Multi-Level Multi-Task Learning for Modeling Cross-Scale Interactions in Nested Geospatial Data

Shuai Yuan
Quora Inc.
syuan@quora.com

Jiayu Zhou
Michigan State University
jiayuz@cse.msu.edu

Pang-Ning Tan
Michigan State University
ptan@cse.msu.edu

Emi Fergus
The National Research Council
emifergus@gmail.com

Tyler Wagner
U.S. Geological Survey
txw19@psu.edu

Patricia A. Soranno
Michigan State University
soranno@anr.msu.edu

***Abstract*—Predictive modeling of nested geospatial data is a challenging problem as the models must take into account potential interactions among variables defined at different spatial scales. These cross-scale interactions, as they are commonly known, are particularly important to understand relationships among ecological properties at macroscales. In this paper, we present a novel, multi-level multi-task learning framework for modeling nested geospatial data in the lake ecology domain. Specifically, we consider region-specific models to predict lake water quality from multi-scaled factors. Our framework enables distinct models to be developed for each region using both its local and regional information. The framework also allows information to be shared among the region-specific models through their common set of latent factors. Such information sharing helps to create more robust models especially for regions with limited or no training data. In addition, the framework can automatically determine cross-scale interactions between the regional variables and the local variables that are nested within them. Our experimental results show that the proposed framework outperforms all the baseline methods in at least 64% of the regions for 3 out of 4 lake water quality datasets evaluated in this study. Furthermore, the latent factors can be clustered to obtain a new set of regions that is more aligned with the response variables than the original regions that were defined a priori from the ecology domain.**

I. Introduction

Predictive modeling of geospatial data is an important problem in many domains. For example, scientists seek to develop models from geospatial data that can help explain natural and anthropogenic factors influencing environmental variability [1]–[3]. However, building a robust geospatial prediction model can be challenging as the underlying processes of the system may interact at different spatial scales. In the lake ecology domain, previous studies have found strong evidence for cross-scale interactions between geospatial driver variables quantified at local and regional spatial scales for predicting lake nutrient concentrations [4]. Cross-scale interactions (CSIs) [5] refer to the coupling between the local and regional variables and their joint effect on the focal response variable. For example, interactions between local wetland cover around a lake and regional agricultural land use have been shown to affect the performance of models predicting total phosphorus concentrations in lakes [6]. Nested geospatial data, containing variables measured at multiple spatial scales, are needed to detect such patterns. For modeling lake nutrients, the predictor variables used include local drivers such as lake depth, lake type, and amount of wetland areas surrounding the lake as well as regional drivers such as climate and land use (see Fig. 1). In this example, the values of the local predictor variables may vary from one lake to another but the values of the regional predictor variables are the same for all lakes within the same region. The nature of such nested data makes it challenging to effectively incorporate both local and regional variables into the model formulation. On one hand, the local and regional variables can be concatenated to form a multi-scale feature vector from which a global regression model can be fitted against the data. Unfortunately, such a strategy may not be effective due to the complex relationship between the predictor and response variables, making it difficult to construct an accurate, one-size-fits-all model for all the regions. On the other hand, a local regression model can be trained to fit the data in each region separately. However, such a model would ignore the regional variables altogether as their values would be identical for all lakes in the same region.

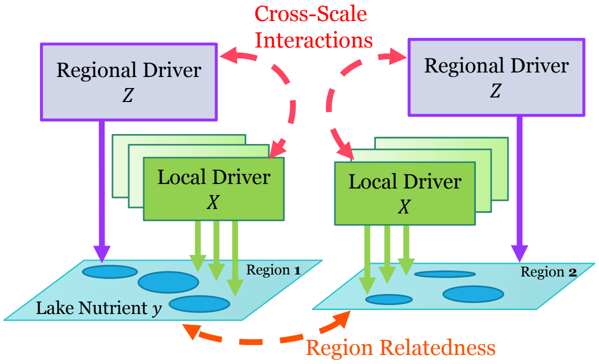


Figure 1: Example of nested lake ecology data with cross-scale interactions between the local and regional predictor variables.

Another challenge in the predictive modeling of nested geospatial data is the unbalanced sample sizes across different regions. If the underlying relationships in the geospatial domain are complex, the predictive models developed for data-poor regions are likely to be inferior compared to those developed for data-rich regions. Finally, the original set of regions from which the nested data were obtained may not be ideal for predictive modeling as they were often defined for other purposes such as political boundaries and management policies. Indeed, it is possible that the lakes from the same region may not share the same relationship between their predictor and response variables. It would be desirable to develop a framework that can identify a set of regions that better capture the relationship between the predictor and response variables of the data.

