A Hidden Markov Contour Tree Model for Spatial Structured Prediction

Arpan Man Sainju, Wenchong He, Zhe Jiang, *Member, IEEE,*

Abstract—Spatial structured models are predictive models that capture dependency structure between samples based on their locations in the space. Learning such models plays an important role in many geoscience applications such as water surface mapping, but it also poses significant challenges due to implicit dependency structure in continuous space and high computational costs. Existing models often assume that the dependency structure is based on either spatial proximity or network topology, and thus cannot incorporate complex dependency structure such as contour and flow direction on a 3D potential surface. To fill the gap, we recently proposed a novel spatial structured model called hidden Markov contour tree (HMCT), which generalizes the traditional hidden Markov model from a total order sequence to a partial order polytree. HMCT also advances existing work on hidden Markov trees through capturing complex contour structures on a 3D surface. We proposed efficient model construction and learning algorithms. This paper extends our initial HMCT model into a post-processor that can refine the classified results from other existing models. We analyzed the theoretical properties of the extended model. Evaluations on real-world flood mapping datasets show that HMCT outperforms multiple baseline methods in classification performance and the HMCT can also effectively enhance the results of other baseline methods. Computational experiments also show that HMCT is scalable to large data sizes (e.g., classifying millions of samples in seconds).

Index Terms—Hidden Markov Contour Tree; 3D Surface Classification; Spatial Structured Prediction

1 Introduction

Spatial structured models are predictive models that capture dependency structure between samples based on their locations. Given data samples in a spatial raster framework with explanatory feature layers and a potential field layer, as well as an independent set of training samples with class labels, the spatial structured learning problem aims to learn a model that can predict sample classes in the same framework [1]. For example, in flood extent mapping from earth imagery, data samples are imagery pixels in a regular grid, the explanatory feature layers are spectral bands, and the potential field can be elevation that controls water flow directions. The goal is to predict the classes (flood or dry) of pixels based on not only the spectral features but also the implicit flow directions on an elevation surface.

The problem is important in many societal applications such as flood extent mapping for disaster response and national water forecasting. Flood extent mapping plays a crucial role in addressing grand societal challenges such as disaster management, national water forecasting, as well as energy and food security. For example, during Hurricane Harvey floods in 2017, first responders needed to know where flood water was in order to plan rescue efforts. In national water forecasting, detailed flood extent maps can be used to calibrate and validate the NOAA National Water Model [2], which can forecast the flow of over 2.7 million rivers and streams through the entire continental U.S. [3]. In current practice, flood extent maps are mostly generated by flood forecasting models, whose accuracy is often unsatisfactory in high spatial details [3]. Other ways to generate flood maps involve sending a field crew on the ground to record high-water marks, or visually interpreting earth observation imagery [4]. However, the process is both expensive and time consuming. With the large amount of high-resolution earth imagery being collected from satellites (e.g., DigitalGlobe, Planet Labs), aerial planes (e.g., NOAA National Geodetic Survey), and unmanned aerial vehicles, the cost of manually labeling flood extent becomes prohibitive. Note that though we use flood mapping as a motivation example, the problem can potentially be applied to many other applications such as water quality monitoring and air pollution mapping in which pollutants are transmitted following flow directions [5], protein structure learning in biochemistry [6], and geometric shape analysis in computer graphics [7].

However, the problem poses several unique challenges that are not well addressed in traditional classification problems. First, implicit spatial dependency structure exists between pixel locations. For example, due to gravity, flood water tends to flow from one location to nearby lower locations. Such dependency structure is complex, following contour patterns on a 3D surface. Second, data contains rich noise and obstacles. For example, high-resolution earth imagery often has noise, clouds and shadows. In addition, the spectral features of image pixels can be insufficient to distinguish classes (also called class confusion) due to heterogeneity. For instance, pixels of tree canopies overlaying flood water have the same spectral features with those trees in dry areas, yet their classes are different. Finally, the problem is also computationally challenging due to the cost of modeling complex spatial structure on a large data volume (e.g., millions of sample locations).

Existing spatial structured models often assume that

- *Corresponding author: Zhe Jiang, zjiang@cs.ua.edu*
- *A.M. Sainju, W. He and Z. Jiang were with the Department of Computer Science, University of Alabama, Tuscaloosa, AL, 35487.*

Manuscript received April 19, 2005; revised August 26, 2015.

the dependency structure is based on spatial proximity or spatial network topology. Models based on spatial proximity assume that adjacent or nearby locations have stronger dependency. The assumption comes from the first law of geography [8], "everything is related to everything else, but near things are more related than distant things". Examples include Markov random field [9] , conditional random field [10], spatial network embedding [11], [12], [13], [14], [15], and convolutional neural networks [16]. Other works incorporate spatial structures by spatial regularization on loss function (to penalize difference between nearby locations) together with efficient optimization [17], [18], [19]. Models based on spatial network topology assume that dependency between sample locations only follows an underlying spatial network structure. These models are useful for samples that are only located along spatial networks (e.g., traffic accidents along road networks). Examples of such models include spatial network Kriging [9], [20] and spatial network autoregressive models [21]. In summary, related works often capture undirected spatial dependency structure based on distance or network topology. Complex dependency structure that is both directed and prevalent at all locations in the continuous space are largely unexplored. Recently, a geographical hidden Markov tree (HMT) model [22] has been proposed, which captures directed spatial dependency based on flow directions across all locations, but it is largely motivated by a one dimensional spatial view and does not consider complex spatial structures such as contours on a 3D surface.

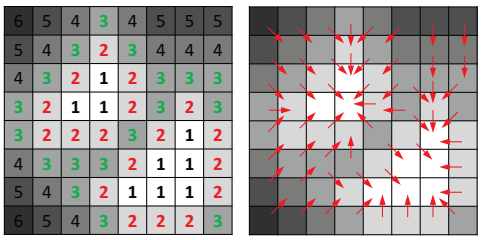
To fill the gap, we recently proposed a novel spatial structure model called hidden Markov contour tree (HMCT) [23]. It is a probabilistic graphical model that generalizes the common hidden Markov model (HMM) from a total order sequence to a partial order polytree. Specifically, the hidden class layer contains nodes (pixels) in a contour tree structure to reflect flow directions between all locations on a 3D surface. We also proposed efficient learning algorithms based on contour tree node collapsing and value pre-aggregation. Preliminary results showed that HMCT significantly outperforms several baseline methods on real-world flood mapping datasets and the proposed model is scalable to a large data volume [23]. This paper extends our preliminary results with the following additional contributions.

- We proposed an extension of the HMCT model from generative to discriminative so that the model can be used as a post-processor (HMCT-PP). We analyzed the theoretical properties of the extended model.
- We conducted evaluations on real-world flood mapping datasets. Results show that HMCT-PP can significantly enhance the results from baseline methods.
- We also conducted sensitivity analysis of the proposed approaches to initial parameters. We also interpreted the results through visualization and analysis.

2 Problem Statement

2.1 Preliminaries

Definition 1. *A* spatial raster framework *is a tessellation of a two dimensional plane into a regular grid of N cells. Spatial neighborhood relationship exists between cells based on cell adjacency. The framework can contain m non-spatial explanatory feature layers (e.g., spectral bands in earth imagery), one potential field layer (e.g., elevation), and one class layer (e.g.,* flood, dry*).*



(a) Elevation values with contours (b) Flow directions

Fig. 1: Example of raster framework and flow dependency

Definition 2. *Each cell in a raster framework is a* spatial data sample*, noted as $\mathbf{s_n} = (\mathbf{x}_n, \phi_n, y_n)$, where $n \in \mathbb{N}, 1 \leq n \leq N$, $\mathbf{x}_n \in \mathbb{R}^{m \times 1}$ is a vector of m non-spatial explanatory feature values with each element corresponding to one feature layer, $\phi_n \in \mathbb{R}$ is a cell's potential field value, and $y_n \in \{0, 1\}$ is a binary class label.*

A raster framework with all samples is noted as $\mathcal{F} = \{\mathbf{s_n} | n \in \mathbb{N}, 1 \leq n \leq N\}$, non-spatial explanatory features of all samples are noted as $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$, the potential field layer is noted as $\mathbf{\Phi} = [\phi_1, ..., \phi_N]^T$, and the class layer is noted as $\mathbf{Y} = [y_1, ..., y_N]^T$.

Definition 3. *A* contour *in a raster framework is a set of contiguous samples whose potential field values are equal. For example, in Figure 1(a), there are two contours for the elevation value* 1*. Note that our definition of contour here is discrete. Original definition based on a continuous field in topology can be found in [24], [25].*

Definition 4. *Spatial flow dependency exists between cells following the gradient on the potential field layer. Formally, a* flow dependency *$\mathbf{s}_i \rightsquigarrow \mathbf{s}_j$ exists if and only if there exist a sequence of neighboring (adjacent) cells $< \mathbf{s}_i, \mathbf{s}_{p_1}, \mathbf{s}_{p_2}, ..., \mathbf{s}_{p_l}, \mathbf{s}_j >$ such that $\phi_i \leq \phi_{p_1}$, $\phi_{p_l} \leq \phi_j$, and $\phi_{p_k} \leq \phi_{p_{k+1}}$ for any $1 \leq k \leq l - 1$. For example, due to gravity, flood water can flow from cells with elevation* 3 *to neighboring cells with elevation* 2 *in Figure 1(a).*

2.2 Formal problem definition

We now formally define the spatial structured learning problem.

Input:
- Spatial raster framework $\mathcal{F} = \{\mathbf{s}_n | n \in \mathbb{N}, 1 \leq n \leq N\}$
- Explanatory features of samples $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$
- Spatial potential field values of samples: $\mathbf{\Phi} = [\phi_1, ..., \phi_N]^T$
- Training samples $\{\mathbf{s}_k | k \in training\ set\}$ outside $\mathcal{F}$

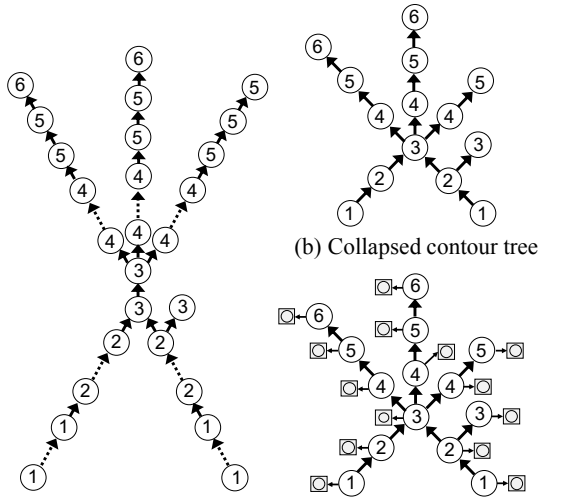
Output: A spatial structured model $f : \mathbf{Y} = f(\mathbf{X})$
Objective: minimize classification errors
Constraint:
- Explanatory feature layers contain noise and obstacles
- Sample classes follow partial order flow dependency
- Sample class is binary, $y_n \in \{0, 1\}$

3 Proposed Approach

The section introduces a novel spatial structured model called hidden Markov contour tree together with effective and efficient learning algorithms.

