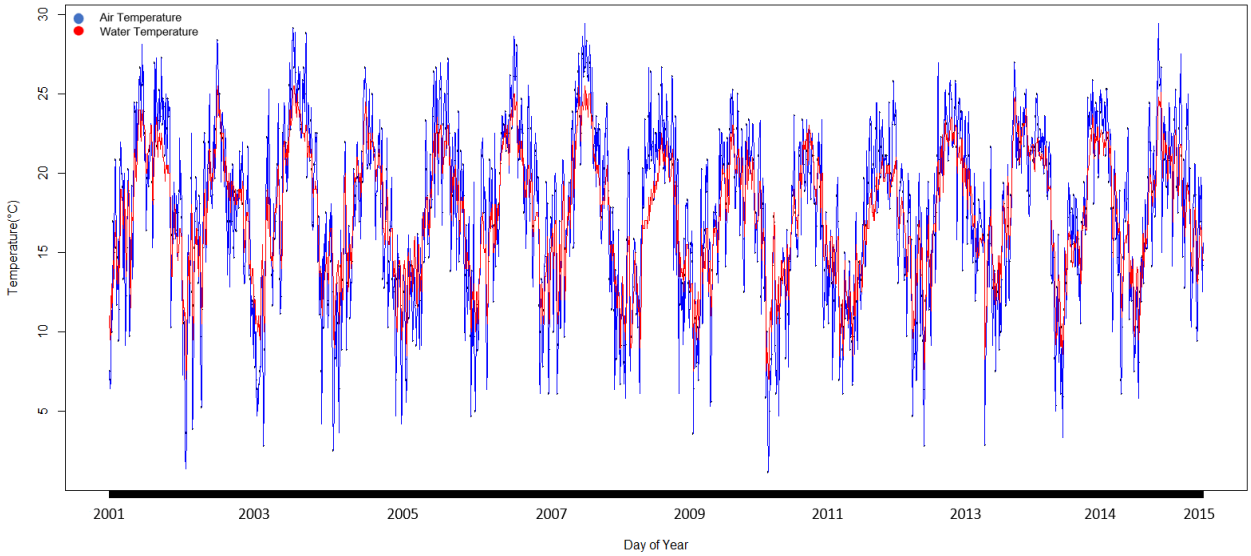
$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(O_i - P_i)}{1 - trace(S)/n}\right]^2 \qquad (4.8)$$

where n is the size of the series studied, O_i is the observed value, P_i is the predicted value, $\bar{O}$ is the average of the original series and trace (S) is the effective number of parameters (Golub *et al.*, 1979).

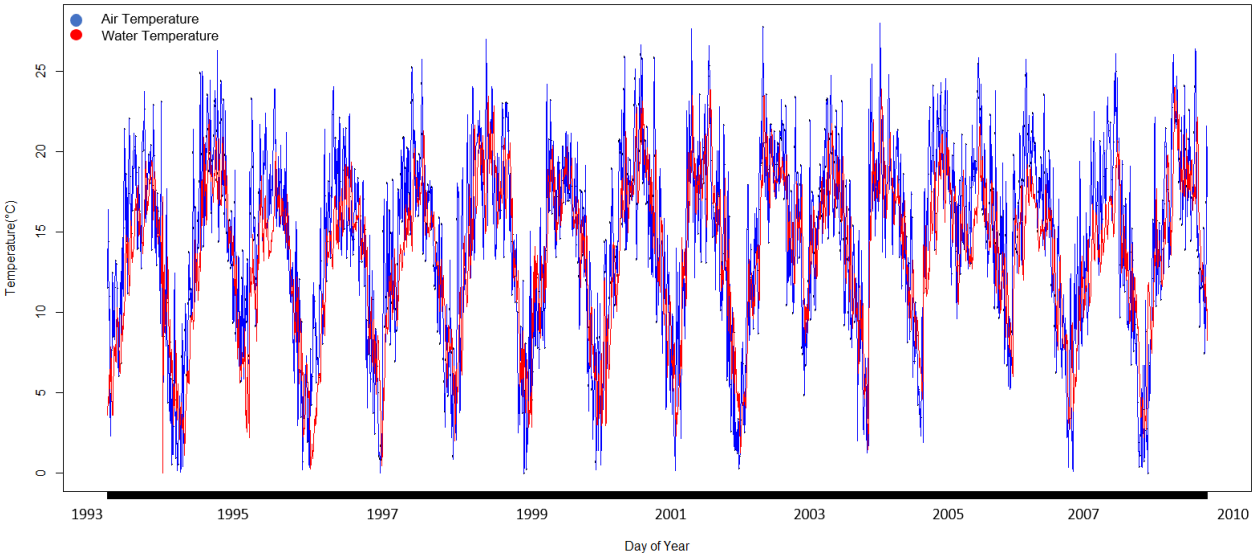
4.3 Results and Interpretation

The daily average water temperature in the river (y) for the four case studies, described in Section 2.1 is hereby modelled using the EMD-R, GAM and Logistic models with air temperature as the input. The parameters of these respective models were estimated using the formulas defined in Section 2.2. The results obtained for two case studies, namely the Missouri River in the United States and Catamaran Brook in Canada, are presented with more details. Results for the Trinity River and Potomac River are similar and are therefore not presented in details (appendix).

According to Figure 4.3, the ranges of variation in air temperature are more pronounced than those of water temperature. The original air temperature data sets for the Missouri River and Catamaran Brook respectively, is characterized by several components at different frequencies. It reveals the presence of a strong seasonality. The amplitudes of the seasonal cycle of air temperature are relatively well synchronized with those of water temperature for the Missouri River and Catamaran Brook.



a) Missouri River Station



b) Catamaran Brook Station

Figure 4.3 : Average daily water and air temperature in Missouri River and Catamaran Brook

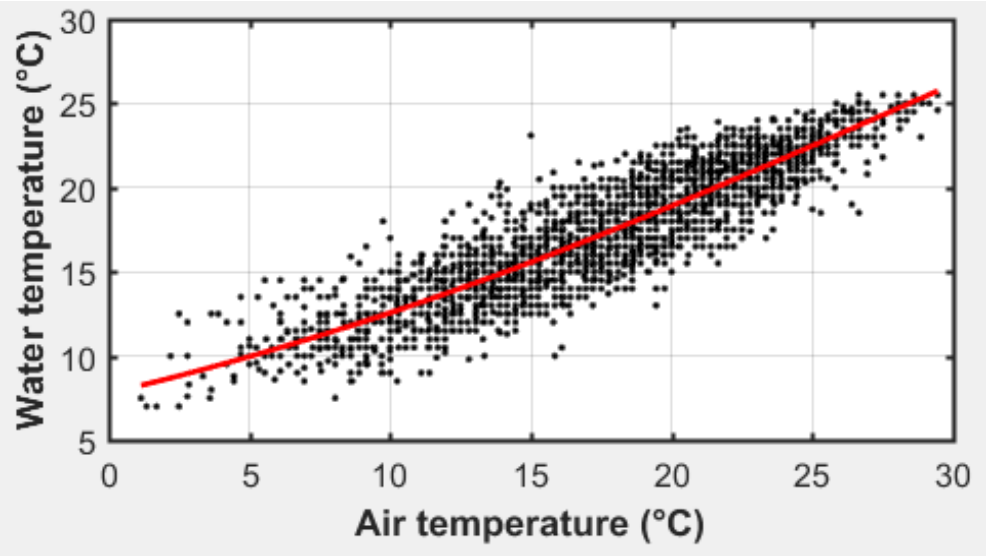
LOGISTIC MODEL RESULTS

Figure 4.4 shows the fitted logistic regression between water and air temperature and the fitted functions described below. There is a strong dispersion between daily average water and air temperature. The application of the sigmoid model gave total explained variances equal to 80.39% (highest of all stations) and 55.30% (lowest) for Missouri River and Catamaran Brook

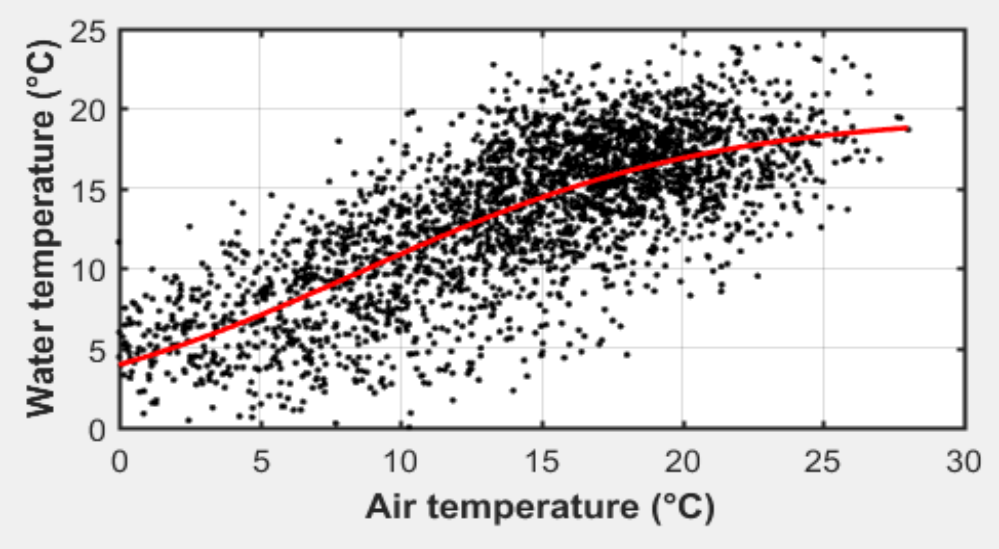
respectively. The resulting model equation for the Missouri River and Catamaran Brook are respectively:

$$y = \frac{48.51}{1+e^{0.06(27.42-Ta)}} \quad (4.9)$$

$$y = \frac{19.63}{1+e^{0.16(8.64-Ta)}} \quad (4.10)$$



a) Missouri River



b) Catamaran Brook

Figure 4.4 : Relationship between daily water and air temperature in (a) Missouri River and (b) Catamaran Brook and a fitted logistic function

GAM RESULTS

The smooth effects of air temperature on water temperature are shown in Figure 4.5. For Catamaran Brook, the estimated relation between air and water temperature is clearly nonlinear with an S-shape, especially between 12.5 °C and 22.5 °C (Figure 4.5b). At extreme values of air temperatures, the smooth effects flatten. On the other hand, for the Missouri River, the smooth effects graph shows a nearly linear relationship between air and water temperature (Figure 4.5a). The analytical results of GAM for the two case studies mentioned in Table 2 show the non-linearity effects of air temperature with a probability-value less than 0.0001. The latter shows that the non-linear component is not negligible.

We notice that the air temperature smoothing function for the GAM in the case of Catamaran Brook is very close to that of the sigmoid model (Figure 4.4 and Figure 4.5).

