International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Using Generalized Additive Models to Assess, Explore and Unify Environmental Monitoring Datasets

Russell G. Richards [a], Rodger Tomlinson [a] and Milani Chaloupka [b]
[a] *Griffith Centre for Coastal Management, Griffith University, Australia*
r.richards@griffith.edu.au
[a] *Griffith Centre for Coastal Management, Griffith University, Australia*
r.tomlinson@griffith.edu.au
[b] *Ecological Modelling Services Pty Ltd, University of Queensland, Australia*
m.chaloupka@uq.edu.au

Abstract:
An on-going challenge for decision makers is the interpretation of temporal trends from monitoring data given that environmental processes often generate complex data that are multivariate and potentially nonlinear. Generalized additive models (GAMs) is a well-suited modelling framework for uncovering such trends and unifying datasets. This approach allows flexible specification of regression splines to represent the functional relationships between a response variable (the parameter of interest) and a suite of temporal and spatial covariates that can be continuous or discrete using a link function and smooth functions of the covariates. We highlight the utility of using GAMs through three case studies. The first highlights the use of a GAM to unify the findings of an established long-term water quality-monitoring program with those of a focused short-term monitoring program. In the second, a GAM is used to evaluate the spatial patterns in a biomonitoring dataset whilst simultaneously accounting for variability in oyster size, which can have a confounding effect on such data. The final case study focuses on a 12 month continuous monitoring program of oceanographic data as part of an evaluation of the environmental conditions for a desalination plant intake pipe. The context for these studies is predominantly water quality in the coastal zone, however the benefits and widespread application to other research areas is clearly evident.

***Keywords**:* GAMs; GAMMs; water quality; nonlinear regression

1. INTRODUCTION

A major challenge to researchers and decision makers alike is interpreting spatial and temporal trends in monitoring data that are multivariate, potentially nonlinear and where spatial structure and dependency might also be important [Bailey et al., 2005]. Generalized additive models (GAMs) and generalized additive mixed models (GAMMs) are well-suited modeling frameworks for uncovering such trends because they allow flexible and non-linear specification of the dependence of some response variable such as the soft-tissue trace metal concentration in an oyster (i.e. biomonitoring data) on a set of temporal and/or spatial covariates without having to specify the model in terms of detailed parametric relationships.

While GAMs have been used extensively in areas such as fisheries [Venables and Dichmont 2004] and species distribution assessments [Guisan et al. 2002], they have been used sparingly on water quality and biomonitoring datasets. Yet these types of monitoring data typically represent a significant proportion of the overall monitoring effort in coastal zone areas. Trend analysis of such data is often assessed through linear models, which is

unrealistic for many applications especially if the behaviour of the response variable is poorly represented by a normal, homoscedastic and additive error term [Venables and Dichmont 2004]. There are various techniques utilizing splines to represent nonlinear and multivariate regression analysis e.g. univariate and additive splines, response surface models. However, GAMs provide a more flexible framework for regression analysis allowing the response variable to be drawn from other distributions of the exponential family including the Poisson, Bionomial and Gamma.

$$g(y_i) = f_1(x_i) + f_2(z_i) + Z_i b + \varepsilon_i \quad (1)$$

The form of a GAM follows (1), where g is some smoothing link function, y_i is the response variable (i.e. the target contaminant), x_i and z_i are some predictor variables, f_1 and f_2 are smooth functions that are estimated and ε_i are independent error terms with a density function described by $N(0,\sigma^2)$. Inclusion of a random effects term, $Z_i b$, extends the GAM to a generalized additive mixed model (GAMM) where Z is a random effects matrix and b is a vector of random effects described by $N(0, \psi_\theta)$ with ψ_θ representing a covariance matrix.

We present here three case studies located in South East Queensland, Australia (Figure 1) that highlight the utility of GAMs/GAMMs for uncovering trends and processes in coastal monitoring data. The first case study shows how a GAM can be used to compare the findings of a focused short-term monitoring program with an established long-term water quality-monitoring program. Case study 2 provides an example of using a GAMM to evaluate the spatial gradients within an oyster-heavy metal biomonitoring dataset whilst simultaneously accounting for the potentially confounding effects of variability in individual oyster size and spatial structure.