3.1 Modeling Spatial Contour Structure

Our goal is to explore a data structure to represent flow dependency between pixel locations on a potential field (e.g., 3D elevation surface). Such dependency structure can be represented by *contour tree* [25], [26]. A contour tree is a polytree (directed acyclic graph whose underlying undirected graph is a tree) that captures the evolution of contours (level sets) of a potential field. Figure 2(b) shows an example based on the elevation surface in Figure 1. Assume we increase an elevation threshold from 1 to 6. For threshold 1, two separate contours appear, corresponding to the two leaf nodes on the bottom of Figure 2(b). For threshold 2, the two contours both grow bigger but remain separated. For threshold 3, the two contours merge into one (corresponding to the central node with number 3), and a new separate contour appears based on the bottom right pixel in Figure 1(a) (corresponding to the tree node 3 on a side branch). For threshold 4, a contour of elevation 3 splits into three. Then, each contour grows separately as the threshold further increases. Contour tree is naturally a good representation for flow direction dependency between locations on an elevation surface. For example, in Figure 2(b), if the central node with elevation 3 is in the *flood* class, all its parent nodes (the ones with elevations 1 and 2 below it) must be in the *flood* class, since flood water flows from a high elevation to nearby lower elevations.



(a) Uncollapsed contour tree (b) Collapsed contour tree (c) Hidden Markov contour tree (circle in box is feature node set)

Fig. 2: Illustration of hidden Markov contour tree

Existing research on contour tree often focuses on the critical nodes (such as local minimum and maximum, saddle point) [24], [25]. Tree construction algorithms often require an input elevation surface to be represented as a mesh with unique values [26] . In order to run the algorithms on an elevation surface with duplicated values, we can use perturbation to enforce an arbitrary order on locations with an equal elevation. This makes the tree unnecessarily large, as shown in Figure 2(a). To address this limitation, we propose a collapsed contour tree, which starts from a contour tree from existing algorithm but then collapses nodes in the same contour. Figure 2(a-b) shows an example. For instance, two connected nodes with elevation 3 in Figure 2(a) are merged into one in Figure 2(b). Node collapsing can dramatically reduce the size of a contour tree, potentially reducing the cost of learning and inference algorithms. Algorithm 1 shows the detailed steps for node collapsing. The main idea is to use breadth-first search to find each collapsable contour component, and then to collapse tree nodes within that component. The time complexity is $O(N)$ where N is the total number of nodes in the uncollapsed contour tree.

Algorithm 1 Collapse Contour Tree

Input:
- An uncollapsed contour tree

Output:
- A collapsed contour tree

Initialize all nodes as *unvisited*
for each tree node $\mathbf{s_n}$ by topological order **do**
  if $\mathbf{s_n}$ is *unvisited* **then**
    Mark $\mathbf{s_n}$ as *visited*
    Breadth First Search (BFS) from $\mathbf{s_n}$ to its contour
    Collapse BFS traversed nodes, mark them *visited*
return the collapsed contour tree

3.2 Hidden Markov Contour Tree (HMCT)

We propose a hidden Markov contour tree (HMCT), a probabilistic graphical model that generalizes the common hidden Markov model from a total order sequence to a partial order polytree. A HMCT model consists of two layers: a hidden class layer in the form of a collapsed contour tree, and an observation feature layer. Each contour tree node in the hidden class layer represents a same unknown hidden class shared over all pixels in the contour. Each contour tree edge represents the class transitional probability between pixels from two contours based on flow dependency. A hidden class node in contour tree is connected to a set of observed feature nodes (circle in boxes in Figure 2(c)), which correspond to the explanatory feature vectors of all pixels in that contour.

The joint distribution of all samples' features and classes can be formulated as Equation 1, where $\mathcal{P}_n$ is the set of parent samples of the nth sample in the dependency tree ($\mathcal{P}_n = \emptyset$ for a leaf node), and $y_{k\in\mathcal{P}_n} \equiv \{y_k | k \in \mathcal{P}_n\}$ is the set of parent classes of node n. $\mathbf{x}_n = \{\mathbf{x}_{n_i} | 1 \le i \le N_n\}$ is the set of feature nodes corresponding to hidden class node n (i.e., $\mathbf{x}_{n_i}$ is the feature vector for a pixel n_i in the nth contour). Note that our notation of $\mathbf{x}_n$ here represents a set, and should not be confused with notations in Section 2.

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y}) = \prod_{n=1}^{N}\prod_{i=1}^{N_n} P(\mathbf{x}_{n_i}|y_n) \prod_{n=1}^{N} P(y_n|y_{k\in\mathcal{P}_n}) \tag{1}$$

For simplicity, observed feature nodes are assumed conditionally independent given their hidden class node, following an i.i.d. Gaussian distribution, as shown in Equation 2,

where $\boldsymbol{\mu}_{y_n}$ and $\boldsymbol{\Sigma}_{y_n}$ are the mean and covariance matrix of feature vector $\mathbf{x}_{n_i}$ for class y_n ($y_n = 0, 1$).

$$P(\mathbf{x}_{n_i}|y_n) \sim \mathcal{N}(\boldsymbol{\mu}_{y_n}, \boldsymbol{\Sigma}_{y_n}) \tag{2}$$

Class transitional probability follows the partial order flow dependency constraint. For example, due to gravity, if any parent's class is *dry*, the child's class must be *dry*; if all parents' classes are *flood*, then the child has a high probability of being *flood* due to spatial autocorrelation. Consider *flood* as the positive class (class 1) and *dry* as the negative class (class 0), the product of all parents' classes is $y_{\mathcal{P}_n} \equiv \prod_{k \in \mathcal{P}_n} y_k$. Class transitional probability can be modeled as Table 1, where ρ is a parameter close to 1. The prior class probability of a node without parents is also shown on the right of Table 1, where π is another parameter.

TABLE 1: Class transition probability and prior probability

$P(y_n\|y_{\mathcal{P}_n})$	$y_{\mathcal{P}_n}=0$	$y_{\mathcal{P}_n}=1$
$y_n=0$	1	$1-\rho$
$y_n=1$	0	ρ

	$P(y_n)$
$y_n=0$	$1-\pi$
$y_n=1$	π

3.3 HMCT Learning and Inference

The parameters of hidden Markov contour tree include the mean and covariance matrix of sample features in each class, prior probability of leaf node classes, and class transition probability for non-leaf nodes. We denote the entire set of parameters as $\boldsymbol{\Theta} = \{\rho, \pi, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 0, 1\}$. Learning the set of parameters poses two major challenges: first, there exist unknown hidden class variables $\mathbf{Y} = [y_1, ..., y_N]^T$, which are non-i.i.d.; second, the number of node variables can be large.

To address these challenges, we propose to use the expectation-maximization (EM) algorithm and message (belief) propagation. Our EM-based approach has the following major steps:

(a) Initialize parameter set $\boldsymbol{\Theta}_0$
(b) Compute posterior distribution of hidden classes: $P(\mathbf{Y}|\boldsymbol{X}, \boldsymbol{\Theta_0})$
(c) Compute posterior expectation of log likelihood: $LL(\boldsymbol{\Theta}) = \mathbb{E}_{\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta_0}} \log P(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta})$
(d) Update parameters: $\boldsymbol{\Theta_0} \leftarrow \arg\max_{\boldsymbol{\Theta}} LL(\boldsymbol{\Theta})$
Return $\boldsymbol{\Theta_0}$ if it's converged, otherwise goto (b)

Algorithm 2 shows the details. First, we initialize parameters either with random values within reasonable range (for ρ and π) or with initial estimates based on training samples (e.g., the mean and covariance of features in each class for $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$). Then we conduct a breadth-first search on the tree from any start node as a root. After this, the algorithm starts the iteration till parameters converge. In each iteration, it propagates messages first from leaves to root (steps 6-7) and then from root to leaves (steps 8-9). Marginal posterior distribution of node classes are then computed (steps 10-11). Based on this, the algorithm updates parameters (step 12). Message propagation is based on the sum and product algorithm [27], [28]. Propagation of message along nodes in a graph (or tree) is equivalent to marginalizing out node variables in the overall joint distribution in Equation 1.

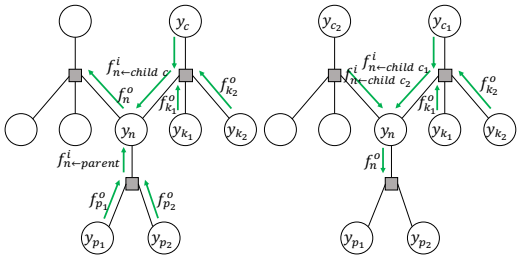
Algorithm 2 EM Algorithm for Hidden Markov Tree

Input:
- $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N]^T$: cell sample feature matrix
- $\mathcal{T}$: a contour tree for spatial dependency
- ϵ: parameter convergence threshold

Output:
- $\boldsymbol{\Theta} = \{\rho, \pi, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 0, 1\}$: set of model parameters

1: Initialize $\boldsymbol{\Theta_0}$, $\boldsymbol{\Theta}$
2: Find a node n_0 in $\mathcal{T}$ as the root
3: Do a breadth-first search (BFS) on $\mathcal{T}$ from root n_0
4: **while** $\|\boldsymbol{\Theta_0} - \boldsymbol{\Theta}\|_\infty > \epsilon$ **do**
5: $\quad\boldsymbol{\Theta_0} \leftarrow \boldsymbol{\Theta}$
6: $\quad$**for each** y_n from leaves to root in BFS **do**
7: $\quad\quad$Compute messages $f_n^i(y_n), f_n^o(y_n)$ by (3)-(6)
8: $\quad$**for each** y_n from root to leaves in BFS **do**
9: $\quad\quad$Compute messages $g_n^i(y_n), g_n^o(y_n)$ by (7)-(12)
10: $\quad$**for each** $y_n, 1 \le n \le N$ **do**
11: $\quad\quad$// Compute marginal distributions: $P(y_n|\mathbf{X}, \boldsymbol{\Theta_0}), P(y_n, y_{k \in \mathcal{P}_n}|\mathbf{X}, \boldsymbol{\Theta_0})$
12: $\quad$Update $\boldsymbol{\Theta}$ based on marginal distributions: $\boldsymbol{\Theta} \leftarrow \arg\max_{\boldsymbol{\Theta}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta_0}} \log P(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta})$ by (13)-(16)
13: **return** Θ

Figure 3 illustrates the forward message propagation process from leaves to root (denoted as f). Each node can have only one outgoing message (denoted as $f_n^o(y_n)$, but can have multiple incoming messages (incoming messages from a child c and the parent side are denoted as $f^i_{n \leftarrow child\ c}(y_n)$ and $f^i_{n \leftarrow parent}(y_n)$ respectively). As illustrated in Figure 3(a), computing $f^i_{n \leftarrow child\ c}(y_n)$ involves integration (sum) over the product of outgoing messages from the child c and c's other parents, together with the conditional probability between c and c's all parents, as specified in (3). The outgoing message from node n has two cases: to its child c or to its parents. In the first case, we first need to compute another incoming message from parents $f^i_{n \leftarrow parent}(y_n)$ based on (4). Then the outgoing message to a child c_0 can be computed based on (5). These are also illustrated in Figure 3(a). In the second case, we only need to compute outgoing message to parent based on (6) (also illustrated in Figure 3(b)). Note that we denote $P(\mathbf{x}_n|y_n) \equiv \prod_{i=1}^{N_n} P(\mathbf{x}_{n_i}|y_n)$ in these equations for brevity.