Tableau 4.2 : GAM results for a) Missouri River, b) Catamaran Brook, c) Trinity River and d) Potomac River

a)

Smoothing functions	Estimated degrees of freedom	Fisher test	P-value
S (Air Temperature)	3.31	725	$< 2.10^{-16}$
R^2-adj	0.805		
Deviance explained	89%		
GCV	1.686		

b)

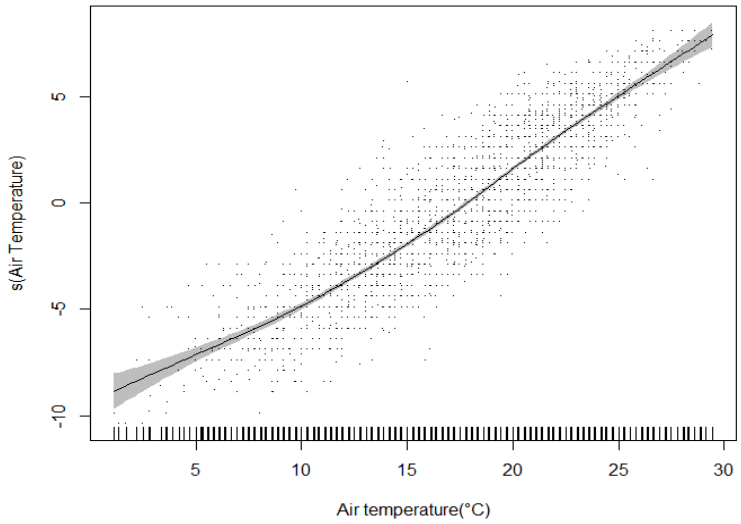
Smoothing functions	Estimated degrees of freedom	Fisher test	P-value
S (Air Temperature)	5.03	557	$< 2.10^{-16}$
R^2-adj	0.56		
Deviance explained	55.8%		
GCV	10.31		

c)

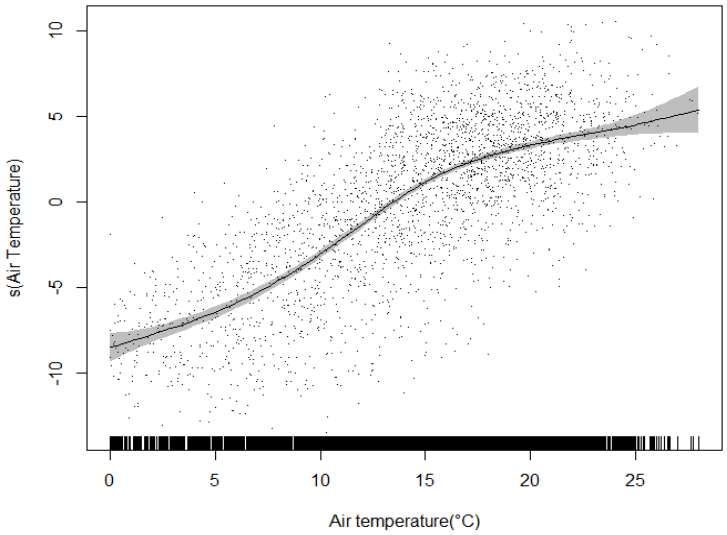
Smoothing functions	Estimated degrees of freedom	Fisher test	P-value
S (Air Temperature)	6.33	1600	$< 2.10^{-16}$
R^2-adj	0.752		
Deviance explained	75.2%		
GCV	7.198		

d)

Smoothing functions	Estimated degrees of freedom	Fisher test	P-value
S (Air Temperature)	5.567	255.243	$< 2.10^{-16}$
R^2-adj	0.639		
Deviance explained	64%		
GCV	6.525		



a) Missouri Station



b) Catamaran Station

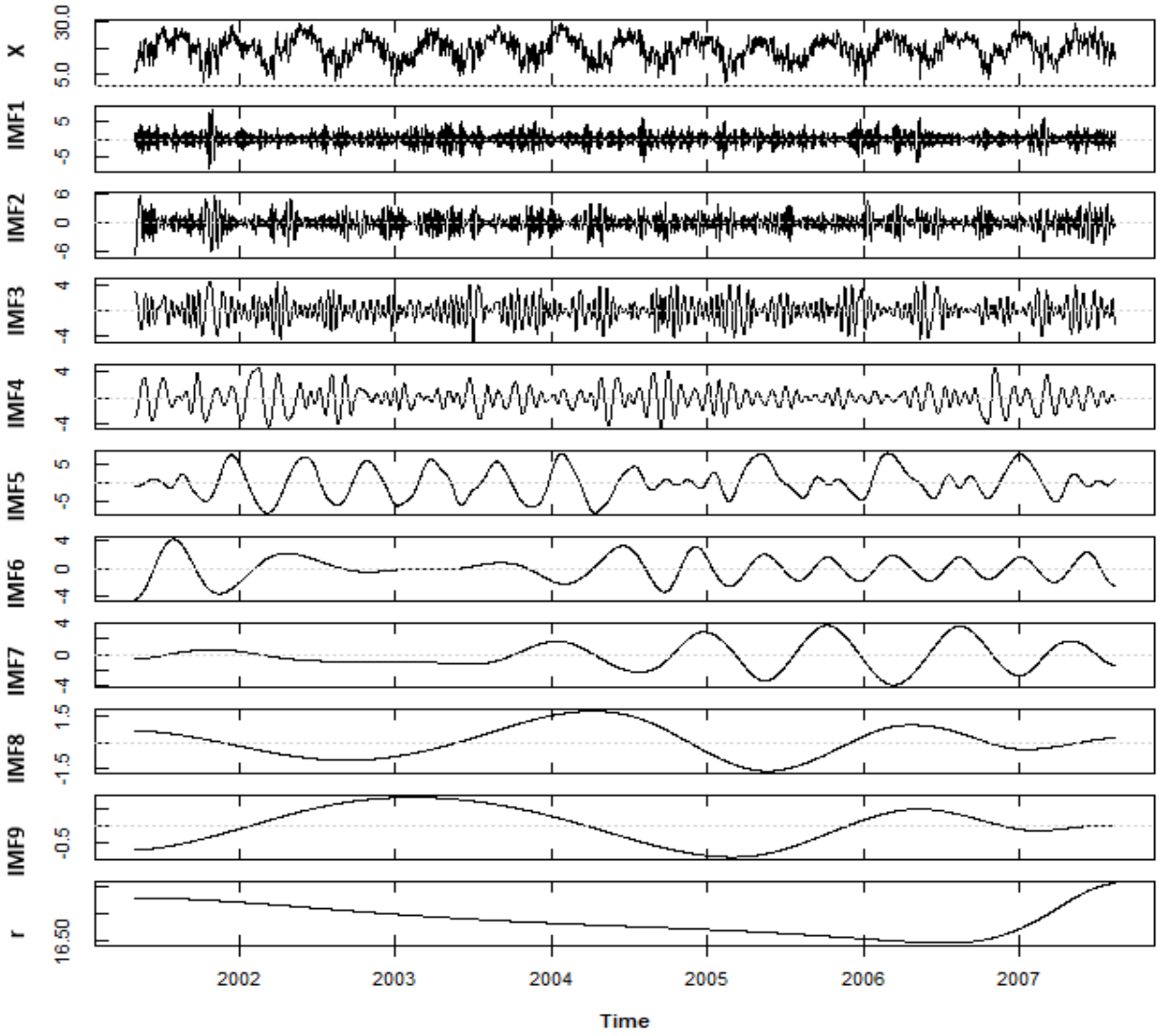
Figure 4.5 : Estimated smooth effect functions for a) the Missouri River & b) Catamaran Brook for the air temperature

EMD-R

The application of the EMD-R method (Figure 4.6) illustrates the decomposition of air temperature using the traditional EMD decomposition method. The components represented in Figure 4.6 are not clearly separated from each other and the low frequency IMFs are mixed together. This indicates the presence of mixed modes.

For our studied cases (Figure 4.7), we apply the EEMD, this version is developed to solve the problem of mode mixing (Abdoulaye Thioune, 2015b). In this article, the parameters of EEMD are chosen with reference to the previous work (Rilling *et al.*, 2003, Rehman *et al.*, 2013). Several combinations of the parameters were tested, each time checking the mode mixing problem. Finally, a single white noise with a variance of 10% as recommended was chosen. While for the number of sets, the largest possible value Ne=1000 was chosen. The two-original series (Missouri and Catamaran) were broken down to reveal 17 IMFs components for Missouri River and 16 components for Catamaran Brook with a residual component as in Figure 4.7. In the latter, it can be seen that the frequency of each IMF for the two case studies is indeed regular, but within each IMF, the amplitude is variable.

The decomposition result shows a general separation of the data into locally non-overlapping time scale components. This shows that the EEMD results give components that can be interpreted. According to Figure 4.7, we notice that for the Missouri River case, we can sum the IMF6, IMF7 and the IMF11 to IMF17 (noted IMF11:17 in Figure 4.7a) since they have the same frequency. The same for the Catamaran case, we can sum IMF6, IMF7 and the IMF11 to IMF 16 (noted IMF11:16 in Figure 4.7b), finally obtaining 10 IMFs components for the Missouri case and the Catamaran case.


X
IMF1
IMF2
IMF3
IMF4
IMF5
IMF6
IMF7
IMF8
IMF9
r
30.0
5.0
5
-5
6
0
-6
4
-4
4
-4
5
-5
4
0
-4
4
0
-4
1.5
-1.5
-0.5
16.50
2002
2003
2004
2005
2006
2007
Time