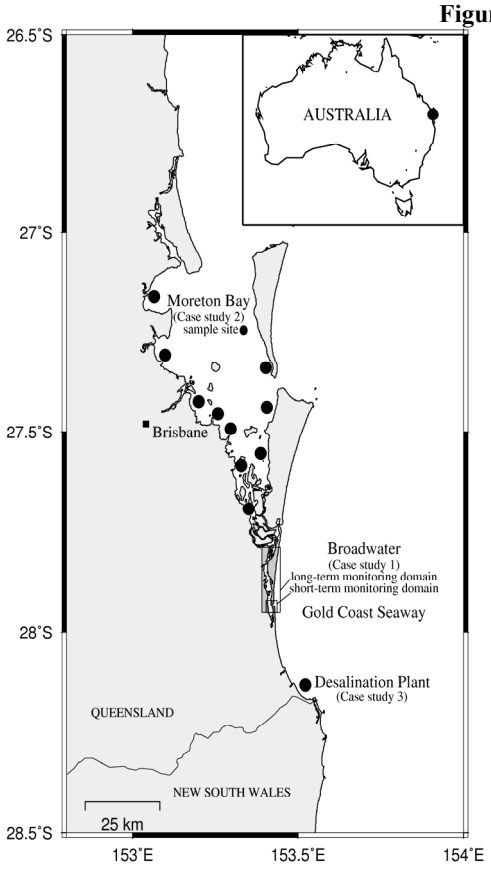
The final case study outlined here provides an insight into the use of a GAMM to explore trends contained within a continuous monitoring program of oceanographic data whilst also accounting for autocorrelation of the error term.

For the three case studies, all continuous variables included in the model(s) were assumed to be potentially non-linear and therefore were fitted with smoother splines. All estimated parameter effects for each model were evaluated by fitting a series of models of decreasing complexity by sequentially removing covariates and comparing the Generalized Cross Validation (GCV) score [Wood 2006]. Diagnostic checks were also carried out to ensure assumptions of independence, normality and constant variance were upheld.

2. CASE STUDY 1: UNIFYING WATER QUALITY DATA FROM TWO MON-ITORING PROGRAMS

The first case study is based on an intensive short-term water quality monitoring program that was carried out in the Gold Coast Broadwater, a shallow estuary located in the South East corner of Queensland (Figure 1). These types of monitoring programs are regularly carried out in coastal zones for the purposes of baseline and/or impact monitoring. This monitoring was conducted over 3 separate days (Day 1, Day 2, Day 3) in early 2009 as part of an

Figure 1. Locations of the three case studies.



assessment of extended release times of recycled water into the Broadwater. Extended release is a management strategy being implemented to cope with the higher treated wastewater loads caused by a growing local population. An outcome of this short-term monitoring program was total nitrogen (TN) concentrations (mean: 326 μg-N l^{-1}) that were significantly higher (t-test; $p<0.001$) than had been measured in the Broadwater (mean: 151 μg-N l^{-1}) during a long-term on-going monitoring program (Ecosystem Health Monitoring Program www.healthywaterways.org). While increased TN concentrations would be expected within the plume of the discharged recycled water, measurements recorded at locations well away from the discharge pipe (ca. 2 km) showed similarly elevated TN concentrations. Assuming that both monitoring programs used identical collection and analytical methods, and even for the sake of this example, the same equipment, personnel and laboratory, comparisons of the results between the two monitoring programs need to be viewed in the context of a range of mechanisms (in addition to the discharge of recycled water) that might be controlling TN concentrations within the Broadwater. Nitrogen concentrations within an estuary typically depend on a range of mechanisms [Eyre and McKee 2002] and the contribution of these will be fluid, both spatially and temporally, and occur at a range of scales [Eyre and McKee 2002]. Consequently, a GAM was used to uncover the functional drivers of TN in the Broadwater based on the water quality data measured during the long-term monitoring (1207 observation sets). The form of the GAM used is shown in (2):

$$g(\mathrm{TN}) = f(\text{conductivity}) + f(\text{year}) + f(\text{month}) + f(\text{lat, long}) + f(\text{TST, by=EBB}) + f(\text{TST, by=FLOOD}) + f(\text{time,wind9am,wind3pm}), \text{ family} = \text{quasi(link=log, variance="mu")} \quad (2)$$

Conductivity was included as a proxy measure of catchment runoff and estuarine mixing. The metrical covariates of *year* and *month* were included as the data may be confounded by seasonal and longer-term non-linear trends [Cox et al. 2005] while the georeferenced coordinates of each monitoring site were included to account for any underlying spatial

effect. To account for the potential effect of tidal current as a mechanism of TN dynamics, a covariate (*TST*) representing the elapsed time between slackwater, whether high or low, and monitoring was included and was conditioned upon whether the tide was ebb or flood flow (i.e. a varying coefficient). Finally, to account for the possible effects of wind-driven re-suspension, a three-term interaction smoother was specified for monitoring time and the local wind speed recorded at 9am and 3pm.