(a) Case 1: f out to child c_1 (b) Case 2: f out to parent

Fig. 3: Illustration of message propagation leaves to root

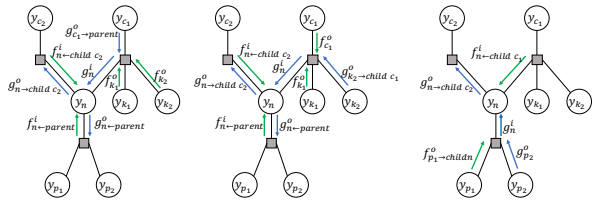
$$f^i_{n \leftarrow child\ c}(y_n) = \sum_{y_c, y_{\{k \in \mathcal{P}_c, k \neq n\}}} P(y_c|y_{k \in \mathcal{P}_c}) f_c^o(y_c) \prod_{k \in \mathcal{P}_c, k \neq n} f_k^o(y_k) \tag{3}$$

$$f^i_{n\leftarrow parent}(y_n) = \begin{cases} P(y_n) & \text{if } n \text{ no parent} \\ \sum\limits_{y_{k\in\mathcal{P}_n}} P(y_n|y_{k\in\mathcal{P}_n}) \prod\limits_{k\in\mathcal{P}_n} f^o_k(y_k) & \text{otherwise} \end{cases} \tag{4}$$

$$f^o_n(y_n) = P(\mathbf{x}_n|y_n) f^i_{n\leftarrow parent}(y_n) \prod_{c\in\mathcal{C}_n, c\neq c_0} f^i_{n\leftarrow child\ c}(y_n) \tag{5}$$

$$f^o_n(y_n) = P(\mathbf{x}_n|y_n) \prod_{c\in\mathcal{C}_n} f^i_{n\leftarrow child\ c}(y_n) \tag{6}$$

Backward message propagation from root to leaves also follows a recursive process. Each node only has one incoming message but several outgoing messages. There are three cases as shown in Figure 4(a-c): incoming message $g^i_n(y_n)$ can be from a child, a child's parent, or a parent. In the first two cases, $g^i_n(y_n)$ is computed from outgoing messages on the child side based on (7) and (8) respectively. The outgoing messages to another child or parent can be computed based on (9) and (10) for both of the first two cases. In the third case, $g^i_n(y_n)$ is computed based on outgoing messages from the parent side, as specified in (11). The outgoing message to a child can be computed based on (12). Since in the third case, the incoming message is from the parent side, we do not need to compute outgoing message to parent $g^o_{n\rightarrow parent}(y_n)$.



(a) Case 1: g from child (b) Case 2: g from child's parent (c) Case 3: g from parent

Fig. 4: Illustration of message propagation leaves to root

$$g^i_n(y_n) = \sum_{y_{c_0}, y_{p\in\mathcal{P}_{c_0}, p\neq n}} g^o_{c_0\rightarrow parent}(y_{c_0}) P(y_{c_0}|y_{p\in\mathcal{P}_{c_0}}) \cdot \prod_{p\in\mathcal{P}_{c_0}, p\neq n} f^o_p(y_p) \tag{7}$$

$$g^i_n(y_n) = \sum_{y_{c_0}, y_{\{p\in\mathcal{P}_{c_0}, p\neq n\}}} g^o_{p_{c_0}\rightarrow child\ c_0}(y_{p_{c_0}}) f^o_{c_0}(y_{c_0}) P(y_{c_0}|y_{k\in\mathcal{P}_{c_0}}) \prod_{p\in\mathcal{P}_{c_0}, p\neq p_{c_0}, p\neq n} f^o_p(y_p) \tag{8}$$

$$g^o_{n\rightarrow child\ c}(y_n) = g^i_n(y_n) P(\mathbf{x}_n|y_n) f^i_{n\leftarrow parent}(y_n) \prod_{c'\in\mathcal{C}_n, c'\neq c_0, c'\neq c} f^i_{n\leftarrow child\ c'}(y_n) \tag{9}$$

$$g^o_{n\rightarrow parent}(y_n) = g^i_n(y_n) P(\mathbf{x}_n|y_n) \prod_{c\in\mathcal{C}_n, c\neq c_0} f^i_{n\leftarrow child\ c}(y_n) \tag{10}$$

$$g^i_n(y_n) = \sum_{y_{k\in\mathcal{P}_n}} P(y_n|y_{k\in\mathcal{P}_n}) g^o_{p_0\rightarrow n}(y_{p_0}) \prod_{p\in\mathcal{P}_n} f^o_p(y_p) \tag{11}$$

$$g^o_{n\rightarrow child\ c}(y_n) = g^i_n(y_n) P(\mathbf{x}_n|y_n) \prod_{c'\in\mathcal{C}_n, c'\neq c} f^i_{n\leftarrow child\ c'}(y_n) \tag{12}$$

After both forward and backward message propagation, we can compute marginal posterior distribution of hidden class variables. The unnormalized marginal posterior distribution of the class of a leaf node (without parents), as well as the class of a non-leaf node (with parents) can be computed by multiplying all messages coming into these nodes. We do not provide details due to space limit. Results can be normalized by dividing them over the total sum of all variable configurations.

After computing the marginal posterior distribution, we can update model parameters by maximizing the posterior expectation of log likelihood (the maximization or M step in EM). Taking the marginal posterior distributions computed above together with the prior and transitional probabilities in Table 1 into the posterior expectation, we can get the following parameter update formulas (the M step in EM).

$$\rho = \frac{\sum\limits_{n|\mathcal{P}_n\neq\emptyset} \sum\limits_{y_n} \sum\limits_{y_{\mathcal{P}_n}} y_{\mathcal{P}_n} y_n P(y_n, y_{\mathcal{P}_n}|\mathbf{X}, \mathbf{\Theta_0})}{\sum\limits_{n|\mathcal{P}_n\neq\emptyset} \sum\limits_{y_n} \sum\limits_{y_{\mathcal{P}_n}} y_{\mathcal{P}_n} P(y_n, y_{\mathcal{P}_n}|\mathbf{X}, \mathbf{\Theta_0})} \tag{13}$$

$$\pi = \frac{\sum\limits_{n|\mathcal{P}_n=\emptyset} \sum\limits_{y_n} y_n P(y_n|\mathbf{X}, \mathbf{\Theta_0})}{\sum\limits_{n|\mathcal{P}_n=\emptyset} \sum\limits_{y_n} P(y_n|\mathbf{X}, \mathbf{\Theta_0})} \tag{14}$$

$$\boldsymbol{\mu}_c = \frac{\sum\limits_{n=1}^{N} \sum\limits_{i=1}^{N_n} \mathbf{x}_{n_i} P(y_n = c|\mathbf{X}, \mathbf{\Theta_0})}{\sum\limits_{n=1}^{N} \sum\limits_{i=1}^{N_n} P(y_n = c|\mathbf{X}, \mathbf{\Theta_0})}, c = 0, 1 \tag{15}$$

$$\mathbf{\Sigma}_c = \frac{\sum\limits_{n=1}^{N} \sum\limits_{i=1}^{N_n} (\mathbf{x}_{n_i} - \boldsymbol{\mu}_c)(\mathbf{x}_{n_i} - \boldsymbol{\mu}_c)^T P(y_n = c|\mathbf{X}, \mathbf{\Theta_0})}{\sum\limits_{n=1}^{N} \sum\limits_{i=1}^{N_n} P(y_n = c|\mathbf{X}, \mathbf{\Theta_0})}, c = 0, 1 \tag{16}$$

Class inference: After learning model parameters, we can infer hidden class variables by maximizing the overall probability. A naive approach that enumerate all combinations of class assignment is infeasible due to the exponential cost. We use a dynamic programming based method called *max-sum* [29]. The process is similar to the sum and product algorithm above. The main difference is that instead of using sum operation, we need to use max operation in message propagation, and also memorize the optimal variable values. We omit the details due to space limit.

3.4 Computational Performance Tuning

Time complexity of Algorithm 2: The cost includes contour tree traversal, message calculation, marginal probability calculation, and parameter update. Tree traversal cost is $O(N_T)$ where N_T is the number of collapsed contour tree nodes. Message calculation for the nth contour tree node is $O(N_n 2^d)$, where N_n is the number of feature nodes for tree node n, and d is the in-degree. The exponential term 2^d comes from the sum of products on all parent combinations. The term N_n comes from the aggregation over individual feature nodes in $P(\mathbf{x}_n|y_n) \equiv \prod_{i=1}^{N_n} P(\mathbf{x}_{n_i}|y_n)$ as well as in updating $\boldsymbol{\mu}_c$ and $\mathbf{\Sigma}_c$. Thus, the cost per iteration is $O(N2^d)$ where N is the total number of feature nodes or pixels ($N \gg N_T$). The overall cost is $O(N2^d I)$ where I is the number of iterations.

From the analysis above, we can see that the total cost is linear to the total number of feature nodes N instead of the number of collapsed contour tree nodes N_T, due

to the need of aggregating over individual feature nodes. To improve computational efficiency, we propose to **pre-aggregate** feature node values for each contour tree node (class node) once, and later to use the pre-aggregated results in message calculation. In this way, the remaining cost after pre-aggregation is $O(N_T)$ instead of $O(N)$. Pre-aggregation can be done during contour tree node collapsing.

The next question is how to do the pre-aggregation correctly (returning the same final results). In fact, there are only three parts involving aggregation over feature nodes: calculating $P(\mathbf{x}_n|y_n)$ in Equation 17, updating $\boldsymbol{\mu}_c$ in Equation 15, and updating $\boldsymbol{\Sigma}_c$ in Equation 16. Our goal here is to compute these three parts based on pre-aggregated feature nodes over each class node ($\mathbf{x}_{n_i}$ with $1 \leq n_i \leq N_n$). The specific process is described in Theorem 1.

$$\log P(\mathbf{x}_n|y_n) = \log \prod_{i=1}^{N_n} P(\mathbf{x}_{n_i}|y_n) = -\frac{N_n}{2}(m\log(2\pi) + \log|\Sigma_{y_n}|) - \frac{1}{2}\sum_{i=1}^{N_n}(\mathbf{x}_{n_i} - \boldsymbol{\mu}_c)^T \Sigma_{y_n}^{-1}(\mathbf{x}_{n_i} - \boldsymbol{\mu}_c) \tag{17}$$

Theorem 1. *Assume* $A_n = \sum_{i=1}^{N_n} \mathbf{x}_{n_i}\mathbf{x}_{n_i}^T$, $B_n = \sum_{i=1}^{N_n} \mathbf{x}_{n_i}$, $C_n = N_n$. *We can calculate Equation 15, Equation 16 and Equation 17 based on pre-aggregated results* A_n, B_n *and* C_n *on each class node* n.