To address the above challenges, this paper presents a novel multi-level multi-task (MLMT) learning framework for the predictive modeling of nested geospatial data from the ecology domain. The framework enables a distinct prediction model to be trained for each region using both its local and regional predictor variables. The framework assumes that the nested geospatial data are characterized by a set of low-rank latent factors, which relates the dependencies between the local and regional predictor variables to the response variable. Instead of building the models for each region independently, the models are jointly trained by inferring their local and regional latent factors. The shared latent factors provide several advantages for our framework. First, they enable the data-poor regions to leverage information from other regions in order to construct more robust models. Second, the latent factors can be used to identify CSIs in the nested geospatial data. Finally, they provide a new feature representation for each lake, which allows us to cluster the lakes into a new set of regions based on the similarity of their local latent factors. Empirical results using four lake water quality datasets from the LAGOS-NE database [7] showed significant performance improvement in the local prediction models when trained on the new set of regions instead of the original, pre-defined regions of the data.

II. Preliminaries

We consider a two-level nested geospatial data set, $\mathcal{D} = \{\mathbf{X}_i, \mathbf{y}_i, \mathbf{z}_i\}_{i=1}^r$, where r is the number of regions. Let $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$ be the design matrix containing the local predictor variables in region i, $\mathbf{y}_i \in \mathbb{R}^{n_i}$ be the corresponding values of response variables, and $\mathbf{z}_i \in \mathbb{R}^k$ be the corresponding regional predictor variables. Here d is the number of local predictors, k is the number of regional predictors, and n_i is the number of geospatial objects (e.g., lakes) in region i.

The goal of geospatial predictive modeling is to learn a target function $f(\mathbf{x}, \mathbf{z})$ that maps the local and regional predictor variables of a geospatial object ($\mathbf{x} \in \mathbb{R}^d, \mathbf{z} \in \mathbb{R}^k$) to its response value y with minimal prediction error. A trivial way to model nested geospatial data is by fitting a single, global model f_{global} to the entire data set $\mathcal{D}$. Unfortunately, the global model may not provide a good fit to the data especially if the relationship between the predictor and response variables may vary by region. Alternatively, one could train an independent, local model f_{local} for each region, but this approach is also not as effective especially for regions that have very few training examples available. Furthermore, the local models will not be able to utilize the regional variables since their values are the same for all the training examples in the same region. Alternative techniques are therefore needed for modeling nested geospatial data.

Multi-level modeling [8] is a widely-used statistical technique for assessing the influence of multi-scale variables on the response variable of interest. For a two-level model, the relationship between the response and predictor variables for a geospatial object $(\mathbf{x}_i, \mathbf{z}_i)$ in region i is given as follows:

$$\begin{aligned} y_i &= \mathbf{w}_i^T \mathbf{x}_i + \epsilon_1, \qquad \epsilon_1 \sim \mathbb{N}(0, \sigma_1^2) \\ \mathbf{w}_i &= \mathbf{G}^T \mathbf{z}_i + \boldsymbol{\epsilon}_2, \qquad \boldsymbol{\epsilon}_2 \sim \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma}_2), \end{aligned} \tag{1}$$

where $\mathbf{G} \in \mathbb{R}^{k \times d}$ is a matrix that captures the CSIs between the local and regional predictors. Specifically, the (i, j)-th element of $\mathbf{G}$ corresponds to the cross-scale interaction term between the i-th regional predictor and the j-th local predictor. It can be shown that the maximum likelihood estimation (MLE) of $\mathbf{G}$ can be found by minimizing the following loss function: $L(\mathbf{G}) = \sum_{i=1}^r \parallel \mathbf{y}_i - \mathbf{X}_i \mathbf{G}^T \mathbf{z}_i \parallel_2^2$. Several variants of the formulation have also been proposed in the literature. For example, Zhao et al. [9] presented a multi-level modeling approach for hierarchical multi-source event forecasting. Lozano et al. [10] presented the following multi-level lasso formulation for multi-task regression:

$$\min_{\mathbf{G}} \ \frac{1}{2} \sum_{i=1}^r \parallel \mathbf{y}_i - \mathbf{X}_i \mathbf{G}^T \mathbf{z}_i \parallel_2^2 + \rho_1 \parallel \mathbf{G} \parallel_1 \tag{2}$$

Since the $\{\mathbf{z}_i\}$ is given, the optimization problem can be solved by using the proximal gradient descent method. During the prediction step, the value of the response variable for a test instance $(\mathbf{x}^*, \mathbf{z}^*)$ can be predicted as follows:

$$\begin{aligned} \hat{y} = \mathbf{z}^{*T} \mathbf{G} \mathbf{x}^* &= G_{11} + \sum_{p=2}^d G_{1p} x_p^* + \sum_{q=2}^k G_{q1} z_q^* \\ &\quad + \sum_{p,q>1} z_q^* G_{qp} x_p^* \end{aligned} \tag{3}$$

In the preceding equation, $x_1^* = z_1^* = 1$ and G_{11} is the intercept term of the model. The second term measures the effect of the local predictors on the response variable whereas the third term measures the effect of the regional predictors. The last term of the equation quantifies the influence due to joint coupling of the local and regional predictors on the response variable y. A non-zero value in G_{qp}, where $p, q > 1$, can thus be regarded as evidence for a CSI between the p-th local and q-th regional predictors.