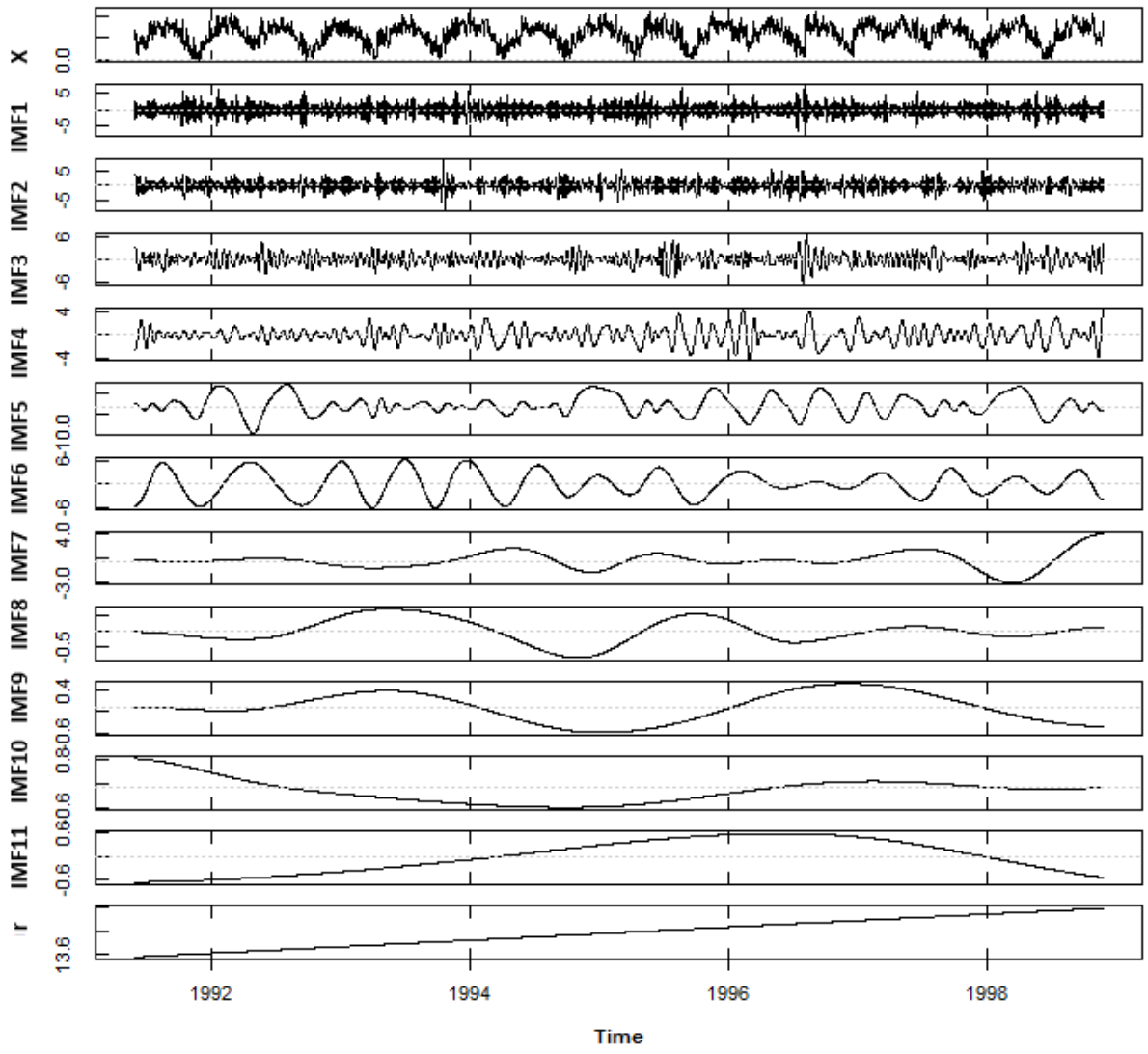
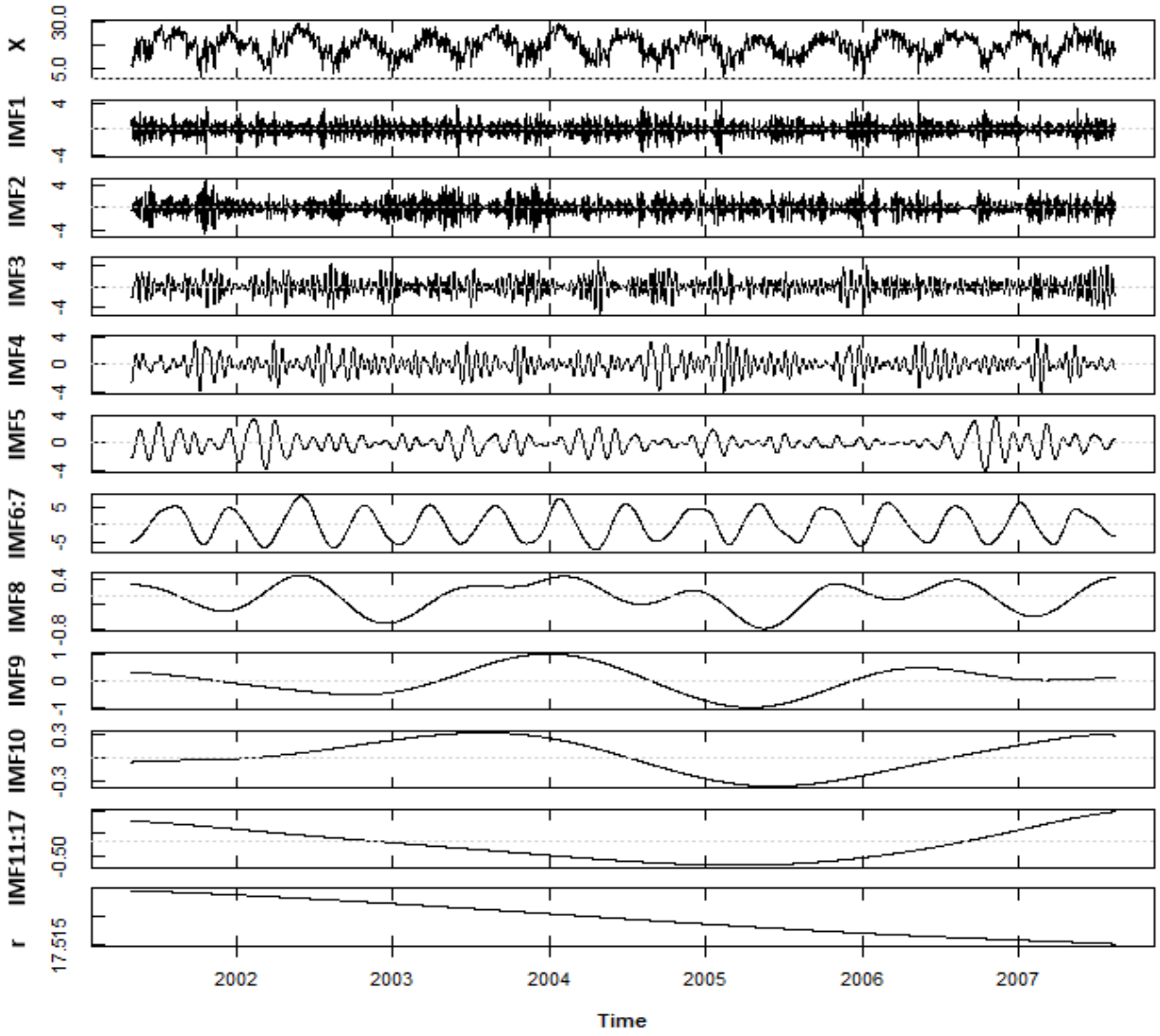


Figure 4.6 : Decomposed air temperature series with the EMD algorithm (Missouri top & Catamaran bottom)


X
IMF1
IMF2
IMF3
IMF4
IMF5
IMF6:7
IMF8
IMF9
IMF10
IMF11:17
r
30.0
5.0
4
-4
4
-4
4
-4
4
0
-4
4
0
-4
5
-5
0.4
-0.8
1
0
-1
0.3
-0.3
-0.50
17.515
2002
2003
2004
2005
2006
2007
Time

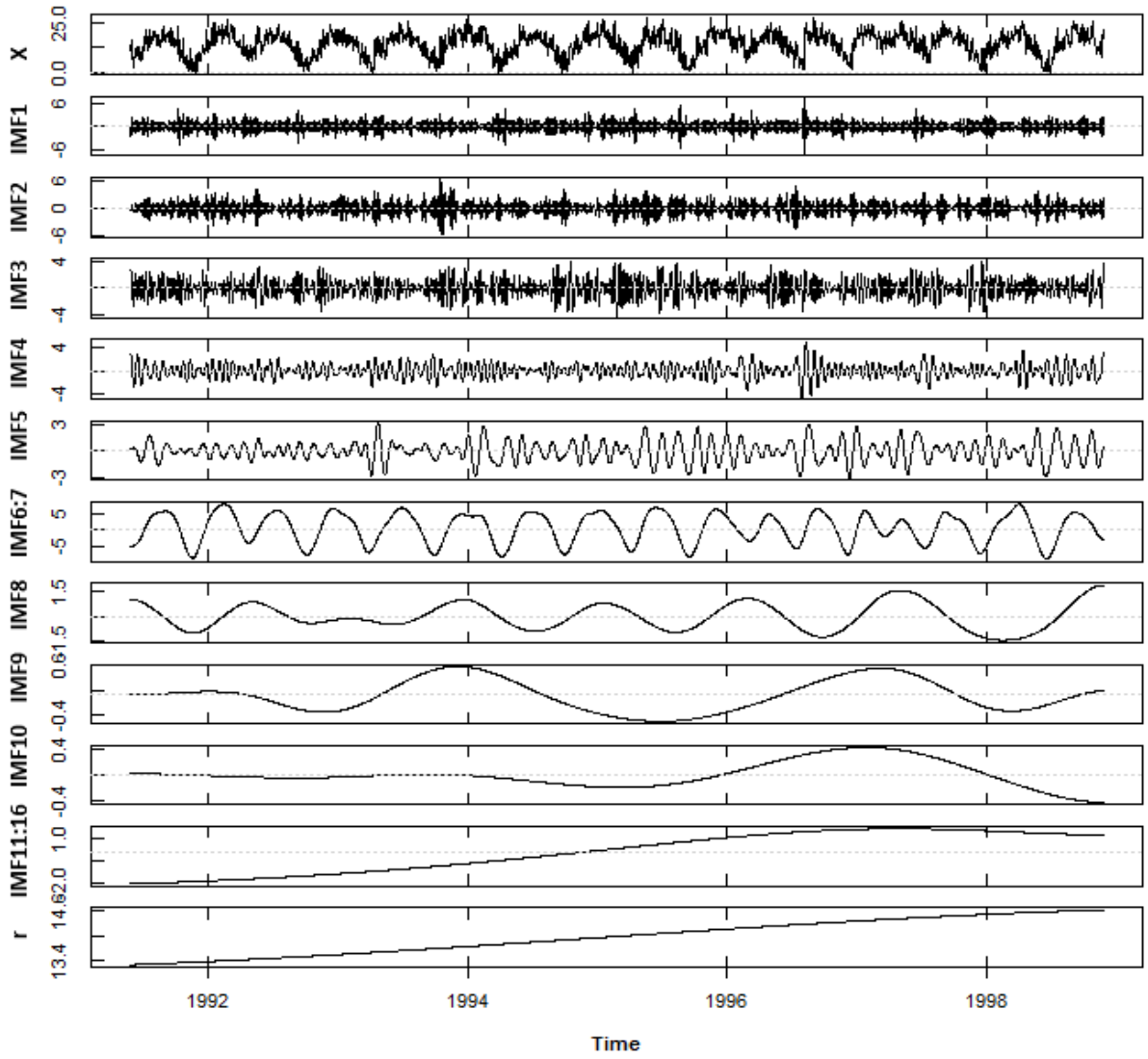


Figure 4.7 : Decomposed air temperature series with the EEMD algorithm a) Missouri top & b) Catamaran bottom

According to Tableau 4.3, it can be seen that for the two case studies, the IMF1 and IMF2 components show quasi-regular peaks with an average period between 3 and 6 days, with an average amplitude varying between 2°C and 3°C for Missouri and between 2.5°C and 3.5°C for Catamaran Creek. These high frequency random oscillations may be related to the hot periods of the summer season when the air temperature records high values. IMF3 and IMF4 have an average period between one and three weeks with an amplitude close to that of the first two components. The IMF5 component has an average period of about 40 days, with a relatively small amplitude compared to the first components.

The IMF6 and IMF7 components are biannual components with an average period of about 6 months and a higher amplitude than the previous components. The causes of these day-length cycles have been attributed to the semi-annual and annual cycles of the atmospheric circulation. The other components represent interannual variations. IMF8 is quasi-biannual, and IMF9 has a mean period slightly longer than three years. For the last two components (IMF10 and IMF11:17 or IMF11:16, the period exceeds three years where the range for IMF10 is around 5°C and for IMF11:16 or IMF11:17 varies between 1.5°C and 2.5°C for both case studies.