Proof. Equation 18, Equation 19, and Equation 20 show how to calculate Equation 15, Equation 16 and Equation 17 based on pre-aggregated values A_n, B_n and C_n. Note that $\langle \cdot, \cdot \rangle$ is inner product between two matrices. We omit details due to space limit.

$$\sum_{i=1}^{N_n} \mathbf{x}_{n_i} = B_n \tag{18}$$

$$\sum_{i=1}^{N_n} (\mathbf{x}_{n_i} - \boldsymbol{\mu}_c)(\mathbf{x}_{n_i} - \boldsymbol{\mu}_c)^T = A_n - B_n\boldsymbol{\mu}_c^T - (B_n\boldsymbol{\mu}_c^T)^T + C_n\boldsymbol{\mu}_c\boldsymbol{\mu}_c^T \tag{19}$$

$$\sum_{i=1}^{N_n} (\mathbf{x}_{n_i} - \boldsymbol{\mu}_{y_n})^T \Sigma_{y_n}^{-1}(\mathbf{x}_{n_i} - \boldsymbol{\mu}_{y_n}) = \langle \Sigma_{y_n}^{-1}, A_n - B_n\boldsymbol{\mu}_{y_n}^T - (B_n\boldsymbol{\mu}_{y_n}^T)^T + N_n\boldsymbol{\mu}_{y_n}\boldsymbol{\mu}_{y_n}^T \rangle \tag{20}$$

□

Time complexity of refined algorithm: Pre-aggregation part (i.e., computing A_n, B_n, and C_n for all contour tree node n) is $O(N)$, but this can be done for only once during contour tree construction and node collapsing. After pre-aggregation, the time complexity for learning and inference part is reduced from $O(2^d N \cdot I)$ to $O(2^d N_T \cdot I)$ where d is node in-degree. Since N_T is the number of contour tree nodes after node collapsing, it is significantly smaller than the total number of pixels N.

3.5 Using HMCT for Post-processing

The original hidden Markov contour tree model is a generative model based on the joint probability distribution of both sample classes and features. Sample feature distribution is conditioned on hidden class nodes (i.e., $P(\mathbf{x}_n|y_n)$).

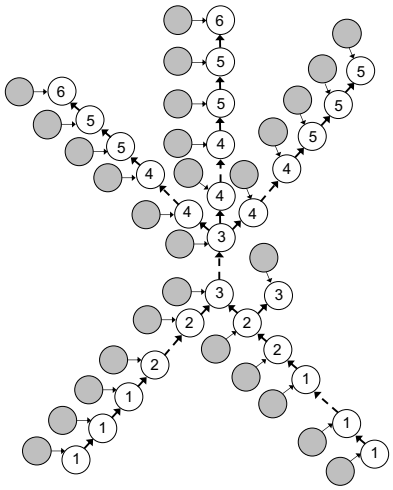


Fig. 5: An illustrative example of the discriminative version of hidden Markov contour tree

Sometimes in practice, we are given initial class probabilities of test samples predicted based on their features (i.e., $P(y_n|\mathbf{x}_n)$) from an existing classifier (without knowing about details in their feature values or training data). Such initial class probabilities can be erroneous due to feature noise and obstacles. In this case, we still hope to incorporate the similar structural dependency between class labels like before but cannot directly apply the HMCT model. Thus, we extend the HMCT model into a post-processing version called **HMCT-PP**. The input of HMCT-PP include initial class probabilities predicted by an existing classifier based on sample features as well as the dependency tree structure between samples, the output is refined class probability after incorporating the structural dependency between classes in the tree. Since HMCT-PP directly models sample classes conditioned on features, the model becomes discriminative. We use the uncollapsed version of contour tree in HMCT-PP because the initial class probability is provided for each individual sample pixel.

3.5.1 Model Structure

An example of the new model architecture (HMCT-PP) is shown in Figure 5. In the figure, white nodes represent hidden classes of pixels and gray nodes represent the corresponding feature values. We can see that the same flow dependency structure exists between hidden class nodes (except that the nodes are uncollapsed), but arrow directions between feature nodes and hidden class nodes are reversed to make the model discriminative. In other words, the hidden class layer is now conditioned on the feature layer. The arrows from a feature node to its corresponding hidden class note indicates the class probability of a sample (pixel) based on relevant feature values. Such a design of the new model structure is to allow the model to input initial class probabilities based on feature values ($P(y_n|\mathbf{x}_n)$). Note that in this example, the hidden class node y_n of a sample is conditioned on only its own features $\mathbf{x}_n$ for simplicity. The idea can be extended to allow initial input class probabilities of samples based on spatial contextual features such as those from the U-Net model. The arrows between class

nodes in the hidden class layer enforce structural dependency between sample classes. The reason why we do not use *collapsed* contour tree in the hidden class layer is that every individual sample may have a unique initial class probability based on relevant feature values in the model input. Even though two samples are on the same contour segment (i.e., their class nodes could have been collapsed in a collapsed contour tree), their initial class probabilities may differ. Using the uncollapsed contour tree in the hidden class layer of HMCT-PP provides flexibility to allow such input cases. It is worth noting that though the model has feature nodes of all samples in the overall structure, the model may not need the actual feature values in the input for class inference, since the initial class probability $P(y_n|\mathbf{x}_n)$ implicitly contains feature information.

3.5.2 Probabilistic Formulation

We now introduce the probabilistic formulation of the revised discriminative model. Based on the conditional independence assumption captured by the model tree structure in Figure 5, the overall conditional probability of all sample classes given all sample features $P(\mathbf{Y}|\mathbf{X})$ can be decomposed into local factors as shown in Equation 21, where $P(y_n|y_{\mathcal{P}_n}, \mathbf{X})$ represents the conditional class probability of a sample given its relevant features and parents' classes. For a sample without parents (i.e., leaf nodes in the contour tree), $y_{\mathcal{P}_n} = \emptyset$. If the input initial class probabilities of samples are produced by an i.i.d. classifier (i.e., the initial class probability of a sample is based on their its own feature), then we can assume $P(y_n|y_{\mathcal{P}_n}, \mathbf{X}) = P(y_n|y_{\mathcal{P}_n}, \mathbf{x}_n)$. Otherwise, if the input initial class probabilities of samples are produced by a spatial contextual classifier, we can assume that $P(y_n|y_{\mathcal{P}_n}, \mathbf{X}) = P(y_n|y_{\mathcal{P}_n}, \mathbf{x}_{S(n)})$, where $\mathbf{x}_{S(n)}$ is the spatial contextual features for sample n. For example, for the U-Net model, the spatial context for a pixel is all the pixels in the input image chunk. However, we still need to estimate the local factors, which we will discuss next.

$$P(\mathbf{Y}|\mathbf{X}) = \prod_n P(y_n|y_{\mathcal{P}_n}, \mathbf{X}) \tag{21}$$

We first discuss how to estimate the local factors in the case that the input class probabilities of pixels are provided by an i.i.d. classifier, i.e., the input class probability of each pixel is based on its own feature ($P(y_n|\mathbf{X}) = P(y_n|\mathbf{x}_n)$). Many i.i.d. classifiers such as decision tree, random forest, and gradient boosted model fall into this category. This also corresponds to the example of model structure in Figure 5. The corresponding estimation of local factors in Equation 21 is shown in Equation 22. The deductions in the second row and the third row are based on Bayes' theorem and the product rule. In the third row, the local factor $\prod_n P(y_n|y_{\mathcal{P}_n}, \mathbf{x}_n)$ is expressed by $P(y_n|\mathbf{x}_n)$, $P(y_{\mathcal{P}_n}|y_n)$ and $P(y_{\mathcal{P}_n}|\mathbf{x}_n)$. In the fourth row, we make an approximation that $P(y_{\mathcal{P}_n}|\mathbf{x}_n) = P(y_{\mathcal{P}_n}$, i.e., the classes of parent nodes $y_{\mathcal{P}_n}$ are independent from the feature of the current node $\mathbf{x}_n$. This approximation is reasonable based on the model structure in Figure 5. In the fifth row, we use Bayes' theorem again to express $P(y_{\mathcal{P}_n}|y_n)$ by $P(y_n|y_{\mathcal{P}_n})$ and $P(y_n)$. Finally, we cancel out the term $P(y_{\mathcal{P}_n})$ in both numerator and denominator and get the expression in the last row. In this expression, the local factor is expressed by the class probability of a sample conditioned on its feature $P(y_n|\mathbf{x}_n)$, class transitional probability from physical constraints $P(y_n|y_{\mathcal{P}_n})$ and the marginal class probability of a sample $P(y_n)$. Among these factors, $P(y_n|\mathbf{x}_n)$ is given as inputs from an existing classifier, e.g., decision tree, random forest or gradient boosted tree. $P(y_n|y_{\mathcal{P}_n})$ can be computed by a hyper-parameter ρ as in the original HMCT model. Based on the physical constraint, we can set ρ close to 1 (e.g., $\rho = 0.99999$). (Note that for leaf nodes without parents, the factor $P(y_n|y_{\mathcal{P}_n})$ in the numerator cancels out with factor $P(y_n)$ in the denominator.) For marginal class probability $P(y_n)$, it is hard to estimate for all individual samples. We make a further approximation that $P(y_n)$ is specified by a fixed parameter for all samples, i.e., $P(y_n = 1) = \pi$, where π can be set as 0.5. In the final expression, the term $P(y_n|\mathbf{x}_n)$ incorporates the local class probability based on a sample's feature itself, and the term $\frac{P(y_n|y_{\mathcal{P}_n})}{P(y_n)}$ incorporates the effect of adding structural dependency on a node's class probability. In this way, the class label of one node is determined not only by the local feature information itself but also the class labels at other locations based on the physical constrains.

$$\begin{aligned}
&\prod_n P(y_n|y_{\mathcal{P}_n}, \mathbf{x}_n) \\
&= \prod_n \frac{P(y_n, y_{\mathcal{P}_n}|\mathbf{x}_n)}{P(y_{\mathcal{P}_n}|\mathbf{x}_n)} \\
&= \prod_n \frac{P(y_n|\mathbf{x}_n)P(y_{\mathcal{P}_n}|y_n)}{P(y_{\mathcal{P}_n}|\mathbf{x}_n)} \\
&\approx \prod_n \frac{P(y_n|\mathbf{x}_n)P(y_{\mathcal{P}_n}|y_n)}{P(y_{\mathcal{P}_n})} \\
&= \prod_n \frac{P(y_n|\mathbf{x}_n)P(y_{\mathcal{P}_n})P(y_n|y_{\mathcal{P}_n})}{P(y_{\mathcal{P}_n})P(y_n)} \\
&= \prod_n \frac{P(y_n|\mathbf{x}_n)P(y_n|y_{\mathcal{P}_n})}{P(y_n)}
\end{aligned} \tag{22}$$

We now discuss the case that the input class probabilities of pixels are provided by a spatial contextual classifier such as the deep learning model U-Net [30]. In this case, $P(y_n|\mathbf{X}) = P(y_n|\mathbf{x}_{S(n)})$, where $S(n)$ is the set of contextual pixels being used in predicting the class $P(y_n)$. In general, we can follow the same deduction as in the case above by replacing $\mathbf{x}_n$ by $\mathbf{x}_{S(n)}$. The only difference is that the approximation in the fourth row is less accurate in theory now since the parents' classes may be dependent on the current sample's feature in a spatial contextual classifier. But the final expression is still very intuitive. The term $P(y_n|\mathbf{x}_{S(n)})$ reflects the initial class probability. The term $\frac{P(y_n|y_{\mathcal{P}_n})}{P(y_n)}$ reflects the effect of adding structural dependency on a node's class probability.