III. Multi-Level Multi-Task Learning (MLMT) Framework

This section presents the proposed MLMT framework for modeling CSIs in nested geospatial data.

A. Objective Function

The traditional multi-level model formulation shown in Equation (1) restricts the regression coefficients for all the regions to lie in the column space of $\mathbf{G}^T$. In contrast, our proposed formulation assumes that the regression coefficients for all the regions share a common set of latent factors. Specifically, each $\mathbf{w}_i$ is decomposed into a product of two terms: a latent factor matrix $\mathbf{U} \in \mathbb{R}^{d\times m}$ that is shared by all the regions and a vector $\mathbf{v}_i \in \mathbb{R}^m$ that is shared by all the geospatial objects in region i, where d is the number of local predictors and m is the number of latent factors. Instead of regressing $\mathbf{w}_i$ directly against the regional variables $\mathbf{z}_i$, we regress the latent factor $\mathbf{v}_i$ against $\mathbf{z}_i$, which leads to the following optimization problem:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{R}} \quad \frac{1}{2}\sum_{i=1}^{r} \parallel \mathbf{y}_i - \mathbf{X}_i\mathbf{U}\mathbf{v}_i \parallel_2^2 + \frac{\rho_1}{2}\sum_{i=1}^{r} \parallel \mathbf{z}_i - \mathbf{R}\mathbf{v}_i \parallel_2^2 + \rho_2 \parallel \mathbf{U} \parallel_1 + \rho_3 \parallel \mathbf{V} \parallel_1 + \rho_4 \parallel \mathbf{R} \parallel_1, \tag{4}$$

where $\mathbf{V} = [\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_r]$ and r is the number of regions. The first term in Equation (4) corresponds to the squared loss prediction error of the model while the second term corresponds to the error in fitting the regional predictors $\mathbf{Z}$ to $\mathbf{V}$. The last 3 terms of the objective function controls the sparsity of the model by enforcing L1-regularization to the latent factors $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{R}$. ρ_1, ρ_2, ρ_3, and ρ_4 are the user-specified parameters. In this formulation, $\mathbf{U}$ represents the latent factors for the local predictors while $\mathbf{R}$ represents the latent factors for the regional predictors.

B. Parameter Estimation

We employ the block coordinate descent approach to minimize the objective function given in Equation (4). Since there are three latent factors ($\mathbf{U}$, $\mathbf{V}$, and $\mathbf{R}$) to be estimated, the algorithm iteratively estimates one of the three latent factors while keeping the other two latent factors fixed. The update formula for each latent factor is given below.

a) Update formula for $\mathbf{V}$*:* Assuming $\mathbf{U}$ and $\mathbf{R}$ are given, the optimization for $\mathbf{V}$ is obtained by minimizing the following objective function:

$$\mathcal{L}(\mathbf{V}) = \frac{1}{2}\sum_{i=1}^{r} \parallel \mathbf{y}_i - \mathbf{X}_i\mathbf{U}\mathbf{v}_i \parallel_2^2 + \frac{\rho_1}{2} \parallel \mathbf{Z}^T - \mathbf{R}\mathbf{V} \parallel_F^2 + \rho_3 \parallel \mathbf{V} \parallel_1 \tag{5}$$

Since $\mathcal{L}(\mathbf{V})$ is not a smooth function, we solve the optimization problem using the proximal gradient descent algorithm. Specifically, $\mathbf{V}$ is iteratively updated by solving the following problem:

$$\mathbf{V}^{(s)} = prox_\lambda(\mathbf{V}^{(s-1)} - \lambda\nabla g(\mathbf{V}^{(s-1)})), \tag{6}$$

where $g(\mathbf{V})$ is the smooth part of the objective function given in Equation (5) and $prox_\lambda(x)$ is a soft thresholding function on x defined as follows: $prox_\lambda(x) = \text{sign}(x)\ \max(x-\lambda, 0)$. The gradient for $g(\mathbf{V})$ is given by: $\nabla g(\mathbf{v_i}) = -(\mathbf{X}_i\mathbf{U})^T(\mathbf{y}_i - \mathbf{X}_i\mathbf{U}\mathbf{v}_i) - \rho_1\mathbf{R}^T(\mathbf{z}_i - \mathbf{R}\mathbf{v}_i)$ which can be plugged into Equation (6) to obtain the new $\mathbf{V}^{(k)}$.

b) Update formula for $\mathbf{U}$*:* Assuming $\mathbf{V}$ and $\mathbf{R}$ are given, the latent factors $\mathbf{U}$ are estimated by minimizing the following objective function:

$$\mathcal{L}(\mathbf{U}) = \frac{1}{2}\sum_{i=1}^{r} \parallel \mathbf{y}_i - \mathbf{X}_i\mathbf{U}\mathbf{v}_i \parallel_2^2 + \rho_2 \parallel \mathbf{U} \parallel_1 \tag{7}$$