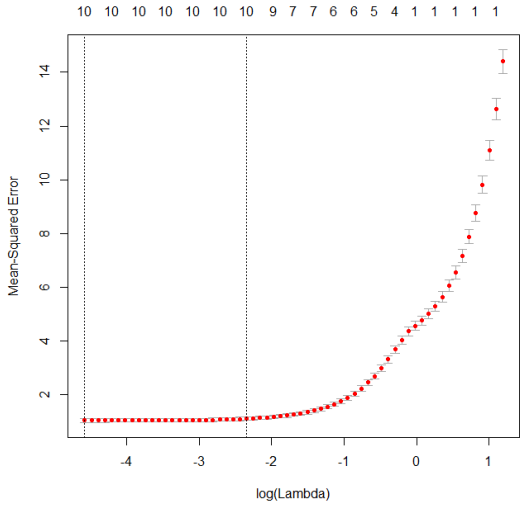
Tableau 4.3 : Mean Period, Mean Amplitude and regression coefficients of Missouri River and Catamaran Brook

	Mean period (day)		Mean amplitude (°C)		Regression coefficients	
	Missouri study	Catamaran study	Missouri study	Catamaran study	Missouri study	Catamaran study
IMF1	3.05	2.98	2.06	2.51	0.007	0
IMF2	5.98	5.74	3.08	3.02	0.282	-0.106
IMF3	11.28	10.72	3.57	3.36	0.403	0.413
IMF4	21.05	20.41	3.10	2.90	0.463	0.486
IMF5	39.91	41.48	2.84	2.76	0.439	0.600
IMF6+7	152.96	162.00	11.44	12.44	0.807	0.868
IMF8	379.27	458.72	0.68	2.03	0.449	0.858
IMF9	989.33	942.80	1.20	0.84	-0.155	-0.858
IMF10	1625.00	1171.50	0.54	0.43	-1.889	0
IMF11:17*	2798.00	-	1.60	-	0.233	
IMF11:16*	-	1128.50	-	2.48	-	0.460

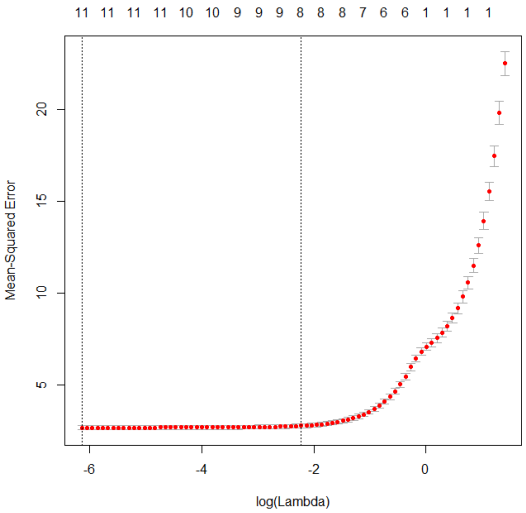
*IMFn:m indicates the summation of IMFs from n to m

Figure 4.8 shows a plot of the Mean Squared Error (MSE) for different values of λ. As the λ value increases, the regression coefficients decrease to zero and the MSE becomes higher, indicating that predictive power of the model is poor. Whereas, as λ decreases, the regression coefficients do not reach zero and the plot appears to flatten. The model having low MSE associated with

the smallest λ (i.e. 0.079 for Missouri River and 0.097 for Catamaran Brook) is identified in Figure 4.8.



a) Missouri Station



b) Catamaran Station

Figure 4.8 : Adjusted validation of a) Missouri & b) Catamaran cases

The red dots are MSE, the vertical lines represent the value for λ selected according to the MSE method and the horizontal axis at the top represents the number of IMFs remaining in the model for the appropriate value of λ. For the Missouri River, the LASSO retaining all the IMFs during the decomposition by giving each a regression coefficient. Whereas in the case of Catamaran Brook, LASSO gave zero for the IMF1 and IMF10, retaining only 8 among the 10 obtained. We note that the IMF6+7 component recorded the highest regression coefficient for the two case studies, which

shows the effect of this component on our regression model obtained. On the other hand, the IMF1 and IMF2 components obtained respectively the lowest regression coefficients in the case of Missouri River and Catamaran Brook, these components have a less important effect than the other IMFs.

4.4 Comparative study and discussion

The logistic model that describes the relationship between water temperature and river air temperature has R^2 = 78.32 % and R^2 = 75.05 % respectively for the Potomac River and Trinity River. Generally, the Logistic model leads to poorer results, with RMSE ranging from 1.72 °C to 3.22°C and GCV coefficient values ranging from 2.96 to 10.37. Indeed, these relatively weak performances may be caused by the fact that this model is deemed better adapted fort weekly time steps (Benyahya et al., 2007), although it has been applied for daily mean water temperatures in the past (e.g. Laanaya et al., 2007).

The application of the GAM resulted respectively in a RMSE of 1.71 °C and 3.20 °C, GCV of 2.95 and 10.31, and a R^2 of 80.5% and 55.8%, for the Missouri River and Catamaran Brook respectively (Table 3). EMD-R performance indicators are presented in Table 4. This model has relatively high coefficients of determination, with R-squared = 92.86 % for Missouri River and R-squared greater than 67% for other case studies.

The performance of the EMD-R, GAM and Logistic models for the four case studies are presented in Table 3. Broadly, the EMD-R performs better than the other models. The EMD-R, R^2 is the highest with explained variance between 87.58% for the Trinity River and 91.41% for the Missouri River. In comparison, the lowest and highest determination coefficients of logistic regression and GAM are respectively around 55% for Catamaran Brook and 80% for Missouri River.

The RMSE criterion, indicates a best performance of the EMD-R with values ranging from 1.01 ° C to 2.38 ° C for the four case studies. We can note that the RMSE values obtained for the GAM and Logistics models are very close but with a slight better result for the GAM. For GCV, EMD-R is again the most performant model for the four case studies with a value of 1.03 for the Missouri River and 5.69 for the Potomac River. While for other cases of comparison, the GCV are very close but the GAM is still better than the Logistic model. For the bias criterion, it is the GAM and Logistic that gave the values closest to zero, but it is justified by the fact that the use of LASSO biases the regression.

Tableau 4.4 : Performance coefficients of the predictive accuracy

Case studies	Model	Coefficient of determination (R^2) (%)	GCV	RMSE (°C)	Biais (°C)
Missouri	EMD-R	**92.86**	**1.03**	**1.01**	-0.41
	GAM	80.50	2.95	1.71	**$-4.14.10^{-14}$**
	Logistic	80.39	2.96	1.72	$-8.24.10^{-4}$
Catamaran	EMD-R	**88.95**	**2.63**	**1.57**	-0.03
	GAM	55.80	10.31	3.20	**$-3.14.10^{-14}$**
	Logistic	55.30	10.37	3.22	-0.012
Trinity	EMD-R	**90.40**	**3.07**	**1.75**	-0.452
	GAM	75.2	7.19	2.67	**$7.70.\ 10^{-15}$**
	Logistic	75.05	7.21	2.68	-0.0019
Potomac	EMD-R	**67.69**	**5.69**	**2.38**	-0.036
	GAM	62.60	6.11	2.47	**$2.69.\ 10^{-11}$**
	Logistic	78.32	6.41	2.53	$-1.62.\ 10^{-5}$

* The bold character indicates the best performance

4.5 Conclusion

The main objective of this paper was to model the daily mean water temperature in four rivers using the average air temperature. We propose to compare a new method, EMD-R to other commonly used methods (GAM and Sigmoid). The EMD-R, GAM and Logistics models were tested using the following performance criteria: R-square, RMSE, GCV and Bias. The EMD-R showed a predictive performance superior to that of GAM and the logistic model in terms of R-square, GCV and RMSE. The EMD-R offers the possibility of exploiting components of the air temperature signal at different frequencies, while maintaining the advantages of non-parametric approaches (e.g. no definition of functions a priori or distributions; no imposition of stationarity). Future work should include studying the potential of EMD-R at sub-daily time steps. As well as with more than two variables, where each variable has a more complex structure that requires more sophisticated and advanced methods. These methods make it possible to describe the actual relationships between the different variables, which are often non-linear.

5 REFFERENCES

Ahmadi-Nedushan B, St-Hilaire A, Ouarda TB, Bilodeau L, Robichaud E, Thiémonge N & Bobée B (2007) Predicting river water temperatures using stochastic models: case study of the Moisie River (Québec, Canada). *Hydrological Processes: An International Journal* 21(1):21-34.

Allen A, Gillooly J & Brown J (2005) Linking the global carbon cycle to individual metabolism. *Functional Ecology* 19(2):202-213.

Bartholow JM, Campbell SG & Flug M (2004) Predicting the thermal effects of dam removal on the Klamath River. *Environmental management* 34(6):856-874.

Beaufort A, Moatar F, Curie F, Ducharne A, Bustillo V & Thiéry D (2016) River temperature modelling by Strahler order at the regional scale in the Loire River basin, France. *River Res Appl* 32(4):597-609.

Bélanger M, El-Jabi N, Caissie D, Ashkar F & Ribi J (2005) Estimation de la température de l'eau de rivière en utilisant les réseaux de neurones et la régression linéaire multiple. *Revue des sciences de l'eau/Journal of Water Science* 18(3):403-421.

Benyahya (2007) *Modélisation statistique de la température de l'eau en rivière et en régime non-hivernal*. (Thèse présentée pour l'obtention du grade de Philosophae Doctor (Ph. D) en …).

Benyahya, Caissie D, St-Hilaire A, Ouarda TB & Bobée B (2007a) A review of statistical water temperature models. *Canadian Water Resources Journal* 32(3):179-192.

Benyahya, St-Hilaire A, Ouarda TBMJ, BobÉE B & Dumas J (2010) Comparison of non-parametric and parametric water temperature models on the Nivelle River, France. *Hydrological Sciences Journal* 53(3):640-655.

Benyahya, St-Hilaire A, Quarda TBMJ, Bobée B & Ahmadi-Nedushan B (2007b) Modeling of water temperatures based on stochastic approaches: case study of the Deschutes River. *J Environ Eng Sci* 6(4):437-448.

Bernard N & Ahmed F (2018) Le LASSO.

Beschta RL, Bilby RE, Brown GW, Holtby LB & Hofstra TD (1987) Stream temperature and aquatic habitat: fisheries and forestry interactions.

Boudraa A-O & Cexus J-C (2007) EMD-based signal filtering. *IEEE transactions on instrumentation and measurement* 56(6):2196-2202.

Bovee KD (1982) A guide to stream habitat analysis using the instream flow incremental methodology. *Information paper* 12.

Bunn SE & Arthington AH (2002) Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. *Environmental management* 30(4):492-507.

Caissie (2006) The thermal regime of rivers: a review. *Freshwater Biol* 51(8):1389-1406.

Caissie, El-Jabi N & St-Hilaire A (1998) Stochastic modelling of water temperatures in a small stream using air to water relations. *Canadian Journal of Civil Engineering* 25(2):250-260.

Caissie, Satish MG & El-Jabi N (2005) Predicting river water temperatures using the equilibrium temperature concept with application on Miramichi River catchments (New Brunswick, Canada). *Hydrological Processes: An International Journal* 19(11):2137-2159.