3.5.3 Class Inference

The class inference can be done based on the probabilistic formulation of the joint conditional distribution $P(\mathbf{Y}|\mathbf{X})$ as expressed in Equation 22. The problem is non-trivial due to the combinatorics of all possible classes of tree nodes. We use the same message propagation method called *max-sum* [29] as in HMCT. The main idea is to use dynamic

programming to only memorize the current maximum class assignment along a tree traversal order. The only difference from HMCT is that the messages are defined differently due to the change of the terms in the local factors. We do not discussed the details of the inference algorithm due to space limit. For class inference, we use fixed hyper-parameters without parameter learning since the model is post-processing on the outputs of an existing classifier. As discussed in the probabilistic formulation section, there are only two hyper-parameters in the revised model HMCT-PP: class transitional probability ρ and approximated marginal class probability π. For the hyper-parameter ρ, we can easily set a value that is close to 1 based on the physical constraint. The hyper-parameter π ($P(y_n = 1) = \pi$) is less straightforward. Since the term appears in the denominator of the term $\frac{P(y_n|y_{\mathcal{P}_n})}{P(y_n)}$, a lower value of π enhances the chance of nodes being inferred as class 1 (flood). In practice, the value of π can be selected based on a rough estimation of the flood extent over the area (π should be small if the flood class is largely under-estimated in the original classifier). For example, in the flood mapping application as tested in this paper, the flood class is often under-estimated by many classifiers because pixels of tree canopies overlaying flood water are often mistakenly classified into the dry (non-flood) class. Thus, we should use a generally smaller π (e.g., 0.1). In the case of lacking such prior knowledge, we can choose to set $\pi = 0.5$ (no preference for either class). We test the sensitivity of the class inference algorithm to the two hyper-parameters in the evaluation.

4 Evaluation

In this section, we compared our proposed method with baseline methods in classification performance on two real world datasets. We also evaluated the computational scalability of our method, particularly on the effect of tree node collapsing. Experiments were conducted on Intel(R) Xeon(R) CPU E5-2687w v4 @ 3.00GHz, 64GB main memory, and Windows 10. Unless specified otherwise, we used default parameters in open source tools for baseline methods. Candidate classification methods include:

- **Non-spatial classifiers with raw features**: We tested decision tree (**DT**), random forest (**RF**), maximum likelihood classifier (**MLC**), and gradient boosted tree (**GBM**) in R packages on **raw** features (red, green, blue spectral bands).
- **Non-spatial classifiers with additional potential field feature (elev.)**: We tested **DT**, **RF** and **MLC** here.
- **Non-spatial classifier with post-processing label propagation (LP):** We tested **DT**, **RF** and **MLC** based on label propagation with 4-neighborhood [31].
- **Markov random field (MRF):** We used open source implementation [32] based on the graph cut method [33].
- **Deep learning:** We used U-Net [30] implemented in Python and Tensorflow (source codes [34]). We used a batch size of 16, a dropout rate of 0.2, and a learning rate of 0.0001. We trained the model in 50 epochs. The model uses double convolution layers with batch-normalization. The downsample path contains five pooling layers. We used a loss function based on the dice coefficient and the Adam optimizer.
- **Hidden Markov Tree (HMT)**: We used HMT [22] in C++.
- **Hidden Markov Contour Tree (HMCT):** This is our proposed methods. Both collapsed (**HMCT-C**) and uncollapsed (**HMCT-UC**) versions were implemented in C++.
- **Hidden Markov Contour Tree - Post Processing (HMCT-PP):** This is our TKDE extension on discriminative HMCT model as post-processing implemented in C++.

Dataset description: We used two flood mapping datasets from the cities of Greensville and Grimesland in North Carolina during Hurricane Mathew in 2016. Explanatory features were red, green, blue bands in aerial imagery from NOAA National Geodetic Survey [35]. The potential field was digital elevation map from the University of North Carolina Libraries [36]. All data were resampled into 2 meter by 2 meter resolution. The number of training and test samples were 5000 per class (flood, dry) in both datasets. The test region size was 1856 by 3149 in Greensville and 2757 by 3853 in Grimesland. Training samples in Greensville were drawn beyond but close to the test region. Training samples in Grimesland dataset were drawn far away from the test region to evaluate model generalizability. Note that for the deep learning method U-Net, we had to provide extra training set in form of contiguous segmented flood imagery. We used 240 image blocks (each of a size of 224 by 224) for training, and 60 blocks for validation.

4.1 Classification Performance Comparison

Results on the Greensville dataset were summarized in Table 2. Decision tree, random forest, gradient boosted tree, and maximum likelihood classifier achieved overall F-scores between 0.69 and 0.72 on raw features. Adding the elevation feature improved the overall F-score dramatically. The improvement was due to the relatively lower elevations among *flood* class pixels. However, the recall of the flood class was still mostly below 0.7 even after adding the elevation feature. The reason is that the right elevation threshold in models differ between training samples and test samples. Post-processing based on label propagation (LP) slightly improved performance of non-spatial classifiers (MLC) due to the removal of salt-and-pepper noise errors. Similar results were found in Markov random field. The deep learning method (U-Net) outperformed non-spatial classifiers (with an F-score of 0.84) due to its capability to learn complex textures from aerial images. HMT performed poorly with very low precision for the dry class and low recall for the flood class, meaning that it mis-classified a significant amount of flood class samples into the dry class. Further investigation showed that many large scale obstacles (tree canopies with dry class features) in the flood area biased the learning and inference of HMT model. In contrast, HMCT models performed significantly better. The reason was that HMCT model flow dependency of locations on two sides of the river in separate tree branches (conditionally independent from each other), mitigating the impact of feature obstacles on model learning and inference.

TABLE 2: Classification on Real Dataset in Greenville, NC

Classifiers	Class	Precision	Recall	F	Avg. F
DT+Raw	Dry	0.65	0.87	0.74	0.69
	Flood	0.80	0.53	0.64	
RF+Raw	Dry	0.64	0.97	0.77	0.69
	Flood	0.94	0.45	0.61	
GBM+Raw	Dry	0.72	0.73	0.72	0.72
	Flood	0.73	0.72	0.72	
MLC+Raw	Dry	0.72	0.88	0.79	0.77
	Flood	0.85	0.66	0.74	
DT+elev.	Dry	0.76	1.00	0.87	0.84
	Flood	1.00	0.70	0.82	
RF+elev.	Dry	0.74	1.00	0.85	0.81
	Flood	1.00	0.64	0.78	
MLC+elev.	Dry	0.70	0.98	0.82	0.78
	Flood	0.97	0.59	0.73	
DT+LP	Dry	0.64	0.96	0.77	0.69
	Flood	0.93	0.46	0.61	
RF+LP	Dry	0.64	1.00	0.78	0.69
	Flood	1.00	0.43	0.60	
MLC+LP	Dry	0.72	0.96	0.82	0.79
	Flood	0.94	0.63	0.75	
MRF	Dry	0.71	0.98	0.82	0.78
	Flood	0.96	0.60	0.74	
U-Net	Dry	0.87	0.79	0.83	0.84
	Flood	0.81	0.90	0.85	
HMT	Dry	0.60	0.99	0.75	0.63
	Flood	0.98	0.34	0.50	
HMCT-UC	Dry	0.96	0.95	0.96	0.96
	Flood	0.95	0.96	0.96	
HMCT-C	Dry	0.97	0.96	0.96	0.96
	Flood	0.96	0.97	0.96	

TABLE 3: Classification on Real Data in Grimesland, NC

Classifiers	Class	Precision	Recall	F	Avg. F
DT+Raw	Dry	0.53	0.92	0.67	0.47
	Flood	0.69	0.17	0.27	
RF+Raw	Dry	0.54	0.98	0.70	0.50
	Flood	0.90	0.18	0.30	
GBM+Raw	Dry	0.69	0.84	0.76	0.73
	Flood	0.79	0.62	0.70	
MLC+Raw	Dry	0.70	0.91	0.79	0.76
	Flood	0.87	0.61	0.72	
DT+elev.	Dry	0.79	0.55	0.65	0.70
	Flood	0.66	0.85	0.72	
RF+elev.	Dry	1.00	0.27	0.43	0.58
	Flood	0.58	1.00	0.73	
MLC+elev.	Dry	0.54	0.88	0.67	0.52
	Flood	0.69	0.26	0.38	
DT+LP	Dry	0.52	0.98	0.67	0.43
	Flood	0.83	0.09	0.17	
RF+LP	Dry	0.53	1.00	0.69	0.45
	Flood	0.96	0.11	0.21	
MLC+LP	Dry	0.71	0.96	0.81	0.77
	Flood	0.93	0.60	0.73	
MRF	Dry	0.70	0.97	0.81	0.76
	Flood	0.96	0.57	0.72	
U-Net	Dry	0.81	0.90	0.85	0.84
	Flood	0.89	0.77	0.83	
HMT	Dry	0.57	0.96	0.72	0.56
	Flood	0.89	0.27	0.41	
HMCT-UC	Dry	0.97	0.92	0.95	0.95
	Flood	0.93	0.97	0.95	
HMCT-C	Dry	0.99	0.94	0.96	0.96
	Flood	0.94	0.99	0.97	

Results on the Grimesland dataset were summarized in Table 3. In this dataset, training samples were drawn far away from the test region to evaluate model generalizability. We can see that most non-spatial classifiers performed poorly except for GBM and MLC (overall F-score 0.73 and 0.76). Adding elevation features improved DT and RF but degraded MLC. This is because training area was overall higher than the entire test region. Thus, elevation thresholds learned from training area were not generalizable to test area. Post-processing also had mixed effect on non-spatial classifiers, because it sometimes mistakenly smoothed out corrected classified samples. U-Net outperformed the other traditional non-spatial models due to its capability of learning complex image textures. HMT again performed poorly due to the same reason as in Greensville dataset. HMCT outperformed others with an overall F-score of 0.94. HMCT showed better generalizability because its flow dependency was based on relative elevation values in test region. In HMCT, training samples were only used for providing a reasonably well initial model parameters $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 1, 2\}$.