Once again, since $\mathcal{L}(\mathbf{U})$ is not a smooth function, we apply proximal gradient descent to update $\mathbf{U}$ as follows:

$$\mathbf{U}^{(s)} = prox_\lambda(\mathbf{U}^{(s-1)} - \lambda\nabla g(\mathbf{U}^{(s-1)})) \tag{8}$$

The gradient of the smooth part of the objective function given in Equation (7) is $\nabla g(\mathbf{U}) = \sum_{i=1}^{r} \mathbf{1}_{m\times 1}(\mathbf{y}_i - \mathbf{X}_i\mathbf{U}\mathbf{v}_i)^T\mathbf{X}_i \odot \mathbf{v}_i\mathbf{1}_{1\times d}$

c) Update formula for $\mathbf{R}$*:* Assuming $\mathbf{U}$ and $\mathbf{V}$ are fixed, the latent factor $\mathbf{R}$ is updated by minimizing the following terms in the objective function that depend on $\mathbf{R}$:

$$\mathcal{L}(\mathbf{R}) = \frac{\rho_1}{2} \parallel \mathbf{Z}^T - \mathbf{R}\mathbf{V} \parallel_F^2 + \rho_4 \parallel \mathbf{R} \parallel_1 \tag{9}$$

The update formula for $\mathbf{R}$ is derived using the proximal gradient descent approach as follows:

$$\mathbf{R}^{(s)} = prox_\lambda(\mathbf{R}^{(s-1)} - \lambda \bigtriangledown g(\mathbf{R}^{(s-1)})) \tag{10}$$

where the gradient of the smooth function $\nabla g(R)$ is given by: $\nabla g(\mathbf{R}) = -\rho_1(\mathbf{Z}^T - \mathbf{R}\mathbf{V})\mathbf{V}^T$

The **MLMT** algorithm updates the model parameters iteratively as follows. First, $\mathbf{W}^{(0)}$ is initialized by applying existing methods such as lasso regression or multi-task learning [11] on the local predictors only. We then factorize $\mathbf{W}^{(0)}$ into a product of $\mathbf{U}^{(0)}$ and $\mathbf{V}^{(0)}$. The initial value for $\mathbf{R}^{(0)}$ is then obtained by solving Equation (9). After initialization, the latent factors are iteratively updated using the formula given in Equations (6), (8), and (10) until one of the the following two stopping conditions are met: (1) if the maximum number of iterations is reached, or (2) the value of the objective function does not change significantly.

C. Cross-scale Interactions (CSIs)

The CSIs identified by the MLMT framework are found by examining the regression coefficients that relate the local and regional predictors of the data, analogous to Equation (3). To illustrate this, we consider a variation of the multi-level modeling method given in Equation (1) by casting its formulation into the following optimization problem:

$$\min_{\mathbf{G},\mathbf{W}} \frac{1}{2}\sum_{i=1}^{R} \parallel \mathbf{y}_i - \mathbf{X}_i\mathbf{w}_i \parallel_2^2 + \frac{\rho_1}{2}\sum_{i=1}^{R} \parallel \mathbf{w}_i - \mathbf{G}^T\mathbf{z}_i \parallel_2^2 \tag{11}$$

In this relaxed multi-level modeling approach, the first term of the objective function penalizes the regression error for each region while the second term fits the regression coefficient to the regional predictors. Taking the partial derivative of the objective function with respect to W and setting it to zero yields the following solution:

$$\mathbf{w}_i = (\mathbf{X}_i^T\mathbf{X}_i + \rho_1\mathbf{I})^{-1}\mathbf{X}_i^T\mathbf{y}_i + \rho_1(\mathbf{X}_i^T\mathbf{X}_i + \rho_1\mathbf{I})^{-1}\mathbf{G}^T\mathbf{z}_i \tag{12}$$

Observe that the first term of the regression coefficient is equivalent to the solution for ridge regression using only the local predictor variables. The second term, on the other hand, is a correction factor due to the regional variables. Given a test example $(\mathbf{x}^*, \mathbf{z}_i)$ from region i, we can predict its response value as follows:

$$\hat{y} = \mathbf{x}^*\mathbf{w}_i = \mathbf{x}^*(\mathbf{X}_i^T\mathbf{X}_i + \rho_1\mathbf{I})^{-1}\mathbf{X}_i^T\mathbf{y}_i + \mathbf{x}^*\hat{\mathbf{G}}_i^T\mathbf{z}_i$$

where $\hat{\mathbf{G}}_i = \rho_1\mathbf{G}(\mathbf{X}_i^T\mathbf{X}_i + \rho_1\mathbf{I})^{-1}$. The predicted value can thus be decomposed into a local prediction term and a cross-scale interactions term involving $\mathbf{x}^*$ and $\mathbf{z}_i$. Therefore, $\hat{\mathbf{G}}_i$ is a modified CSI term for the multi-level modeling formulation given in Equation (11). Using the same strategy, the CSI term for MLMT is given by the following theorem.