Caissie., El-Jabi N & Satish MG (2001) Modelling of maximum daily water temperatures in a small stream using air temperatures. *Journal of Hydrology* 251(1-2):14-28.

Chebana F, Charron C, Ouarda TB & Martel B (2014) Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of Hydrometeorology* 15(6):2418-2428.

Chu H, Wei J & Qiu J (2018) Monthly Streamflow Forecasting Using EEMD-Lasso-DBN Method Based on Multi-Scale Predictors Selection. *Water* 10(10):1486.

Cluis (1972) Relationship between stream water temperature and ambient air temperaturea simple autoregressive model for mean daily stream water temperature fluctuations. *Hydrology Research* 3(2):65-71.

Council NR (2004) *Managing the Columbia River: Instream flows, water withdrawals, and salmon survival.* National Academies Press,

Crisp DT & Howson G (1982) Effect of air temperature upon mean water temperature in streams in the north Pennines and English Lake District. *Freshwater Biol* 12(4):359-367.

Cunjak RA, Caissie D & El-Jabi N (1990) *Projet de recherche sur l'habitat du ruisseau Catamaran: description et champs d'etude general.* La Division,

Demars BO, Russell Manson J, Olafsson JS, Gislason GM, Gudmundsdottír R, Woodward G, Reiss J, Pichler DE, Rasmussen JJ & Friberg N (2011) Temperature and the metabolic balance of streams. *Freshwater Biol* 56(6):1106-1121.

Dominici F, McDermott A, Zeger SL & Samet JM (2003) Airborne particulate matter and mortality: timescale effects in four US cities. *Am J Epidemiol* 157(12):1055-1065.

Dupuis AP & Hann BJ (2009) Climate change, diapause termination and zooplankton population dynamics: an experimental and modelling approach. *Freshwater Biol* 54(2):221-235.

Durocher M, Lee TS, Ouarda TB & Chebana F (2016) Hybrid signal detection approach for hydro-meteorological variables combining EMD and cross-wavelet analysis. *Int J Climatol* 36(4):1600-1613.

Edwards R, Densem J & Russell P (1979) An assessment of the importance of temperature as a factor controlling the growth rate of brown trout in streams. *The Journal of Animal Ecology*:501-507.

Erickson TR & Stefan HG (2000) Linear Air/Water Temperature Correlations for Streams during Open Water Periods. *Journal of Hydrologic Engineering* 5(3):317-321.

Fan G-F, Peng L-L, Zhao X & Hong W-C (2017) Applications of Hybrid EMD with PSO and GA for an SVR-Based Load Forecasting Model. *Energies* 10(11):1713.

Ficklin DL, Stewart IT & Maurer EP (2013) Effects of climate change on stream temperature, dissolved oxygen, and sediment concentration in the Sierra Nevada in California. *Water Resources Research* 49(5):2765-2782.

Golub GH, Heath M & Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21(2):215-223.

Greenberg JA, Hestir EL, Riano D, Scheer GJ & Ustin SL (2012) Using LiDAR Data Analysis to Estimate Changes in Insolation Under Large-Scale Riparian Deforestation 1. *JAWRA Journal of the American Water Resources Association* 48(5):939-948.

Grégoire Y, Trencia G & Faune S (2007) Influence de l'ombrage produit par la végétation riveraine sur la température de l'eau.

Gu RR & Li Y (2002) River temperature sensitivity to hydraulic and meteorological parameters. *J Environ Manage* 66(1):43-56.

Guillemette N, St-Hilaire A, Ouarda TBMJ, Bergeron N, Robichaud É & Bilodeau L (2009) Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in a multivariate space. *Journal of Hydrology* 364(1-2):1-12.

Hadzima-Nyarko M, Rabi A & Šperac M (2014) Implementation of Artificial Neural Networks in Modeling the Water-Air Temperature Relationship of the River Drava. *Water Resources Management* 28(5):1379-1394.

Hastie T & Tibshirani R (1986) Generalized additive models Statistical science.

Hedger RD, Sundt-Hansen LE, Forseth T, Ugedal O, Diserud OH, Kvambekk ÅS & Finstad AG (2013) Predicting climate change effects on subarctic–Arctic populations of Atlantic salmon (Salmo salar). *Can J Fish Aquat Sci* 70(2):159-168.

Huang NE, Shen Z & Long SR (1999) A new view of nonlinear water waves: the Hilbert spectrum. *Annual review of fluid mechanics* 31(1):417-457.

Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen N-C, Tung CC & Liu HH (1998a) 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(1971):903-995.

Huang NE, Shen Z, Long SR, Wu MC, Shih HH, Zheng Q, Yen N-C, Tung CC & Liu HH (1998b) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences.* The Royal Society, p 903-995.

Huang NE & Wu Z (2008) A review on Hilbert-Huang transform: Method and its applications to geophysical studies. *Reviews of geophysics* 46(2).

Iddrisu WA, Nokoe KS, Luguterah A & Antwi EO (2017) Generalized Additive Mixed Modelling of River Discharge in the Black Volta River. *Open Journal of Statistics* 7(04):621.

Isaak D, Wollrab S, Horan D & Chandler G (2012) Climate change effects on stream and river temperatures across the northwest US from 1980–2009 and implications for salmonid fishes. *Climatic Change* 113(2):499-524.

Jeong DI, Daigle A & St-Hilaire A (2013) Development of a stochastic water temperature model and projection of future water temperature and extreme events in the Ouelle River basin in Québec, Canada. *River Res Appl* 29(7):805-821.

Johnson & Belk M (2004) Temperate Utah chub form valid otolith annuli in the absence of fluctuating water temperature. *Journal of fish Biology* 65(1):293-298.

Johnson & Jones JA (2000) Stream temperature responses to forest harvest and debris flows in western Cascades, Oregon. *Can J Fish Aquat Sci* 57(S2):30-39.

Karacor AG, Sivri N & Ucan ON (2007) Maximum stream temperature estimation of Degirmendere River using artificial neural network. *J Sci Ind Res India* 66(5):363-366.

KisÌ§i Ãzr (2009) Wavelet regression model as an alternative to neural networks for monthly streamflow forecasting. *Hydrological Processes* 23(25):3583-3597.

Krider LA, Magner JA, Perry J, Vondracek B & Ferrington Jr LC (2013) Air-water temperature relationships in the trout streams of southeastern Minnesota's carbonate-sandstone landscape. *JAWRA Journal of the American Water Resources Association* 49(4):896-907.

Küçük M & Ağirali˙oğlu N (2006) Wavelet Regression Technique for Streamflow Prediction. *J Appl Stat* 33(9):943-960.

Laanaya F (2015) *Modélisation de la température de l'eau en rivière à l'aide du modèle additif généralisé et comparaison avec d'autres approches statistiques*. (Université du Québec, Institut national de la recherche scientifique).

Laanaya F, St-Hilaire A & Gloaguen E (2017) Water temperature modelling: comparison between the generalized additive model, logistic, residuals regression and linear regression models. *Hydrological Sciences Journal* 62(7):1078-1093.

Langan SJ, Johnston L, Donaghy MJ, Youngson AF, Hay DW & Soulsby C (2001) Variation in river water temperatures in an upland stream over a 30-year period. *Sci Total Environ* 265(1-3):195-207.

Larson LL & Larson SL (1996) Riparian shade and stream temperature: a perspective. *Rangelands Archives* 18(4):149-152.

Lee & Ouarda T (2010) Long-term prediction of precipitation and hydrologic extremes with nonstationary oscillation processes. *Journal of Geophysical Research: Atmospheres* 115(D13).

Lee & Ouarda T (2012) An EMD and PCA hybrid approach for separating noise from signal, and signal in climate change detection. *Int J Climatol* 32(4):624-634.

Lessard JL & Hayes DB (2003) Effects of elevated water temperature on fish and macroinvertebrate communities below small dams. *River Res Appl* 19(7):721-732.

Li J, Duan Z & Huang J (2018) Multi-scale fluctuation analysis of precipitation in Beijing by Extreme-point Symmetric Mode Decomposition. *Proceedings of the International Association of Hydrological Sciences* 379:187-192.

Lio P (2003) Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* 19(1):2-9.

Liu B, Yang D, Ye B & Berezovskaya S (2005) Long-term open-water season stream temperature variations and changes over Lena River Basin in Siberia. *Global and Planetary Change* 48(1-3):96-111.

Loh C-H, Wu T-C & Huang NE (2001) Application of the empirical mode decomposition-Hilbert spectrum method to identify near-fault ground-motion characteristics and structural responses. *Bulletin of the seismological Society of America* 91(5):1339-1357.

Maheu A (2015) Développement d'outils de caractérisation et de modélisation du régime thermique des rivières naturelles et régulées. *(Université du Québec, Institut national de la recherche scientifique, Centre Eau-Terre-Environnement)*:226.

Marceau P, Cluis D & Morin G (1986) Comparaison des performances relatives à un modèle déterministe et à un modèle stochastique de température de l'eau en rivière. *Canadian Journal of Civil Engineering* 13(3):352-364.

Masselot P, Chebana F, Belanger D, St-Hilaire A, Abdous B, Gosselin P & Ouarda T (2018) EMD-regression for modelling multi-scale relationships, and application to weather-related cardiovascular mortality. *Sci Total Environ* 612:1018-1029.

Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JF, Stouffer RJ & Taylor KE (2007) The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American meteorological society* 88(9):1383-1394.

Mohseni & Stefan HG (1999) Stream temperature/air temperature relationship: a physical interpretation. *Journal of Hydrology* 218(3-4):128-141.

Mohseni, Stefan HG & Eaton JG (2003) Global warming and potential changes in fish habitat in US streams. *Climatic Change* 59(3):389-409.

Mohseni, Stefan HG & Erickson TR (1998a) A nonlinear regression model for weekly stream temperatures. *Water Resources Research* 34(10):2685-2692.

Mohseni O, Stefan HG & Erickson TR (1998b) A nonlinear regression model for weekly stream temperatures. *Water Resources Research* 34(10):2685-2692.

Morrill JC, Bales RC & Conklin MH (2005) Estimating Stream Temperature from Air Temperature: Implications for Future Water Quality. *J Environ Eng* 131(1):139-146.

Neumann DW, Rajagopalan B & Zagona EA (2003) Regression Model for Daily Maximum Stream Temperature. *J Environ Eng* 129(7):667-674.

Olden JD & Naiman RJ (2010) Incorporating thermal regimes into environmental flows assessments: modifying dam operations to restore freshwater ecosystem integrity. *Freshwater Biol* 55(1):86-107.

Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ & Osuch M (2015) Comparing various artificial neural network types for water temperature prediction in rivers. *Journal of Hydrology* 529:302-315.

Poff NL & Zimmerman JK (2010) Ecological responses to altered flow regimes: a literature review to inform the science and management of environmental flows. *Freshwater Biol* 55(1):194-205.

Poirel A, Gailhard J & Capra H (2010) Influence des barrages-réservoirs sur la température de l'eau : exemple d'application au bassin versant de l'Ain. *La Houille Blanche* (4):72-79.