Sensitivity of HMCT to initial parameters: We conducted sensitivity of our HMCT (with node collapsing) model to different initial parameter values on prior class probability π and class transitional probability ρ (the parameters of $\{\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c | c = 1, 2\}$ were initialized based on maximum likelihood estimation on the training set). Due to space limit, we only showed results on the Greensville dataset. First, we fixed initial $\rho = 0.99$ and varied initial π from 0.1 to 0.9. Results of converged values of ρ together with their corresponding final F-scores were shown in Figure 6(a-b). It can be seen that our HMCT model was quite stable with different initial π values. Similarly, we fixed initial $\pi = 0.5$, and varied initial ρ from 0.2, 0.3, to 0.99. Results in Figure 6(c-d) showed the same trend. In practice, we can select an initial π value around 0.5 and a relatively high initial ρ value such as 0.9 (because *flood* pixels' neighbor is more likely to be *flood* due to spatial autocorrelation).

Parameter iterations and convergence in HMCT: Here we fixed the initial $\pi = 0.5$ and initial $\rho = 0.99$, and measured the parameter iterations and convergence of HMCT (with node collapsing) on the Greensville dataset. Our convergence threshold was set 0.001%. The values of π and ρ at each iteration were summarized in Figure 7 (we omitted $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$ due to the large number of variables. The parameters converged after 14 iterations. The converged value of ρ was close to 1 as expected.

4.2 Computational Performance Comparison

We evaluated the computational performance of the proposed HMCT model. We particularly compared the time costs of HMCT-UC (without node collapsing) and HMCT-C to understand the effect of node collapsing and pre-aggregation. The time cost of HMT should be very close to HMCT-UC since they both have the same number of tree nodes and same set of node operations in learning and inference. Thus, we did not include the cost of HMT to avoid redundancy. For HMCT-C, we further controlled the maximum in-degree (the number of parents) for each tree node. This was done simply by stopping the collapsing operations on a node when its in-degree reached the maximum. We

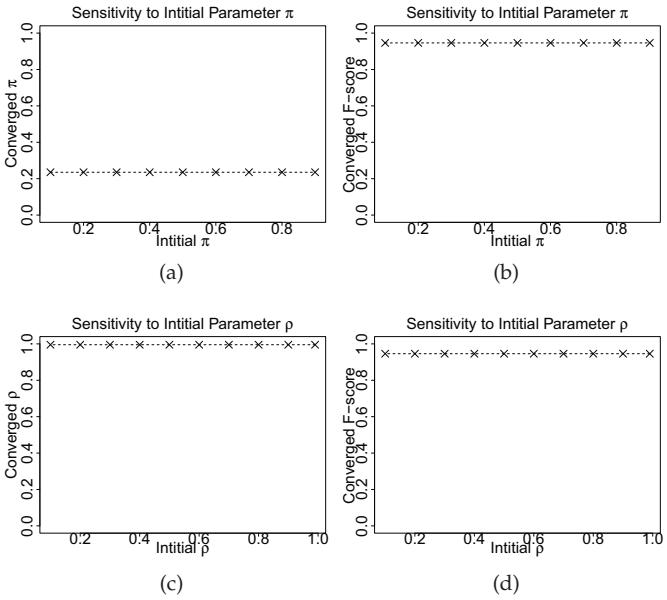


Fig. 6: Sensitivity of HMCT to initial parameters π and ρ

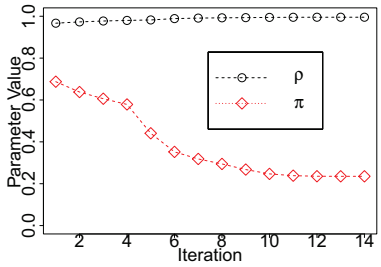


Fig. 7: Parameter iterations and convergence in HMCT

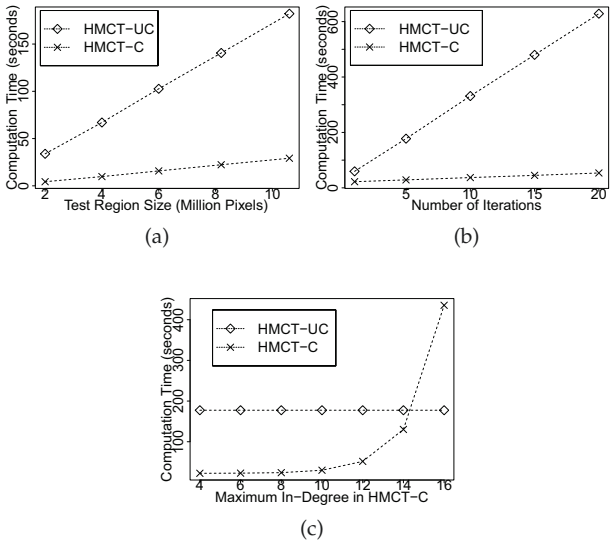


Fig. 8: Computational performance of HMCT-C versus HMCT-UC on varying test region sizes (a), number of iterations (b), and maximum node in-degree (c)

TABLE 4: Time costs of individual components (seconds)

	HMCT-UC	HMCT-C	HMCT-C
Max in-degree	N/A	4	16
Tree construction	13.47	13.47	13.47
Node Collapsing	N/A	6.64	6.64
Pre-aggregation	N/A	0.28	0.28
Parameter learning	152.51	1.80	394.01
Class inference	10.71	0.21	21.09
Total time	176.69	22.41	435.49

compared the two methods on different input data sizes, different number of iterations in parameter learning, and different settings of maximal in-degree per node (in HMCT-C only).

Results were summarized in Figure 8. We fixed input parameters $\rho = 0.99$ and $\pi = 0.3$. First, we chose the number of iterations as 5 and the maximum allowed in-degree in HMCT-C as 5. We varied the size of the test region from around 2 million pixels to around 10.6 million pixels. Results in Figure 8(a) showed that the time costs of both HMCT-UC and HMCT-C increased almost linearly with the test region size. The rate of increase in HMCT-C was smaller. This was consistent with the $O(N_T)$ term in our cost model. Note that N_T is much smaller in HMCT-C due to node collapsing. HMCT-C ran very fast, costing less than 30 seconds on over 10 million pixels. In the second experiment, we chose a test region with around 10.6 million pixels, and set the maximum allowed in-degree in HMCT-C as 10. We varied the number of iterations in parameter learning from 1 to 20. Results in Figure 8(b) showed that the costs of both methods increased almost linearly with the number of iterations. This is consistent with our cost model. In the third experiment, we chose a test region with around 10.6 million pixels, and set the number of iterations as 5. We varied the maximum allowed in-degree in HMCT-C from 4 to 16 (we did not choose values smaller than 4 because the maximum in-degree of uncollapsed tree nodes is already 4 due to the four-neighborhood we used). Results in Figure 8(c) showed that as the maximum allowed in-degree increases, the cost of HMCT-C grew exponentially, exceeding the cost of HMCT-UC when the maximum in-degree was 16. This is consistent with the 2^d term in our cost model.

Table 4 showed detailed time costs for individual components in HMCT algorithms. We chose an input data with a test region of 10.6 million pixels, and the number of iterations as 5, and a max in-degrees of 4 and 16 for HMCT respectively. From the table, we can see that when the max in-degree is 4, learning and inference in HMCT-C were significantly faster than HMCT-UC due to node collapsing. However, when the max in-degree is 16, the costs of HMCT-C was almost twice as that of HMCT-UC, due to the exponential cost in terms of in-degree (2^d). In practice, we can set the maximum allowed in-degree as 4.

We also measured the computational cost of U-Net model on one NVIDIA Tesla P100 GPU. The training set contains 240 image chunks with an equal size of 224 by 224 pixels. The validation set contain 60 image chunks. The total time of model training is 208 seconds for 50 epochs. The time costs of predicting classes on test images are 2.44 seconds for the Grimesland dataset (204 image chunks) and

1.04 seconds for the Greensville dataset (112 image chunks).

4.3 Evaluation of HCMT as Post Processing

Classification performance: We now evaluate the HMCT as a post-processor on top of class probability predictions from five existing methods: decision tree, random forest, maximum likelihood classifier, gradient boosted tree and U-Net. Note that we did not run HMCT-PP on top of HMCT itself because HMCT class inference is based on the maximum likelihood method instead of drawing a threshold (e.g., 0.5) on class probabilities of individual pixels. In other words, the inference algorithm of HMCT only tells which class assignment of all pixels has the highest overall likelihood without providing the class probabilities of individual pixels. Results on the two datasets were summarized in Table 5 and Table 6 respectively. For initial parameters, we set $\rho = 1 - 10^{-9}$ and chose π based on the situation. If the pre-classified flood extent is largely under-estimated, then we set π to a low value (e.g., $\pi = 0.1$ for DT, MLC, GBM) to enlarge the extent. If the pre-classified flood extent is moderate, then we set a moderate value of π (e.g., $\pi = 0.4$ for U-Net). From the two tables, we can find that HMCT-PP significantly enhanced the performance of base classifiers, improving F-cores from below 0.8 to above 0.9. Among all the base methods, HMCT-PP with U-Net achieved the best performance with an F-score of 0.99 and 0.96 in two datasets. The reason for the significant improvement is that HMCT can utilize the high-confident predictions of flood areas from base methods and leverage the class dependency structure from topography constraint to fill the gap of missing flood extent or fine tune the correct flood extent boundary.

One important observation is that HMCT-PP with U-Net outperforms HMCT-C itself. The reason is that HMCT assumes that explanatory feature values of samples in each class follows an i.i.d. Gaussian distribution. This assumption can be over simplistic for our real-world datasets. The main advantage of HMCT is that it captures spatial structural dependency between different class nodes in the contour tree. In other words, HMCT models $P(y_n|y_{P_n}$ in a sophisticated contour tree structure but makes a simplistic assumption on $P(\mathbf{x}_n|y_n)$ (i.e., i.i.d. Gaussian). In contrast, HMCT-PP with U-Net combines the strength of HMCT in modeling $P(y_n|y_{P_n}$ in a contour tree structure and the strength of U-Net in learning $P(y_n|\mathbf{X})$ with spatial contexts of feature X. Thus, we observed that HMCT-PP with U-Net has an average F-score of 0.99, higher than the average F-score of 0.96 in HMCT-C.