Theorem 1: Let $\mathbf{U}$ be the latent factors associated with the local predictors and $\mathbf{R}$ be the latent factors associated with the regional predictors for the multi-level multi-task learning framework given in Equation (4). The CSI term for the formulation is

$$\bar{\mathbf{G}}_i = \rho_1 \mathbf{R}(\mathbf{U}^T \mathbf{X}_i^T \mathbf{X}_i \mathbf{U} + \rho_1 \mathbf{R}^T \mathbf{R})^{-1} \mathbf{U}^T. \qquad (13)$$

Proof: Ignoring the $L1$-regularization terms, the objective function can be re-written as follows:

$$\frac{1}{2}\sum_{i=1}^{r} \| \mathbf{y}_i - \mathbf{X}_i \mathbf{U} \mathbf{v}_i \|_2^2 + \frac{\rho_1}{2} \| \mathbf{Z}^T - \mathbf{R}\mathbf{V} \|_F^2$$

Taking the partial derivative of the objective function with respect to $\mathbf{V}$ and setting it to zero yields the following:

$$\mathbf{v}_i = \Big[(\mathbf{X}_i \mathbf{U})^T (\mathbf{X}_i \mathbf{U}) + \rho_1 \mathbf{R}^T \mathbf{R}\Big]^{-1} \Big[\mathbf{U}^T \mathbf{X}_i^T \mathbf{y}_i + \rho_1 \mathbf{R}^T \mathbf{z}_i\Big]$$

Thus, the predicted value for a test example $(\mathbf{x}^*, \mathbf{z}_i)$ can be computed as follows:

$$\hat{y} \;=\; \mathbf{x}^* \mathbf{U} \mathbf{v}_i = \mathbf{x}^* \mathbf{S} \mathbf{y}_i + \mathbf{x}^* \bar{\mathbf{G}}_i^T \mathbf{z}_i,$$

where $\mathbf{S} = \mathbf{U}\big[\mathbf{U}^T \mathbf{X}_i^T \mathbf{X}_i \mathbf{U} + \rho_1 \mathbf{R}^T \mathbf{R}\big]^{-1} \mathbf{U}^T \mathbf{X}_i^T$ and $\bar{\mathbf{G}}_i = \rho_1 \mathbf{R}(\mathbf{U}^T \mathbf{X}_i^T \mathbf{X}_i \mathbf{U} + \rho_1 \mathbf{R}^T \mathbf{R})^{-1} \mathbf{U}^T$. The first term corresponds to the value predicted using the local predictors only whereas the second term $\mathbf{x}^* \bar{\mathbf{G}}_i^T \mathbf{z}_i$ corresponds to the CSIs between the local and regional variables. ■

At first glance, it appears that the CSI term $\hat{\mathbf{G}}_i$ may vary from one region to another based on the covariance matrix $\mathbf{X}_i^T \mathbf{X}_i$. The varying $\hat{\mathbf{G}}_i$ is the result of our modeling assumption that ρ_1 is fixed for all the regions (see Equation (4)). As ρ_1 is related to the covariance structure of the noise levels for all the tasks, it can be calibrated separately for different regions to obtain a common $\mathbf{G}$. Alternatively, explicit calibration techniques for multi-task learning, such as those proposed in [12], [13] can be implemented for this purpose. However, such techniques require modification to the objective function, and thus, will be a subject for future research.

IV. Experimental Evaluation

We evaluate the performance of the proposed MLMT framework using nested datasets from the lake ecology domain.

A. Datasets

The lake water quality datasets used for our experiments were obtained from the LAGOS-NE database [7]. We selected four water quality metrics as response variables, including total phosphorus (TP), total nitrogen (TN), chlorophyll-a (chla) and Secchi depth (Secchi). The sampling years for the variables span from 2000 to 2013. For each lake, we extracted the sample data from the summer months of June, July, and August, and took their average values over all the sampling years to represent the true values for each response variable. We also selected 13 variables, including lake hydrogeomorphic variables and land cover/use data from 2001, as the local predictors. Ecological Drainage Units (EDUs) [14] were used to define the spatial regions of the study. We extracted 8 regional predictors, including the hydrogeomorphic and land cover/use variables measured at the coarser EDU-level. All the local and regional predictors are standardized to have zero mean and unit standard deviation while the response variables are log-transformed similar to the approach used in [15]. As shown in Table I, the number of instances (lakes) in each region (EDU) with ground truth data available varies from one response variable to another.

Table I: Summary statistics for 4 lake water quality data.