The model was developed and run on the software platform R (http://cran.r-project.org/) utilising the software package mgcv [Wood 2006] with a quasi family and logarithmic link function specified. The model was checked for over-fitting by randomly selecting 90% of the data to fit the data, which was then used to predict for the other 10% [Wood 2006]. The proportion of deviance explained for the fitted and the predicted model were consistent indicating that over-fitting was not occurring.

All variables, except *TST* conditioned on the flood flow, were significant ($p<0.05$) and explained 83.3% of the deviance. Of specific interest in this study were the effects of conductivity, because it was used in the GAM as a proxy measure of catchment loading, along with sampling year and the interaction term between monitoring time and wind speed. A strong negative effect of conductivity was obvious (Figure 2a) and indicated that estuary mixing and/or freshwater catchment flows were a prominent driver of TN concentrations. This is not unexpected because the Broadwater catchment is strongly urbanized and is a significant source of nutrient loading [Moss and Cox 1999; Burton et al. 2004]. However, re-running the GAM with catchment rainfall as a covariate instead of conductivity decreased the performance of the GAM even when potential lag effects between rainfall and TN were accounted for by integrating rainfall over a 1-week period. Possible explanations are that rainfall levels measured at a single measuring station do not adequately represent the catchment rainfall and/or the base flow contribution might be significant. Monitoring year (Figure 2b) is characterised by a strong non-linear smooth with a generally positive effect for 2000-2004, a sharply decreasing effect from 2004-2005, a negative effect for 2005-2008 and a sharply increasing effect for 2009. The period of negative effect broadly coincides with El Niño conditions (2005-2006/2007), which would bring drier conditions to the east coast of Australia [Chiew et al. 1998] and presumably an associated reduced catchment loading. The interaction term between monitoring time and recorded local wind speed (9am, 3pm) (Figure 3) indicated that higher wind speeds (> ca. 25 kph) were important effects on TN. In particular, the 3pm wind speed had an increase positive effect on TN, which increased as the monitoring time occurred later in the day. Conversely, the effect of the morning wind speed (as measured at 9am) became more negative as monitoring time occurred later in the day.

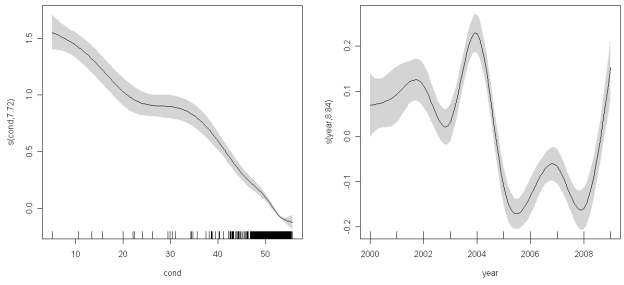


Figure 2. Smoother functions for (a) conductivity and (b) monitoring year. 'Rug' marks along the *x*-axis indicate raw data.

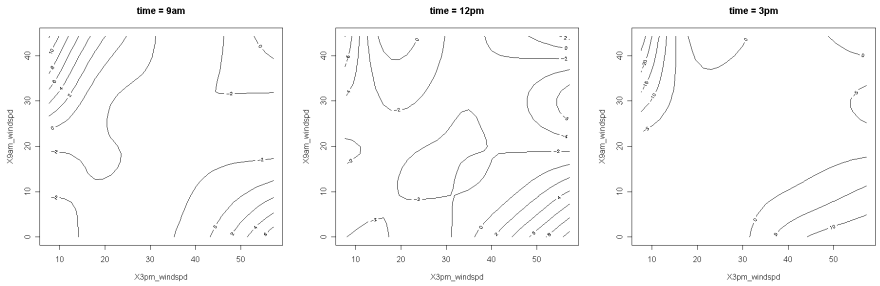


Figure 3. Smoother functions for the interaction term of monitoring time and 9am and 3pm wind speeds recorded at Southport weather station. 'Slices' shown (left to right) for monitoring times of 9am, 12pm and 3pm.