Poole GC & Berman CH (2001) An ecological perspective on in-stream temperature: natural heat dynamics and mechanisms of human-causedthermal degradation. *Environmental management* 27(6):787-802.

Prats J, Val R, Dolz J & Armengol J (2012) Water temperature modeling in the Lower Ebro River (Spain): Heat fluxes, equilibrium temperature, and magnitude of alteration caused by reservoirs and thermal effluent. *Water Resources Research* 48(5).

Qin L, Ma S, Lin J-C & Shia B-C (2016) Lasso Regression Based on Empirical Mode Decomposition. *Communications in Statistics - Simulation and Computation* 45(4):1281-1294.

Rahman, Charron C, Ouarda TB & Chebana F (2018) Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Stoch Env Res Risk A* 32(1):123-139.

Rehman N, Park C, Huang NE & Mandic DP (2013) EMD via MEMD: multivariate noise-aided computation of standard EMD. *Advances in Adaptive Data Analysis* 5(02):1350007.

Rilling G (2007a) *Décompositions Modales Empiriques. Contributions à la théorie, l'algorithmie et l'analyse de performances*.).

Rilling G (2007b) *Décompositions Modales Empiriques. Contributions à la*

théorie, l'algorithmie et l'analyse de performances. (Ecole normale supérieure de lyon - ENS LYON).

Rilling G, Flandrin P & Goncalves P (2003) On empirical mode decomposition and its algorithms. *IEEE-EURASIP workshop on nonlinear signal and image processing.* NSIP-03, Grado (I), p 8-11.

Salter M, Ratkowsky D, Ross T & McMeekin T (2000) Modelling the combined temperature and salt (NaCl) limits for growth of a pathogenic Escherichia coli strain using nonlinear logistic regression. *International journal of food microbiology* 61(2-3):159-167.

Sandersfeld T, Mark FC & Knust R (2017) Temperature-dependent metabolism in Antarctic fish: Do habitat temperature conditions affect thermal tolerance ranges? *Polar Biology* 40(1):141-149.

Sifuzzaman M, Islam M & Ali M (2009) Application of wavelet transform and its advantages compared to Fourier transform.

Singer EE & Gangloff MM (2011) Effects of a small dam on freshwater mussel growth in an Alabama (USA) stream. *Freshwater Biol* 56(9):1904-1915.

St-Hilaire A, Morin G, El-Jabi N & Caissie D (2000) Water temperature modelling in a small forested stream: implication of forest canopy and soil temperature. *Canadian Journal of Civil Engineering* 27(6):1095-1108.

St-Hilaire A, Ouarda TB, Bargaoui Z, Daigle A & Bilodeau L (2012) Daily river water temperature forecast model with ak-nearest neighbour approach. *Hydrological Processes* 26(9):1302-1310.

Thioune (2015a) *Décomposition modale empirique et décomposition spectrale intrinsèque : applications en traitement du signal et de l'image*. (Université Paris-Est).

Thioune A (2015b) *Décomposition modale empirique et décomposition spectrale intrinsèque: applications en traitement du signal et de l'image*. (Paris Est).

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267-288.

Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(3):273-282.

Van Vliet M, Ludwig F, Zwolsman J, Weedon G & Kabat P (2011) Global river temperatures and sensitivity to atmospheric warming and changes in river flow. *Water Resources Research* 47(2).

Wang Z-Y, Qiu J & Li F-F (2018) Hybrid Models Combining EMD/EEMD and ARIMA for Long-Term Streamflow Forecasting. *Water* 10(7):853.

Webb B (1996) Trends in stream and river temperature. *Hydrological processes* 10(2):205-226.

Webb B, Clack P & Walling D (2003) Water–air temperature relationships in a Devon river system and the role of flow. *Hydrological processes* 17(15):3069-3084.

Webb B & Nobilis F (1997) Long-term perspective on the nature of the air–water temperature relationship: a case study. *Hydrological Processes* 11(2):137-147.

Wehrly KE, Brenden TO & Wang L (2009) A comparison of statistical approaches for predicting stream temperatures across heterogeneous landscapes. *JAWRA Journal of the American Water Resources Association* 45(4):986-997.

Wood (2006) Generalized Additive Models: An Introduction with R.,(Chapman and Hall: CRC Press, Boca Raton, FL.).

Wood (2017) *Generalized additive models: an introduction with R.* CRC press,

Wu C, Chau K & Li Y (2009) Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resources Research* 45(8).

Yang AC, Fuh JL, Huang NE, Shia BC, Peng CK & Wang SJ (2011a) Temporal associations between weather and headache: analysis by empirical mode decomposition. *Plos One* 6(1):e14612.

Yang AC, Tsai SJ & Huang NE (2011b) Decomposing the association of completed suicide with air pollution, weather, and unemployment data at different time scales. *J Affect Disord* 129(1-3):275-281.

Zhang J, Yan R, Gao RX & Feng Z (2010) Performance enhancement of ensemble empirical mode decomposition. *Mech Syst Signal Pr* 24(7):2104-2123.

Zhu S, Heddam S, Nyarko EK, Hadzima-Nyarko M, Piccolroaz S & Wu S (2019) Modeling daily water temperature for rivers: comparison between adaptive neuro-fuzzy inference systems and artificial neural networks models. *Environ Sci Pollut Res Int* 26(1):402-420.

Zhu S, Nyarko EK & Hadzima-Nyarko M (2018) Modelling daily water temperature from air temperature for the Missouri River. *PeerJ* 6:e4894.