Response variable	TP	TN	Chla	Secchi
# regions (EDUs)	86	83	87	88
# instances (lakes)	4352	1946	5592	5796
# instances/region	1 - 369	1 - 236	1 - 575	1 - 583
mean value	37.58	739.25	17.19	2.78
standard deviation	66.75	1015.99	29.56	1.87

B. Experimental Setup

We compare the performance of our framework against the following four baseline methods:

- *Global-L*: This method trains a global, lasso regression model to the local predictors of the training data from all regions, while ignoring the regional predictors.
- *Global-LR*: This method trains a global, lasso regression model to fit both the local and regional predictors of the training data from all regions.
- *STL*: This method applies lasso regression independently to each region using only the local predictors of the regions.
- *MLM*: This method applies L1-regularization to the multi-level modeling formulation (see Equation (1)) to build a separate model for each region [10]. It assumes that the regression coefficients for each region are related to the regional variables via the CSI term, $\mathbf{G}$.

The source code for MLMT and other baselines are available at https://github.com/shuaiyuan-msu/csi-mlmt.

We employ two metrics to evaluate the performance of the different methods: root-mean square error (RMSE) and predicted R-squared. RMSE measures the deviation between the observed and predicted values, i.e., RMSE $= \sqrt{\sum_{i=1}^{N}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2/N}$, where $\hat{\mathbf{y}}_i$ is the predicted value and N is the number of predicted instances. The predicted R-squared measures the variance in the predicted values of the response variable explained by the model and is calculated as follows: $R^2 = 1 - \frac{\sum_i (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{\sum_i (\mathbf{y}_i - \bar{\mathbf{y}})^2}$, where $\bar{y}$ is the mean of the observed response variable values.

C. Experimental Results

1) **Performance Comparison for All Regions:** We partitioned each dataset into separate training and test sets, using $2/3$ of the data for training and the remaining $1/3$ for testing. We further divide the training set into two halves, one for training and the other for validation (hyperparameter tuning). We repeated this 10 times with different training and test partitions and reported the average and standard deviation of

Table II: Results for 4 lake water quality data.

	TP	TN	Chla	Secchi
GlobalL	0.330±0.003	0.231±0.007	0.426±0.013	0.274±0.019
GlobalLR	0.310±0.004	0.214±0.006	0.413±0.018	0.258±0.022
STL	0.564±0.475	0.546±0.033	0.529±0.195	0.260±0.015
MLM	0.302±0.005	0.210±0.006	0.423±0.044	0.242±0.011
MLMT	**0.286±0.004**	**0.203±0.006**	**0.381±0.014**	**0.231±0.010**

(a) RMSE results

	TP	TN	Chla	Secchi
GlobalL	0.414±0.009	0.515±0.018	0.359±0.031	0.351±0.093
GlobalLR	0.485±0.014	0.584±0.023	0.399±0.044	0.421±0.101
STL	0.095±0.075	0.011±0.035	0.056±0.033	0.275±0.021
MLM	0.511±0.016	0.599±0.020	0.364±0.140	0.494±0.046
MLMT	**0.560±0.010**	**0.624±0.024**	**0.489±0.030**	**0.540±0.044**

(b) R^2 results

RMSE and R^2 values in Table II. The results in this table suggest that single task learning (STL) performs the worst among all five competing methods on 3 of the 4 datasets. We also observe that global-L is worse than global-LR, which suggests the value of incorporating regional predictors into the predictive modeling framework. Nevertheless, the performances of the global models are inferior compared to the multi-level modeling (MLM) approach since both global-L and global-LR apply the same model to all the regions. Finally, the proposed MLMT framework consistently outperforms all the baseline methods on all four datasets.

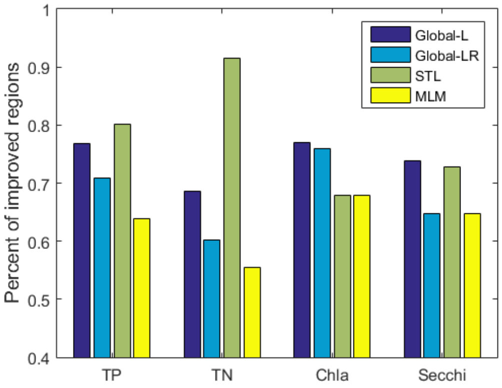
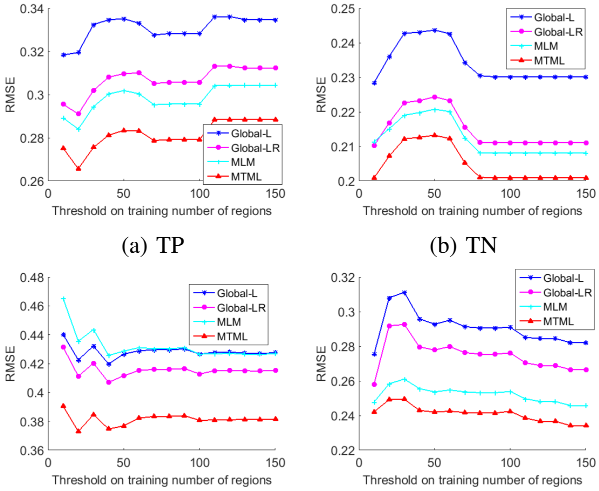


Figure 2: Percentage of regions in which MLMT performs better than baseline methods.