The strong negative relationship between conductivity and TN suggests that freshwater loading into the catchment is an important determinant of observed TN concentrations. Day 3 of the monitoring exhibited the lowest conductivities with a mean of 49 mS cm^{-1} and individual measurements as low as 43 mS cm^{-1} while the mean conductivity levels for Day 1 and Day 2 were above 52 mS cm^{-1}. Coincidently, Day 3 was the only monitoring day that had any significant rainfall (ca. 300mm) recorded in the preceding 3 weeks. Days 1 and 2 were preceded by relatively low rainfall levels. TN concentrations in the Broadwater appear influenced by the timing of the measurements as the short-term monitoring took place in 2009 when a stronger annual effect is expected (refer Figure 2b). Finally, the effect of wind-driven resuspension might be a significant factor in the TN concentrations observed during the short-term monitoring, especially for measurements obtained in the afternoon. In particular, the interaction term (refer Figure 3) indicated that higher TN concentrations would be expected when there was relatively low 9am wind speed (< ca. 15 kph) and relatively high 3pm wind speeds (> ca. 25 kph). These conditions were observed on Day 1 with 9am and 3pm wind speeds of 13.0 and 26.0 kph respectively suggesting that on this date, wind-driven resuspension was an important factor in the observed TN concentrations.

3 CASE STUDY 2: HEAVY METAL BIOMONITORING DATA

Oysters and other bivalves are commonly-used as biomonitors for water quality assessments because they generally fit the criteria for an effective biomonitor. Specifically, they provide an easily measurable and time-integrated indication of contaminant bioavailability. However, it is often difficult to sample oysters of uniform size both within and across sampling sites [Robinson et al., 2005] and various studies have highlighted a significant influence of oyster size on metal bioaccumulation [Mackay et al., 1975; Hayes et al., 1998]. Consequently, a challenge is to unravel the potentially confounding effect of variable oyster size from biomonitoring data so that meaningful and accurate information can be obtained. Here, this potentially confounding effect was addressed by using a generalized additive mixed modelling (GAMM) approach. This enabled size effects and structured and unstructured (random) spatial effects to be considered simultaneously in the biomonitoring data as shown in (3):

$$g(\text{Metal conc.}) = f(\text{oyster weight}) + f(\text{latitude,longitude}) + \text{random effects, family} = \text{normal} \quad (3)$$

The GAMM assessment was carried out on native oysters collected from the intertidal zone of Moreton Bay, a large semi-enclosed embayment located in southeast Queensland, Australia (Figure 1) and consisted of 59 observations covering 10 sampling sites. It is important to note that the western catchment of the Bay is dominated by grazing and forestry [Capelin et al. 1998] and the state capital city of Brisbane (population ca. 840,000) while the eastern catchment is dominated by the sparsely populated Moreton and North Stradbroke Islands. This context is supported by the long-term monitoring carried out as part of the Ecosystem Health Monitoring Program (EHMP) in Moreton Bay, which

highlights that the overall water quality consistently exhibits a broad west-east (high-low) gradient in indicators such as chlorophyll-a, nutrients and turbidity. Six adult oysters were collected from the intertidal zone at ten locations around the Bay (Figure 1) during a one-off survey in July 2001 and the soft-tissue of individual oysters were analysed for Al, Cu, Fe, Mg, Mn and Zn [Richards and Chaloupka 2008]. The GAMMs were fitted in a Bayesian inference framework using Markov chain Monte Carlo (MCMC) simulation techniques implemented in the software program Bayes X [Brezger et al., 2005]. Specific details regarding implementation of this modelling including the specification of priors are described in Richards and Chaloupka [2008].