TABLE 5: HMCT as post-processing on Greenville dataset

Classifiers	Class	Precision	Recall	F	Avg. F
HMCT-PP+DT	Dry	0.97	0.90	0.94	0.94
	Flood	0.91	0.98	0.94	
HMCT-PP+MLC	Dry	0.95	0.91	0.93	0.93
	Flood	0.92	0.95	0.94	
HMCT-PP+RF	Dry	0.85	0.99	0.91	0.91
	Flood	0.99	0.82	0.90	
HMCT-PP+GBM	Dry	0.90	0.96	0.93	0.93
	Flood	0.96	0.89	0.93	
HMCT-PP+U-Net	Dry	0.98	0.98	0.99	0.99
	Flood	0.99	0.98	0.99	

TABLE 6: HMCT as post-processing on Grimesland dataset

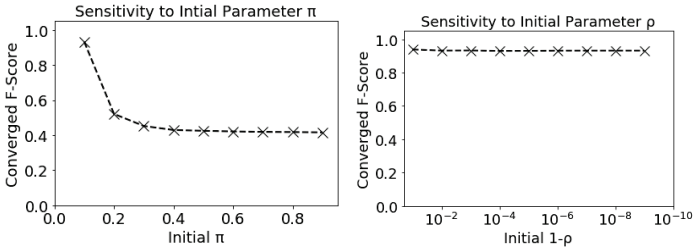
Classifiers	Class	Precision	Recall	F	Avg. F
HMCT-PP+DT	Dry	0.86	0.98	0.92	0.91
	Flood	0.98	0.84	0.90	
HMCT-PP+MLC	Dry	0.88	0.99	0.93	0.93
	Flood	0.97	0.87	0.92	
HMCT-PP+RF	Dry	0.88	0.99	0.93	0.93
	Flood	0.99	0.86	0.92	
HMCT-PP+GBM	Dry	0.95	0.96	0.96	0.96
	Flood	0.95	0.95	0.96	
HMCT-PP+U-Net	Dry	0.95	0.98	0.96	0.96
	Flood	0.98	0.94	0.96	

Sensitivity to initial parameters: We conducted sensitivity analysis of our HMCT-PP on the five base models to different values of input parameters ρ and π. Due to space limit, we only showed results on the Grimesland dataset. When evaluate the sensitivity to π, we fixed the parameter $\rho = 1 - 10^{-10}$, and varied the value of π from 0.1 to 0.9. The results on five base methods are shown in the left sub-figures in Figure 9, Figure 10, Figure 11, Figure 12, and Figure 13 respectively. From the results, we can find that the initial value of π made a significant effect on the final classification performance. For HMCT-PP on DL and RF, only $\pi = 0.1$ is effective. The reason is that the initial classified flood map has very low recall (under-estimation). Thus, we need to set π small enough to compensate for the errors. For HMCT-PP with MLC and GBM, the range of effective π values are wider (0.1 to 0.4 for MLC, and 0.1 to 0.3 for GBM). The reason is likely that the under-estimation in initial flood map is not very significant, and thus we do not necessary set a very small value of π to compensate it. For U-Net model, we can find that the performance is not very sensitivity to π ($\pi = 0.4$ showed the best performance). The reason is that the initial classification in U-Net is mostly confident and correct (F-score is 0.84), and errors are largely on the flood extent boundary. From the analysis, values of π can be chosen based on an initial guess on whether flood is under-estimated or not.

Visualized maps: The above analysis is further confirmed by the visualization of classified flood extent maps from different approaches shown in Figure 14 and Figure 15 respectively. The left column of sub-figures show the flood maps from base classifiers of DT, MLC, RF, GBM, and U-Net respectively, where brown color represents flood and green color represents dry. From the results, we can see that all methods except U-Net shows a significant under-estimation of flood extent due to the tree canopies in the area. This confirms the low recalls in Table 2 and Table 3. The post-processed flood maps by HMCT-PP filled a significant portion of the gap. We can also find that the classified flood maps from U-Net successfully identified the flood areas, but the flood extent boundaries were too coarse. The reason is that U-Net classified the image patch-wise, and very often an entire patch was predicted as flood. The results of HMCT-PP with U-Net showed better flood extent boundaries based on topography.

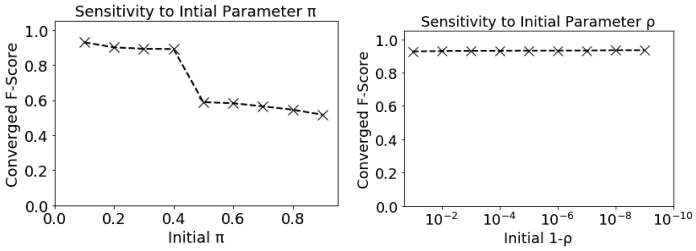
5 Conclusions and Future Works

In this paper, we propose hidden Markov contour tree (HMCT), a spatial structured model that can capture flow



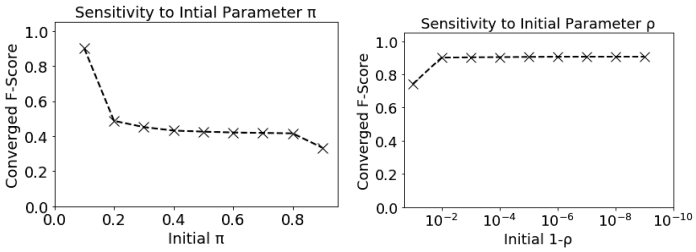
(a) Fix $\rho = 1 - 10^{-9}$ (b) Fix $\pi = 0.1$

Fig. 9: Sensitivity of HMCT-PP+DT to π and ρ



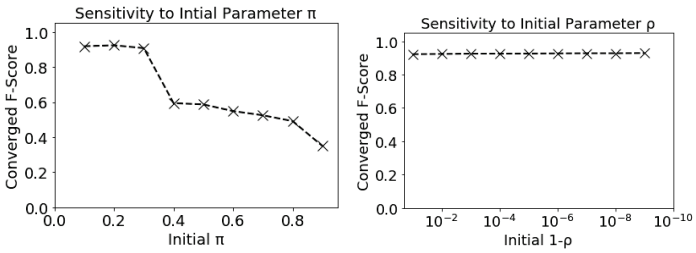
(a) Fix $\rho = 1 - 10^{-9}$ (b) Fix $\pi = 0.1$

Fig. 10: Sensitivity of HMCT-PP+MLC to π and ρ



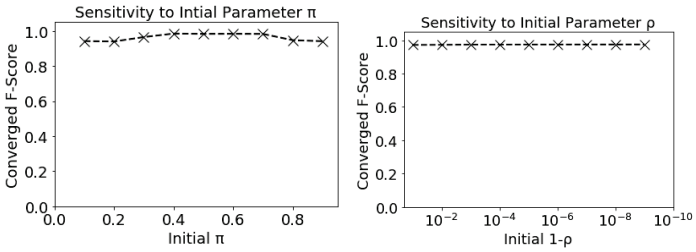
(a) Fix $\rho = 1 - 10^{-9}$ (b) Fix $\pi = 0.1$

Fig. 11: Sensitivity of HMCT-PP+RF to π and ρ



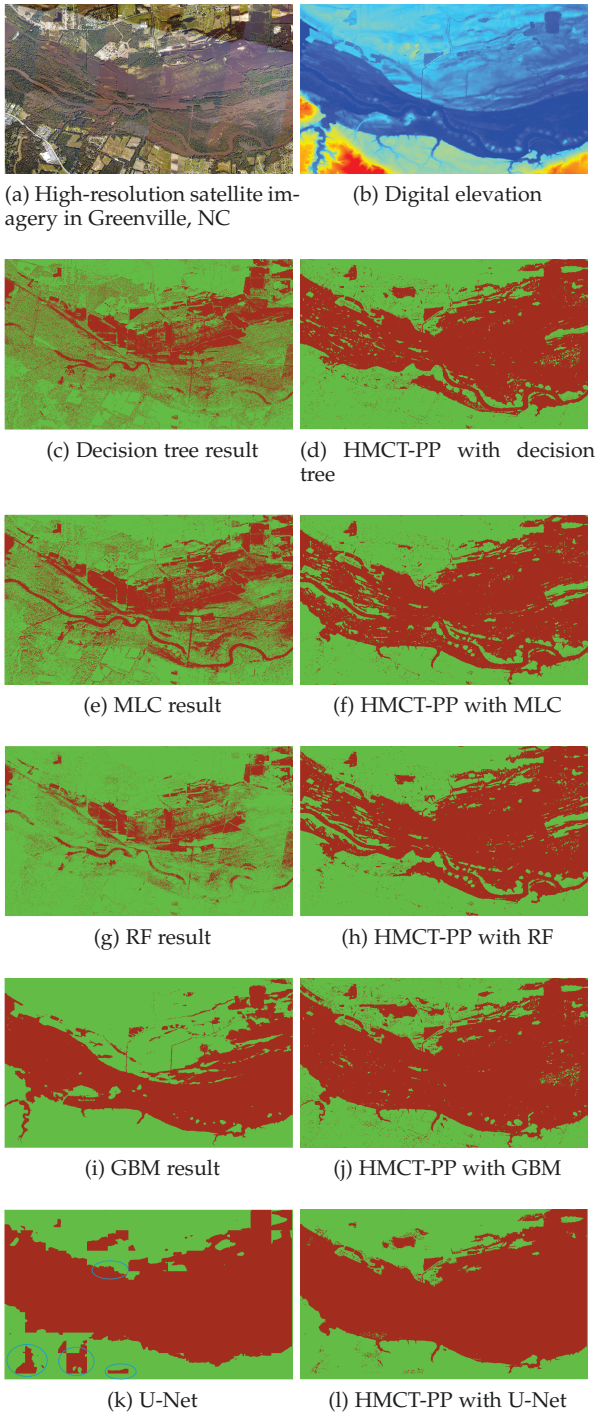
(a) Fix $\rho = 1 - 10^{-9}$ (b) Fix $\pi = 0.1$

Fig. 12: Sensitivity of HMCT-PP+GBM to π and ρ



(a) Fix $\rho = 1 - 10^{-9}$ (b) Fixed π = 0.4

Fig. 13: Sensitivity of HMCT-PP+U-Net to π and ρ



(a) High-resolution satellite imagery in Greenville, NC (b) Digital elevation

(c) Decision tree result (d) HMCT-PP with decision tree

(e) MLC result (f) HMCT-PP with MLC

(g) RF result (h) HMCT-PP with RF

(i) GBM result (j) HMCT-PP with GBM

(k) U-Net (l) HMCT-PP with U-Net

Fig. 14: Results on Greenville flood mapping (flood in brown, dry in green, best viewed in color)

dependency on a potential field surface. We propose efficient algorithms for model parameter learning and class inference. We also extend the HMCT model into a post-processor that can enhance the classified results from other models based on the topography constraints. Evaluations on real world data show that our HMCT algorithms are scalable to large data sizes, and can achieve higher classification performance over existing methods for hydrological applications such as flood mapping.