Appendix

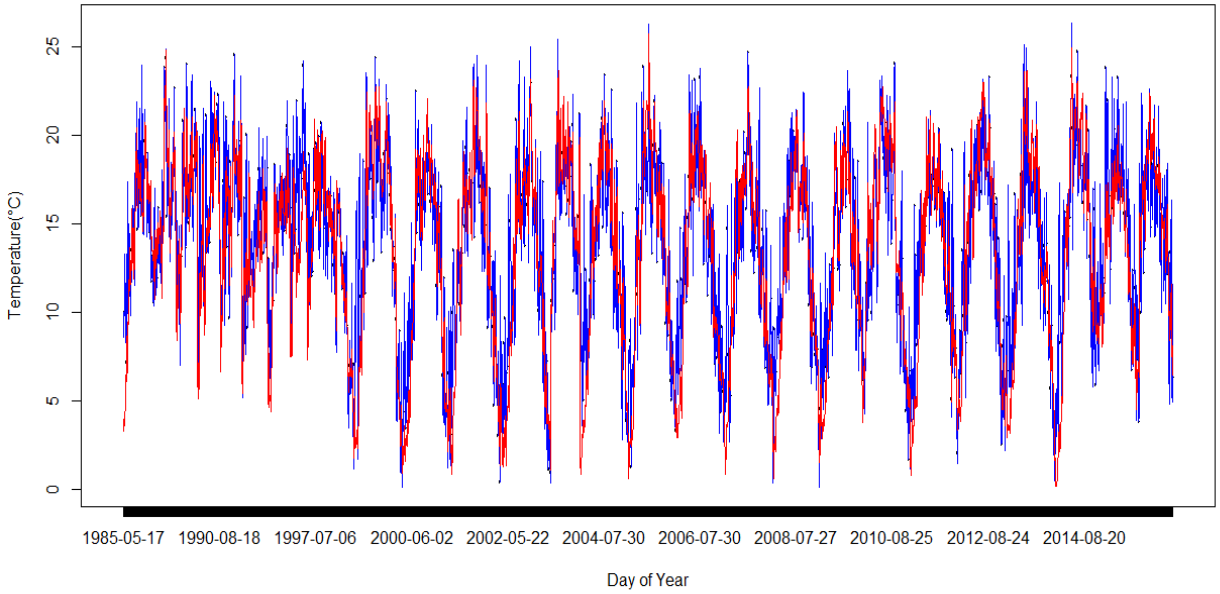


Figure 5.1 Average daily water and air temperature in Trinity River

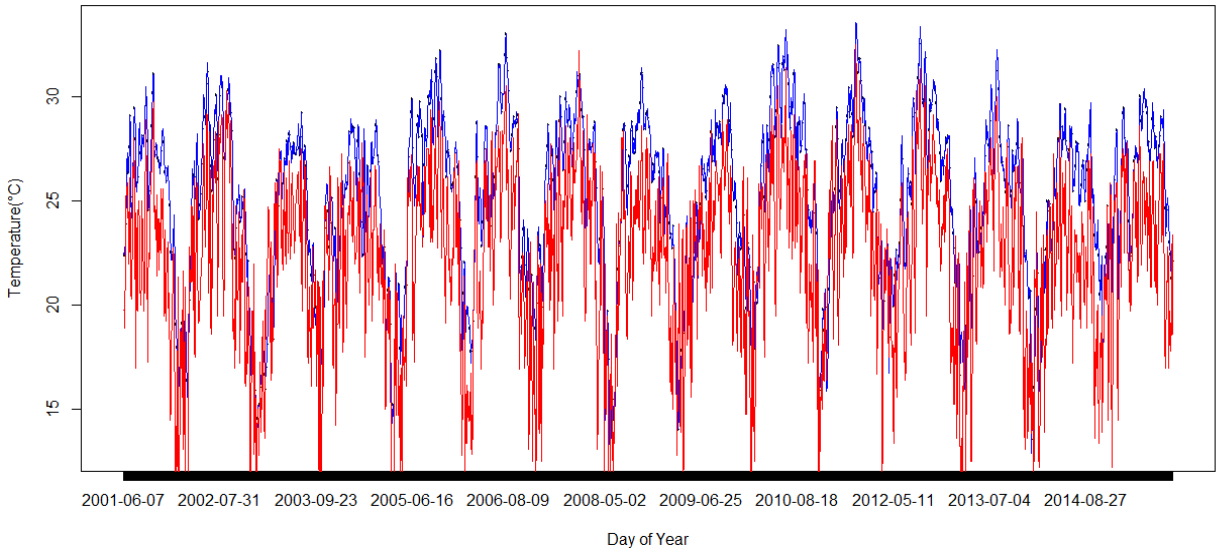


Figure 5.2 Average daily water and air temperature in Potomac River

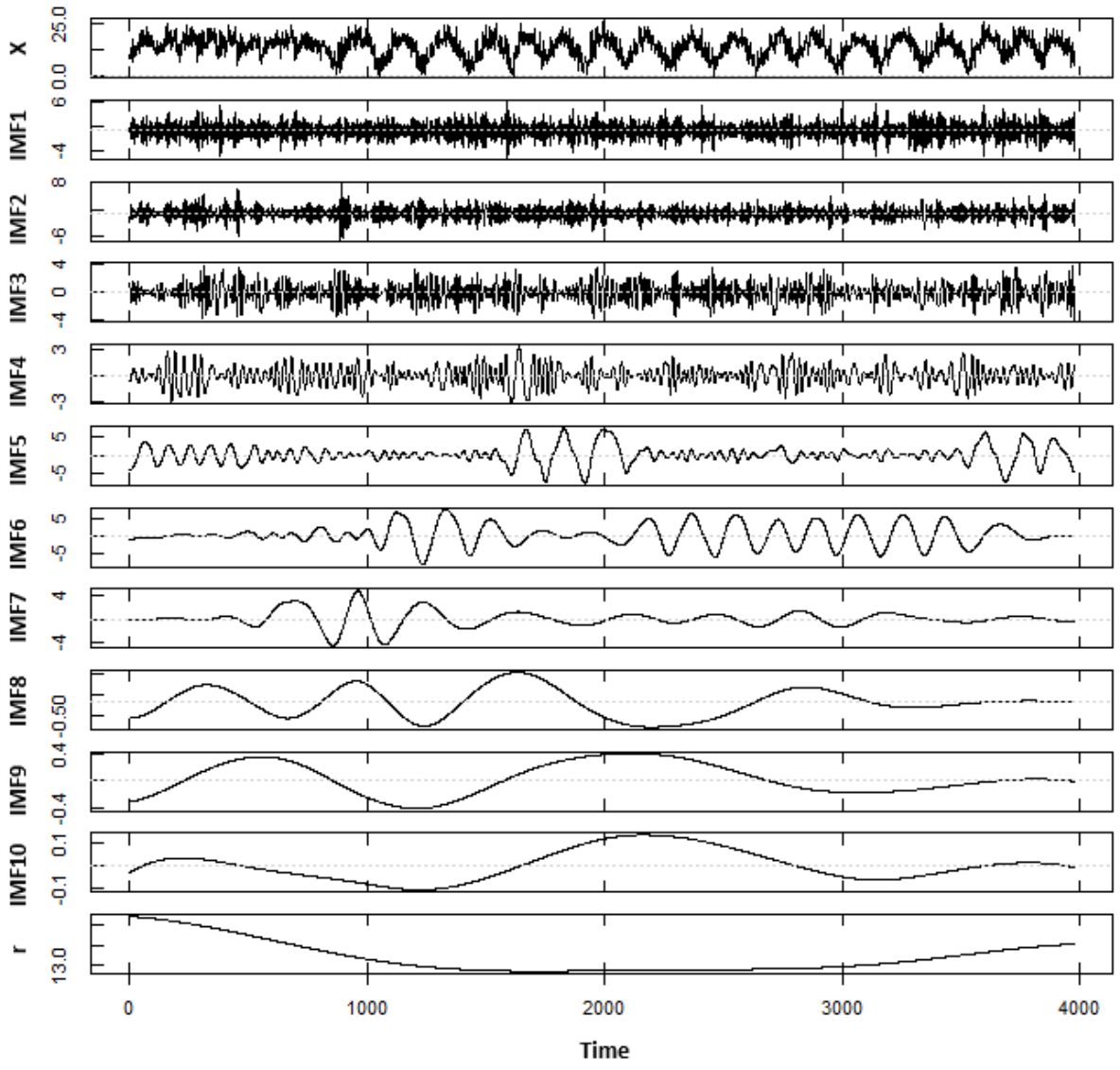


Figure A3 Decomposed air temperature series with the EMD algorithm (Trinity)

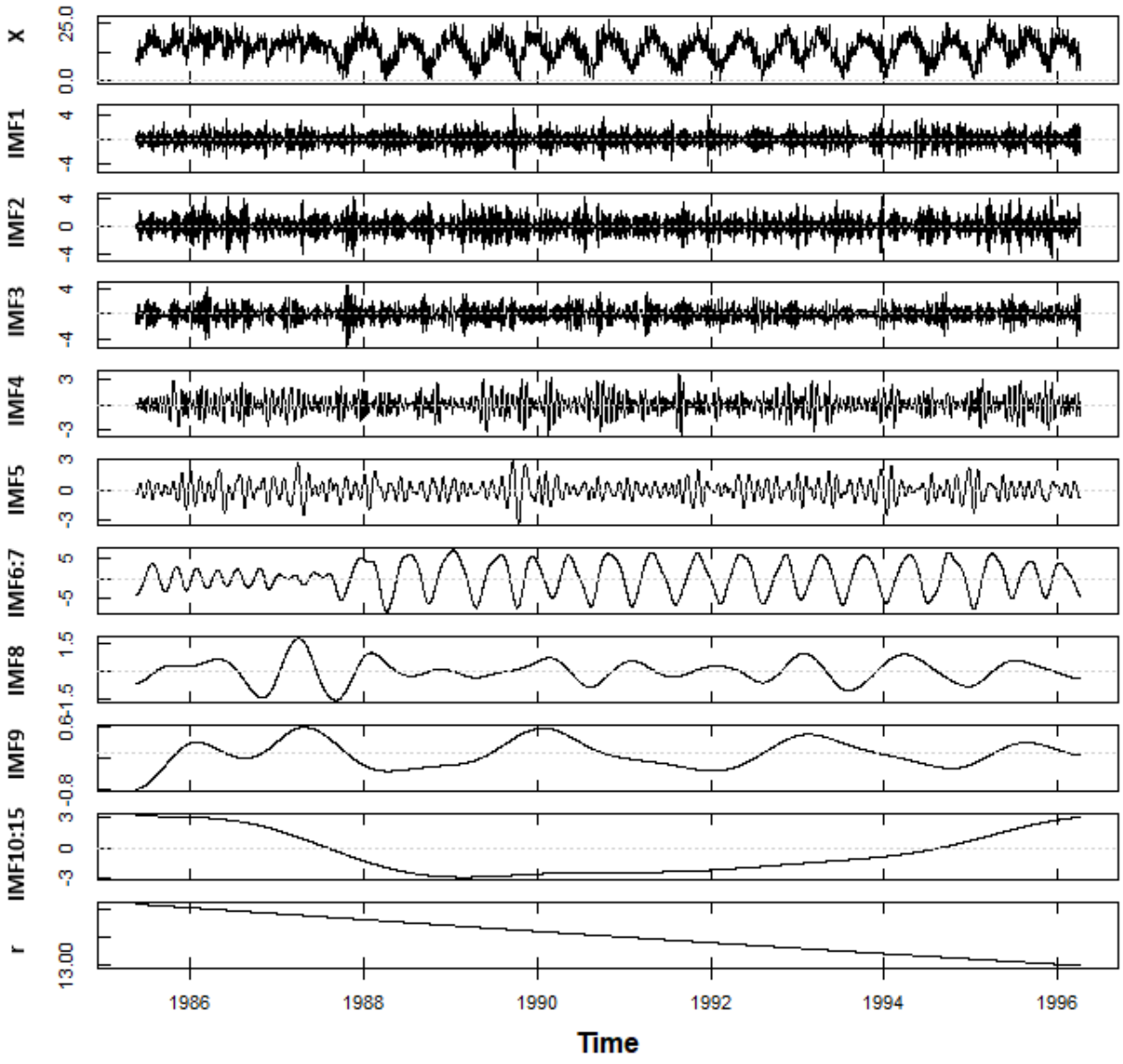


Figure A4 Decomposed air temperature series with the EEMD algorithm (Trinity)

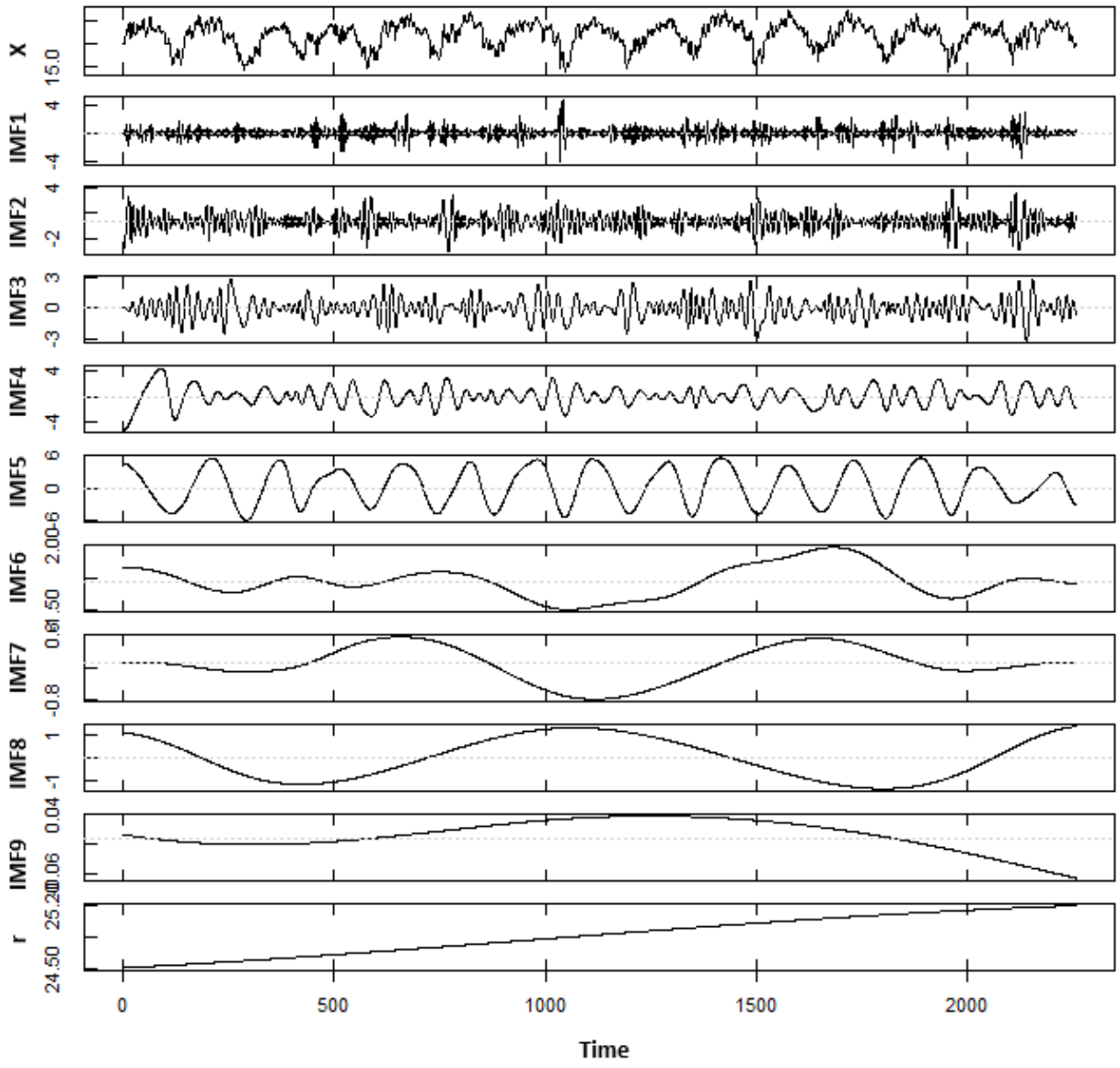


Figure A5 Decomposed air temperature series with the EMD algorithm (Potomac)