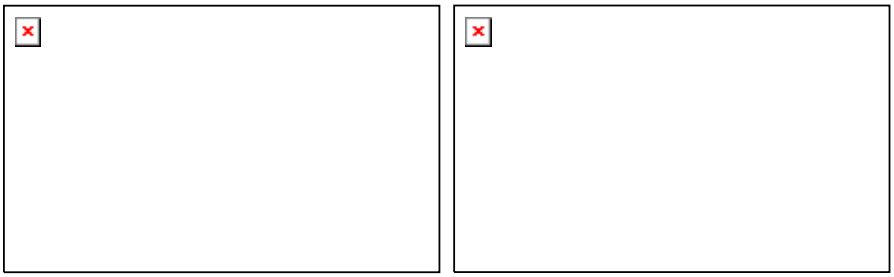


Figure 4. Estimated non-parametric functions of oyster soft-tissue weight (dry weight basis) for Fe and Mg. Posterior mean within the 95% credible region shown.

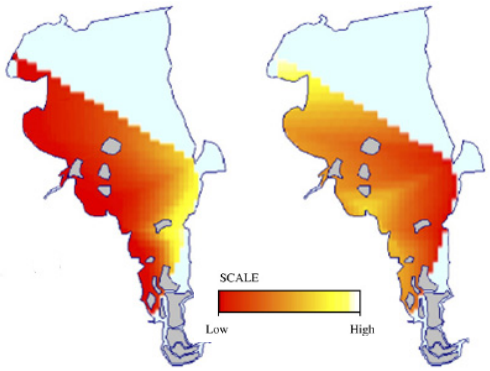


Figure 5. Correlated spatial effect for Mg (left panel) and Mn (right panel).

The posterior plots of soft-tissue Al, Cu, Fe, Mn and Zn concentrations were characterised by broad 95% credible regions that could each be fitted with a zero-gradient line, indicating no significant relationship between oyster soft-tissue mass and trace metal concentration (see Figure 4 for an example for Fe). Mg (Figure 4) was the only trace metal tested that displayed a clear negative effect of oyster size. The correlated spatial effect for each trace metal highlighted a pronounced west-east (high-low) gradient for Cu, Mn (Figure 5) and Zn, which resembles the spatial pattern previously observed for chlorophyll-a [Dennison et al. 1999]. Opposite east-west (high-low) gradients were observed for Al and Mg (Figure 5) while Fe was characterised by 'hotspot' concentrations at Deception Bay to the north and Redland Bay to the south. In this instance, the effect of oyster size as represented by the soft-tissue weight was not a significant effect in the regression modelling for five of the six metals tested. However, there is understandable utility in using a generalized additive modelling approach that explicitly conditions the bioaccumulated metal concentrations on the size of the oysters themselves. This was exemplified for Mg, which was found to have a significant negative size effect and which failure to account for might have resulted in a different spatial gradient and might have resulted in inappropriate management decisions/strategies.

4 CASE STUDY 3: ASSESSMENT OF OCEANOGRAPHIC DATA

The final case study presented here is focused on the results of a continuous monitoring campaign that was carried out to characterise the typical physical oceanographic processes occurring in the vicinity of a desalination inflow and diffuser pipes, South East Queensland, Australia (Figure 1). The implementation of desalination plants in Australia are likely to increase with water demand and a drier environment and therefore understanding the oceanographic processes operating at the intake is an important part in validating the operational design of the plant itself. This monitoring program included the measurement of current speed and direction at 1-meter depth intervals through the water column along with water depth, near-seabed water temperature and turbidity. Current velocity, water depth and temperature were measured using an acoustic Doppler current profiler (ADCP) and turbidity was measured using a YSI sonde. In this case study, a generalized additive mixed model (GAMM) was developed to compliment the assessment of oceanographic processes and help determine the relationship between intake water quality and the ambient environmental conditions. All data was averaged to 30 minute intervals as a means of overcoming the different duty cycles used for the equipment throughout the monitoring program. For example, ADCP velocity measurements were recorded every six, 12 and 20 minutes at various stages of the monitoring.

A subset of the monitoring dataset comprising measurements recorded over a two-month period (January-February 2008) was selected to provide a manageable dataset for the GAMM and this consisted of 1336 observation sets. Seabed turbidity measured by a YSI was selected as the response variable. Turbidity was log-transformed so that it could be sampled from a gaussian distribution, which reduces the computational effort of the model. The predictors initially trialled were the depth-averaged easting (*VelocityE*) and northing (*VelocityN*) current velocities and water depth as measured by a seabed ADCP and represented by the Julian day respectively. The two velocity vectors entails that current direction does not have to be specified. Thin-plate tensor product smooths were specified for the four-predictor variables. Autocorrelation was explicitly modelled to address the potential autoregression in the errors due to the small time intervals of monitoring. The model structure is shown in (4):

$$\log(\text{Turbidity}) \sim f(\text{VelocityE}) + f(\text{VelocityN}) + f(\text{Julian Day}) + f(\text{Depth}), \text{ family} = \text{quasi}(\text{link} = \log, \text{variance}=\text{"mu"}), \text{corr}=\text{corAR1}() \quad (4)$$

The best-fit model was found to be the original initial model less the *VelocityN* covariate. The dominant effect was *Julian Day* and highlighted a decreasing metrical effect (Figure 6).