We also compare the number of regions in which MLMT outperforms the baseline methods. As can be seen from the results shown in Figure 2, MLMT outperforms all the baseline methods in more than 64% of the regions in 3 of the 4 datasets. The percentage increases to over 70% of the regions when compared against STL. For the TN dataset, which has fewer instances available, MLMT still performs better than MLM in more than 55% of the regions.

2) **Performance Comparison for Data-Poor Regions:** We also compare the performance of all the methods for regions with small training set sizes. To identify such regions, we define a maximum sample size threshold τ and calculate the RMSE values for the test examples located in regions that have less than τ training examples. We vary τ from 10 to 150 and plot the results in Figure 3. The results suggest that MLMT has consistently lower RMSE compared to all the baseline methods in the data-poor regions. This validates our assertion that the shared latent factors enable the data-poor regions to leverage information from other regions in order to construct more effective models.



(a) TP

(b) TN

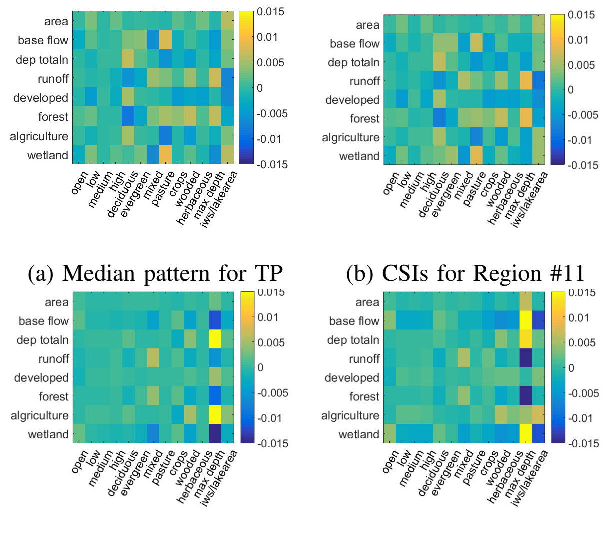
(c) Chla

(d) Secchi

Figure 3: Performance comparison for regions with limited number of training data.

3) **Cross-scale Interactions:** We examine the CSIs found by MLMT that contribute to the prediction of the lake water quality variables by visualizing the $\bar{\mathbf{G}}_i$ matrices given in Equation (13). We plot both median pattern and the pattern that is least correlated with the median pattern. As noted in Section III-C the variability observed in $\bar{\mathbf{G}}_i$ is due to the covariance matrices observed for different regions. For TP, there appears to be no significant difference between the median pattern (Figure 4(a)) and the region with least correlated CSI pattern (Figure 4(b)). All 86 regions appear to follow a similar CSI pattern, as evidenced by the high average correlation (0.991) between the $\bar{\mathbf{G}}$ matrices of all the regions. The median CSI pattern for TP can be compared against previous results reported in the literature. Our analysis showed that the interaction term in $\bar{\mathbf{G}}$ for regional agriculture and local wetland (wooded) is negative, which agrees with previous findings reported in [4], [6]. The CSI pattern suggests that when the proportion of agricultural land use in a region is low, the wetland-TP relationship is positive. In contrast, when the proportion of agricultural land use in a region is high, the wetland-TP relationship is negative. An explanation to this is that in regions with little agriculture, wetlands may be the source of phosphorus to lakes (positive slope), but when agriculture increases, wetland effects on lake phosphorus becomes negative since the wetlands may be retaining phosphorus from getting into lakes. The median pattern also indicates there are other potential CSIs (e.g., negative relationship between proportion of wetland cover and max depth) affecting TP in a region. These CSIs can be used to develop new hypotheses for scientists to further explore and validate.

The median CSI pattern for Secchi depth is shown in Figure 4(c). Many regions have CSI patterns that are very similar to



(a) Median pattern for TP (b) CSIs for Region #11

(c) Median pattern for Secchi (d) CSIs for Region #88

Figure 4: CSIs between local and regional predictors for the prediction of total phosphorous (a)-(b) and Secchi depth (c)-(d). The horizontal axis in each plot denotes local predictors while the vertical axis denotes regional predictors.

Table III: RMSE comparison for original and new regions.

Response variable	TP	TN	Chla	Secchi
Original regions (EDUs)	0.564	0.546	0.529	**0.260**
New regions	**0.403**	**0.519**	**0.487**	0.267

the median pattern. However, there are a few regions with CSI patterns that are considerably different than the median pattern. For example, Figure 4(d) shows the CSI pattern for region #88. For this region, some relationships such as those between regional base flow and local max depth and between regional wetland and local max depth have the opposite sign compared to the relationships shown by the median CSI pattern.