In future work, we plan to explore integration of deep learning framework with our HMCT. We also plan to ex-

plore parallelization of the proposed HMCT learning and inference algorithms based on OpenMP in multi-core platforms. For example, we can potentially design a thread pool to maintain (independent) tree node operations from different sub-branches in the tree. Another potential future work is to explore generalization of the proposed model formulation to other applications scenarios such as analyzing potential energy landscapes in material science [37], [38] and protein surface topology in biochemistry [6], [39]. Further studies are needed to explore how to incorporate the domain knowledge into the customization of our model structure (e.g., formulation of class transitional probability in a contour tree).

ACKNOWLEDGEMENT

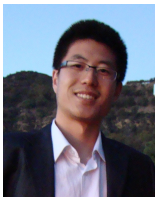
This material is based upon work supported by the National Science Foundation under Grants No. 1850544 and the University Corporation for Atmospheric Research (UCAR) under Contract SUBAWD000837.

REFERENCES

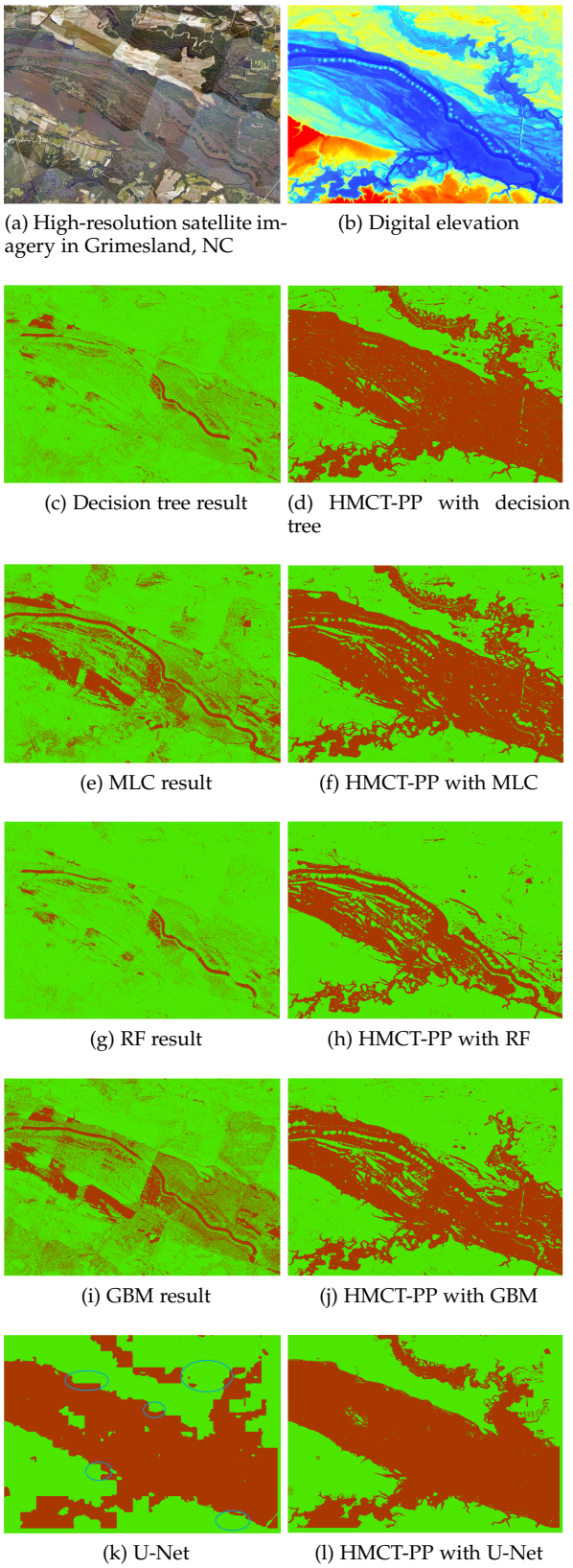
[1] Z. Jiang, "A survey on spatial prediction methods," *IEEE Transactions on Knowledge and Data Engineering*, 2018.
[2] National Oceanic and Atmospheric Administration, "National Water Model: Improving NOAA's Water Prediction Services," http://water.noaa.gov/documents/wrn-national-water-model.pdf, 2018.
[3] D. Cline, "Integrated water resources science and services: an integrated and adaptive roadmap for operational implementation," National Oceanic and Atmospheric Administration, Tech. Rep., 2009.
[4] P. Brivio, R. Colombo, M. Maggi, and R. Tomasoni, "Integration of remote sensing data and gis for accurate mapping of flooded areas," *International Journal of Remote Sensing*, vol. 23, no. 3, pp. 429–441, 2002.
[5] A. Kuhn, W. Engelke, M. Flatken, H.-C. Hege, and I. Hotz, "Topology-based analysis for multimodal atmospheric data of volcano eruptions," in *Topological Methods in Data Analysis and Visualization*. Springer, 2015, pp. 35–50.
[6] D. Günther, R. A. Boto, J. Contreras-Garcia, J.-P. Piquemal, and J. Tierny, "Characterizing molecular interactions in chemical systems," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2476–2485, 2014.
[7] V. Pascucci, X. Tricoche, H. Hagen, and J. Tierny, *Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications*. Springer Science & Business Media, 2010.
[8] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.
[9] L. Anselin, *Spatial econometrics: methods and models*. Springer Science & Business Media, 2013, vol. 4.
[10] J. Lafferty, A. McCallum, F. Pereira *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning (ICML), IEEE*, vol. 1, 2001, pp. 282–289.
[11] H. Wang and Z. Li, "Region representation learning via mobility flow," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2017, pp. 237–246.
[12] Z. Yao, Y. Fu, B. Liu, W. Hu, and H. Xiong, "Representing urban functions through zone embedding with human mobility patterns." in *IJCAI*, 2018, pp. 3919–3925.
[13] Y. Zhang, Y. Fu, P. Wang, X. Li, and Y. Zheng, "Unifying inter-region autocorrelation and intra-region structures for spatial embedding via collective adversarial learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1700–1708.
[14] P. Wang, Y. Fu, H. Xiong, and X. Li, "Adversarial substructured representation learning for mobile user profiling," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 130–138.
[15] P. Wang, Y. Fu, J. Zhang, X. Li, and D. Lin, "Learning urban community structures: A collective embedding perspective with periodic spatial-temporal mobility graphs," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 6, pp. 1–28, 2018.
[16] H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in *2018 AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.
[17] K. Subbian and A. Banerjee, "Climate multi-model regression using spatial smoothing," in *SIAM International Conference on Data Mining (SDM)*, 2013, pp. 324–332.
[18] A. Karpatne, A. Khandelwal, S. Boriah, and V. Kumar, "Predictive learning in the presence of heterogeneity and limited training data," in *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, 2014, pp. 253–261.
[19] J. Stoeckel and G. Fung, "SVM feature selection for classification of SPECT images of alzheimer's disease using spatial information," in *ICDM*, 2005, pp. 410–417.
[20] S. Banerjee, B. P. Carlin, and A. E. Gelfand, *Hierarchical modeling and analysis for spatial data*. CRC Press, 2014.
[21] P. A. Viton, "Notes on spatial econometric models," *City and regional planning*, vol. 870, no. 03, pp. 9–10, 2010.
[22] M. Xie, Z. Jiang, and A. M. Sainju, "Geographical hidden markov tree for flood extent mapping," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. ACM, 2018, pp. 2545–2554.
[23] Z. Jiang and A. M. Sainju, "Hidden markov contour tree: A spatial structured model for hydrological applications," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 804–813.
[24] J. R. Munkres, *Topology*. Pearson, January 2000.
[25] H. Edelsbrunner and J. Harer, *Computational topology: an introduction*. American Mathematical Soc., 2010.
[26] H. Carr, J. Snoeyink, and U. Axen, "Computing contour trees in all dimensions," *Computational Geometry*, vol. 24, no. 2, pp. 75–94, 2003.
[27] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, 2001.
[28] O. Ronen, J. Rohlicek, and M. Ostendorf, "Parameter estimation of dependence tree models using the em algorithm," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 157–159, 1995.
[29] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[31] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," 2002.
[32] The Middlebury Computer Vision Pages, "C++ Source Code of MRF," http://vision.middlebury.edu/MRF/code/, 2018.
[33] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields," in *European conference on computer vision*. Springer, 2006, pp. 16–29.
[34] ZFTurbo, "ZF_UNET_224 Pretrained Model," https://github.com/ZFTurbo/ZF_UNET_224_Pretrained_Model, 2018.
[35] National Oceanic and Atmospheric Administration, "Data and imagery from noaa's national geodetic survey," https://www.ngs.noaa.gov.
[36] NCSU Libraries, "LIDAR Based Elevation Data for North Carolina," https://www.lib.ncsu.edu/gis/elevation, 2018.
[37] D. J. Wales, M. A. Miller, and T. R. Walsh, "Archetypal energy landscapes," *Nature*, vol. 394, no. 6695, pp. 758–760, 1998.
[38] D. Wales *et al.*, *Energy landscapes: Applications to clusters, biomolecules and glasses*. Cambridge University Press, 2003.
[39] H. Edelsbrunner and P. Koehl, "The geometry of biomolecular solvation," *Combinatorial and computational geometry*, vol. 52, pp. 243–275, 2005.

Arpan Man Sainju is a Ph.D. student in the department of Computer Science at the University of Alabama, Tuscaloosa. He received his B.E. degree in computer engineering from Tribhuvan University, Nepal. He has worked on spatial big data algorithms on GPU. His research interests include data mining, spatial database, big data analytics, parallel computing and deep learning.

Wenchong He is a Ph.D. student in the department of Computer Science at the University of Alabama. He received his B.S. degree from University of Science and Technology of China (USTC) in 2017, and Master degree from the College of William Mary in 2019. His research interests include machine learning, data mining and deep learning.



Zhe Jiang is an assistant professor in the department of Computer Science at the University of Alabama, Tuscaloosa. He received his Ph.D. in computer science from the University of Minnesota, Twin Cities in 2016, and B.E. in electrical engineering from the the University of Science and Technology of China, Hefei, China, in 2010. His research interests include spatial big data analytics, spatial and spatio-temporal data mining, spatial database, geographic information system, and their interdisciplinary applications in earth science, transportation, public safety, etc. He is a member of IEEE.



(a) High-resolution satellite imagery in Grimesland, NC (b) Digital elevation (c) Decision tree result (d) HMCT-PP with decision tree (e) MLC result (f) HMCT-PP with MLC (g) RF result (h) HMCT-PP with RF (i) GBM result (j) HMCT-PP with GBM (k) U-Net (l) HMCT-PP with U-Net

Fig. 15: Results on Grimesland flood mapping (flood in brown, dry in green, best viewed in color)