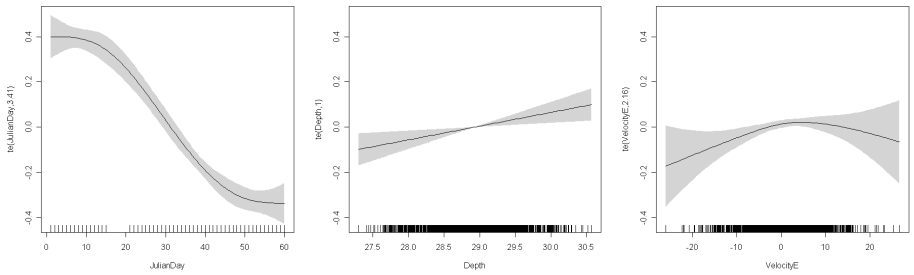


Figure 6. Smooth functions (a) Julian day, (b) water column depth and (c) depth-averaged current easting current velocity.

Depth was found to have a negative linear effect on TN while there was a negative effect of the magnitude of the easting velocity (*VelocityE*). A surprising outcome of the analysis was the negative correlation between turbidity and *VelocityE*, which appears opposite to conventional wisdom of increased current velocity leading to increased turbidity through re-suspension of sediment. However, this observed trend might indicate that there is a turbid layer close to the seabed during 'calm' periods that is resuspended into the water

column when current velocity increases with a net decrease in the turbidity levels at or near the seabed, which has implications for the ongoing performance of the desalination plant.

5 CONCLUSION AND RECOMMENDATIONS

We have presented here three case studies that have highlighted the utility and flexibility of GAMs/GAMMs as techniques for evaluating coastal datasets. Their ability to uncover simultaneous nonlinear functional relationships coupled with their relaxation of the assumption of normality in the residual make them particularly powerful for investigating spatio-temporal trends in environmental datasets and tailoring regional- and local-specific water quality guidelines. This latter point is a key philosophy to setting appropriate water quality guidelines in Australia [Cox et al. 2005]. We have found that GAMs/GAMMs often uncover new or emerging knowledge in the study area as exemplified by all three case studies presented here. In Case Study 1, the flexible specification of the GAM allowed the nonlinear effects of rainfall, tidal flow and wind effects to be accounted for when assessing the main drivers of total nitrogen concentrations. This was despite these predictor variables being poorly aligned with the response variable. We applied a combination of interaction terms (wind effect), proxy variables (conductivity for rainfall) and varying coefficient (tidal effect) approach to address this challenge. Furthermore, we simultaneously conditioned the TN data for annual effects, which had important implications for comparing TN concentrations across different years. In Case Study 2, we used a GAMM to condition a biomonitoring dataset for oyster size, correlated spatial effects and uncorrelated spatial effects so that the underlying bioaccumulation effects could be compared. Failure to account for these combined effects can easily lead to incorrect conclusions about such biomonitoring data, which can propagate through to policy decisions. Finally, in Case Study 3, we used a GAMM to assess a continuous monitoring dataset that had been generated for the assessment of ambient environmental conditions at a desalination plant intake. Key concerns in this instance were the relationships between turbidity and current velocity. As measurements were recorded at short time intervals (e.g. 30 minutes) there was a considerable challenge in assessing such continuous data because independence of the error terms is unlikely. We were able to condition the data for autocorrelation by using a GAMM, with the resulting observation of a negative relationship between easting current velocity and turbidity having potentially important implications for the operation of the desalination plant. Further, the GAMs allowed for straightforward application of formal diagnostic tests during their development. These tests included the comparison of different models using Generalized Cross Validation scores, as well as tests of normality, constant variance, model overfitting and autoregression; these aspects are of crucial importance in model development but are often overlooked. Finally, the context for the studies presented here, including the biomonitoring study, is water quality in the coastal zone. However, the benefits and widespread application of GAMs to other research areas, whether it is the terrestrial or marine environment, is clearly evident.