4) **Comparison Between the New and Original Regions:** Since the original regions (EDUs) were created for other purposes, we hypothesized that a better set of regions can be derived for predictive modeling using the latent factors associated with the lakes. To do this, we first compute the latent feature representation of each lake, which is given by $\mathbf{XU}$. We then apply k-means clustering to generate the new set of regions. For a fair comparison, we set the number of clusters to be the same as the number of regions (EDUs) in the original data. A lasso regression model is independently trained for each new region using only their local predictor variables. Similarly, we also train lasso regression models for each of the original regions. We then compare the performances of the models for the new regions against those for the original regions. Table III summarizes their RMSE values. The results in this table suggest that the local models trained on the new regions have a lower RMSE compared to the local models trained on the original (EDU) regions in 3 out of the 4 datasets. This supports our hypothesis that the new regions created by MLMT can be used to build more accurate local prediction models compared to the original regions.

V. Conclusions

This paper presents a novel framework called MLMT for modeling nested geospatial data. The framework jointly trains a set of models that can incorporate both the local and regional predictors into a unified formulation. We also show how cross-scale interactions can be derived using the proposed framework. Experimental results suggest that MLMT outperforms four other baseline methods on the lake water quality datasets evaluated in this study. The latent factors of MLMT can also be used to create a new set of regions for building more accurate local prediction models compared to the original regions that were defined a priori from the domain.

VI. Acknowledgements

This research was supported in part by the NSF under grant EF-1065786, EF-1638679, IIS-1615612 and IIS-1615597. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

[1] C. E. Fergus, A. O. Finley, P. A. Soranno, and T. Wagner, “Spatial variation in nutrient and water color effects on lake chlorophyll at macroscales,” *Plos one*, vol. 11, no. 10, p. e0164592, 2016.

[2] P. De Marco, J. A. F. Diniz-Filho, and L. M. Bini, “Spatial analysis improves species distribution modelling during range expansion,” *Biology Letters*, vol. 4, no. 5, pp. 577–580, 2008.

[3] J. Franklin, J. M. Serra-Diaz, A. D. Syphard, and H. M. Regan, “Global change and terrestrial plant community dynamics,” 2016.

[4] P. A. Soranno *et al.*, “Cross-scale interactions: quantifying multi-scaled cause–effect relationships in macrosystems,” *Frontiers in Ecology and the Environment*, vol. 12, no. 1, pp. 65–73, 2014.

[5] D. P. C. Peters, R. A. P. Sr., B. T. Bestelmeyer, C. D. Allen, S. Munson-McGee, and K. M. Havstad, “Cross-scale interactions, nonlinearities, and forecasting catastrophic events,” *Proc National Academy of Science*, vol. 101, no. 42, pp. 15 130–15 135, 2004.

[6] E. Fergus, P. Soranno, K. Cheruvelil, and M. T. Bremigan, “Multiscale landscape and wetland drivers of lake total phosphorus and water color,” *Limnology and Oceanography*, vol. 56, no. 6, pp. 2127–2146, 2011.

[7] P. A. Soranno *et al.*, “Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science through data reuse,” *Giga Science*, 2015.

[8] T. A. Snijders and R. J. Bosker, *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2012.

[9] L. Zhao, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting,” in *Proc of ACM SIGKDD*, 2016, pp. 2085–2094.

[10] A. C. Lozano and G. Swirszcz, “Multi-level lasso for sparse multi-task regression,” in *Proc of Int’l Conf on Machine Learning*, 2012.

[11] J. Liu, S. Ji, and J. Ye, “Multi-task feature learning via efficient l2, 1-norm minimization,” in *Proc of Conf. on Uncertainty in Artificial Intelligence*, 2009, pp. 339–348.

[12] H. Liu, L. Wang, and T. Zhao, “Calibrated multivariate regression with application to neural semantic basis discovery.” *Journal of Machine Learning Research*, vol. 16, pp. 1579–1606, 2015.

[13] P. Gong, J. Zhou, W. Fan, and J. Ye, “Efficient multi-task feature learning with calibration,” in *Proc of ACM SIGKDD*, 2014, pp. 761–770.

[14] J. V. Higgins, M. T. Bryer, M. L. Khoury, and T. W. Fitzhugh, “A freshwater classification approach for biodiversity conservation planning,” *Conservation Biology*, vol. 19, no. 2, pp. 432–445, 2005.

[15] T. Wagner, P. A. Soranno, K. E. Webster, and K. S. Cheruvelil, “Landscape drivers of regional variation in the relationship between total phosphorus and chlorophyll in lakes,” *Freshwater Biology*, vol. 56, no. 9, pp. 1811–1824, 2011.