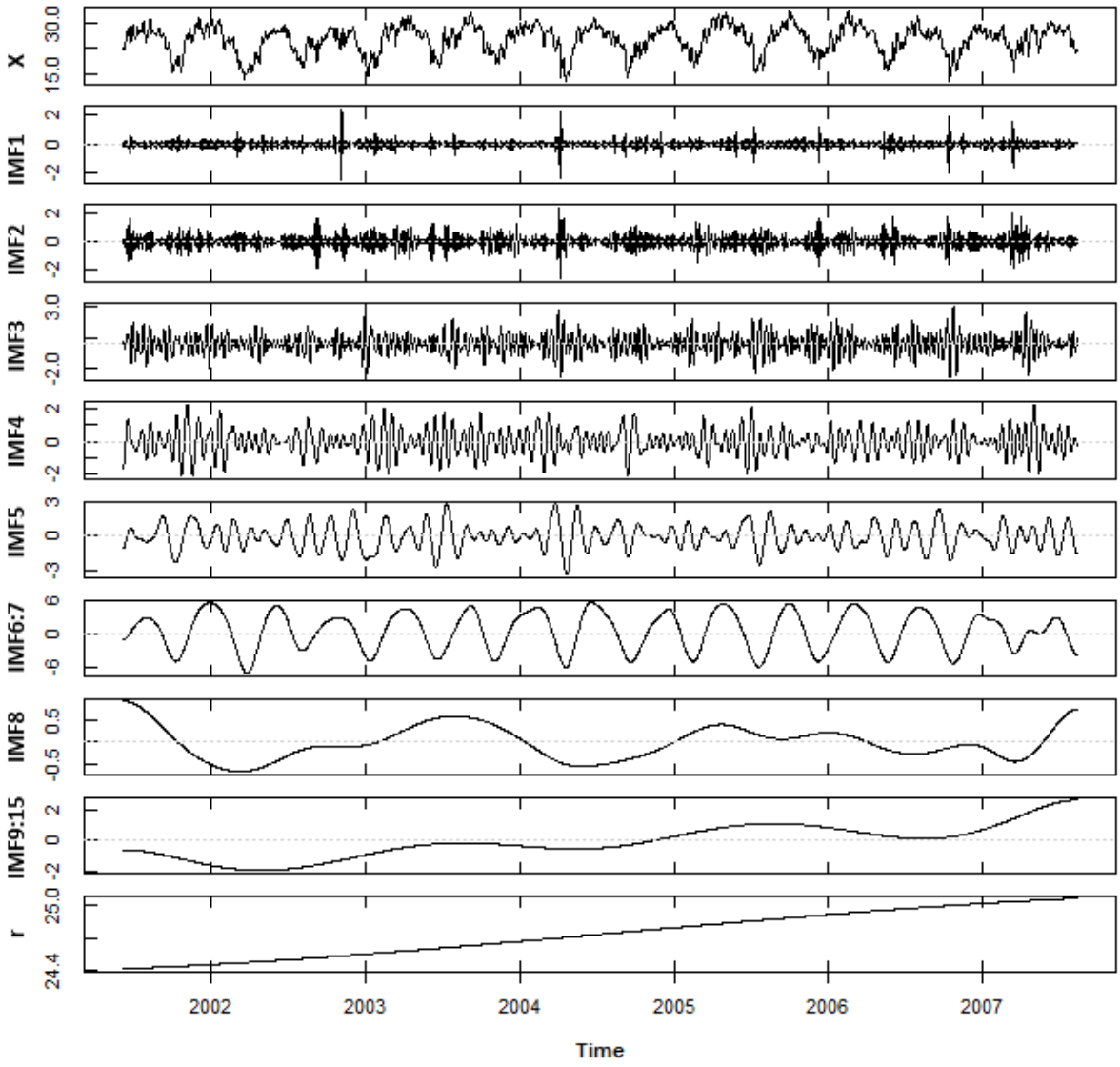


Figure A6 Decomposed air temperature series with the EEMD algorithm (Potomac)

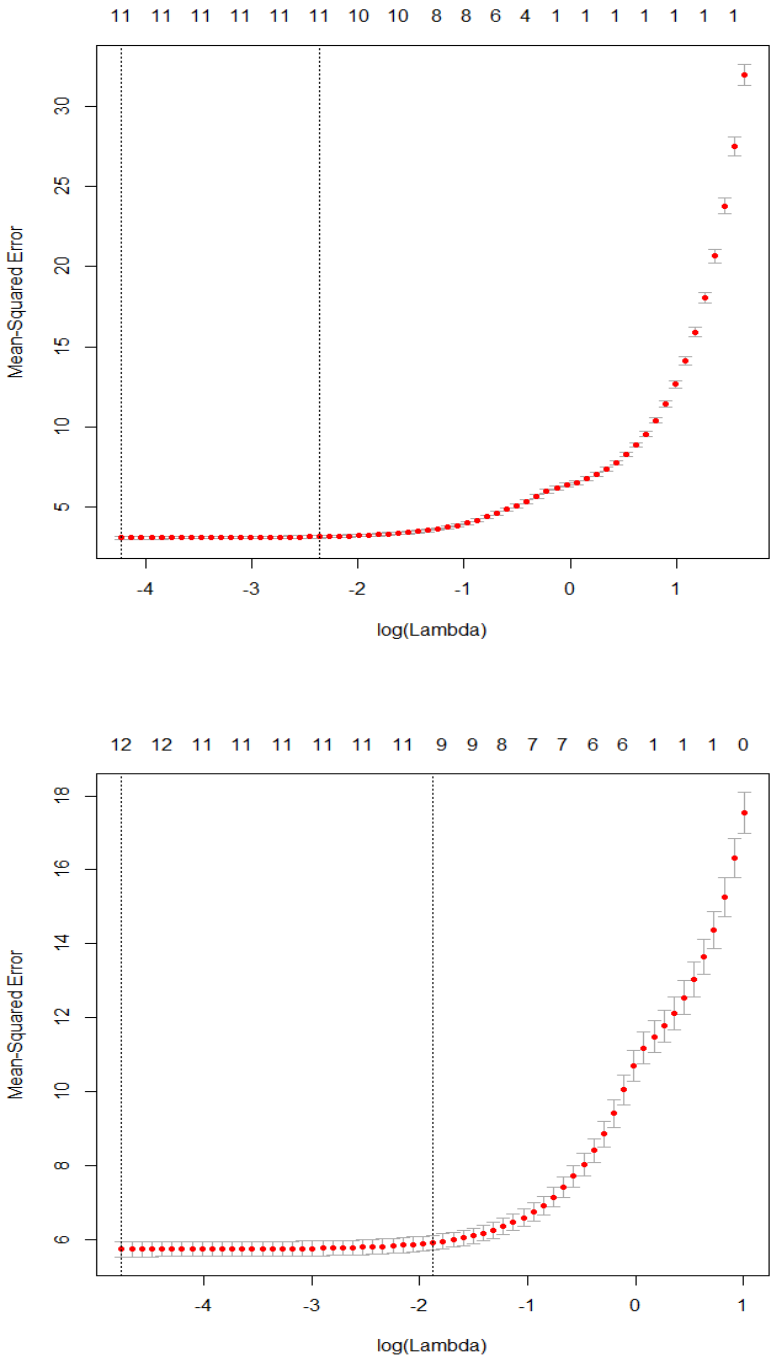
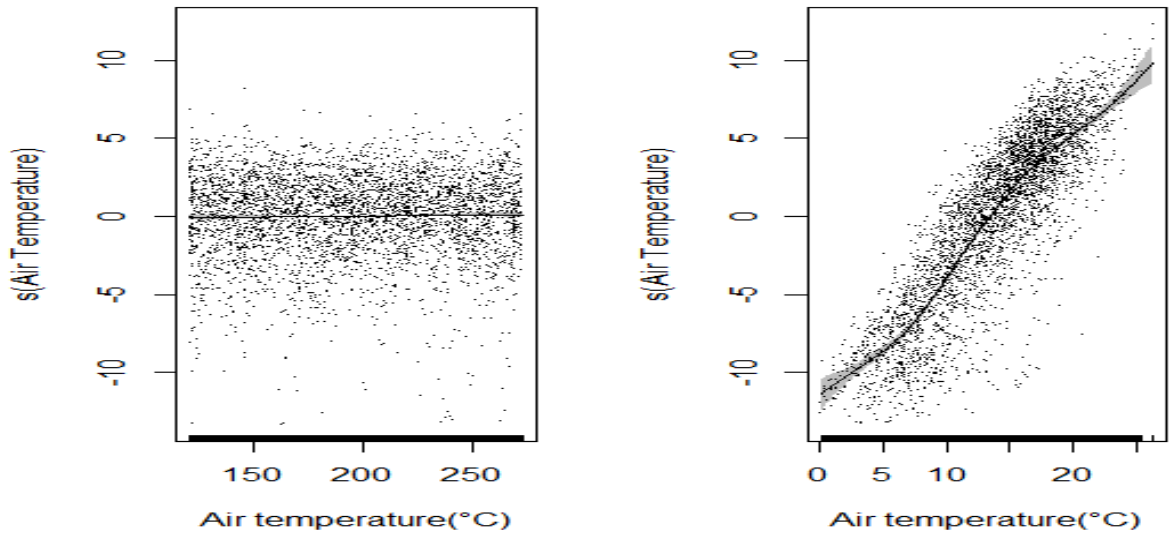


Figure A7 : Adjusted validation of Trinity (Boudraa et al.) & Potomac (bottom)cases

TRINITY



POTOMAC

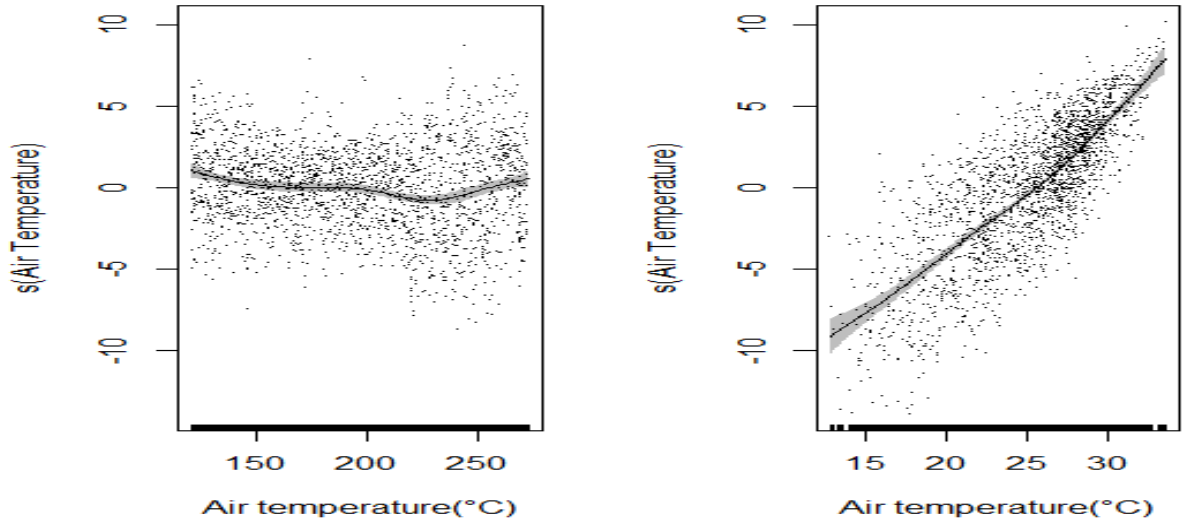


Figure A8 Estimated smooth effect functions with GAM for the Trinity River (Boudraa et al.) & the Potomac River (bottom) for the Julian day of year and the air temperature