ACKNOWLEDGEMENTS

We thank Waterway Management, Gold Coast Water (Case Study 1), the CRC for Coastal Zone Estuary (Case Study 2) and Gold Coast Desalination Alliance (Case Study 3) for providing financial assistance. We are also grateful to the Ecosystem Health Monitoring Program project team for their help in providing access to EHMP water quality data for Case Study 1 and 2. Finally, we would like to thank Peter Bell, Lawrence Hughes, Sally Kirkpatrick, Darrell Strauss and Greg Stuart for their assistance in one or more of the case studies presented here.

REFERENCES

Bailey, T.C., Barcellos, C. and Krzanowski, W.J., Use of spatial factors in the analysis of heavy metals in sediments in a Brazilian coastal region, *Environmetrics*, 16, 563-572, 2005.

Brezger, A., Kneib, T. and Lang, S., BayesX: analyzing Bayesian structured additive regression models, *Journal of Statistical Software*, 14, 2005.

Burton, E.D., Phillips, I.R. and Hawker, D.W., Trace metals and nutrients in bottom sediments of the Southport Broadwater, Australia, *Marine Pollution Bulletin*, 48, 378-402, 2004.

Capelin, M., Kohn, P. and Hoffenberg, P., Land use, land cover and land degradation in the catchment of Moreton Bay. In: Tibbetts IR, Hall NJ, Dennison WC, editors. Brisbane, school of marine sciences. The University of Queensland, p55–66, 1998.

Chiew, F.H.S., Piechota, T.C., Dracup, J.A., McMahon, T.A., El Nino/Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting, *Journal of Hydrology*, 204, 138-149, 1998.

Cox, M.E., Moss, A. and Smyth, G.K., Water quality condition and trend in North Queensland waterways, *Marine Pollution Bulletin*, 51, 89–98, 2005.

Dennison, W.C., O'Neil, J.M., Duffy, E.J., Oliver, P.E. and Shaw, G.R., Blooms of the cyanobacterium *Lyngbya majuscule* in coastal waters of Queensland, Australia. *Bulletin of the Institute of Oceanography and Fisheries*, 19, 501-506, 1999.

Eyre, B.D. and McKee, L.J., Carbon, nitrogen, and phosphorus budgets for a shallow subtropical coastal embayment (Moreton Bay, Australia). *Limnology and Oceanography*, 47, 1043-1055, 2002.

Guisan, A., Edwards, T.C. Jr and Hastie, T., Generalized linear and generalized additive models in studies of species distributions: setting the scene, Ecological Modelling, 157, 89-100, 2002.

Hayes, W.J., Anderson, I.J., Gaffoor, M.Z. and Hurtado, J., Trace metals in oysters and sediments of Botany Bay, Sydney, *Science of the Total Environment*, 212, 39–47, 1998.

Mackay, N.J., Williams, R.J., Kacprzac, J.L., Kazacos, M.N., Collins, A.J. and Auty, E.H., Heavy metals in cultivated oysters (*Crassostrea commercialis* = *Saccostrea cucullata*) from the estuaries of New South Wales, *Australian Journal of Marine and Freshwater Research*, 26, 31–46, 1975.

Moss, A. and Cox, M., Southport Broadwater and Adjacent Pacific Ocean: Water Quality study 1997-1998, Queensland Protection Agency, 1999.

Richards, R. and Chaloupka M., Does oyster size matter for modelling trace metal bioaccumulation? *Science of the Total Environment*, 389, 539–544, 2008.

Robinson, W.A., Maher, W.A., Krikowa, F., Nell, J.A. and Hand, R., The use of the oyster Saccostrea glomerata as a biomonitor of trace metal contamination: intra-sample, local scale and temporal variability and its implication for biomonitoring, *Journal of Environmental Monitoring*, 7, 208–23, 2005.

Venables, W.N. and Dichmont, C.M., GLMs, GAMs and GLMMs: an overview of theory for applications in fisheries research. *Fisheries Research*, 70, 319–337, 2004.

Wood, S.N., Generalized Additive Models. An Introduction with R. Chapman and Hall, Boca Raton, 2006.