Hydrologic Predictions using Probabilistic Relational Models

Max Metzger
Charles River Analytics.
625 Mt. Auburn St.
Cambridge, MA 02138

Alison O'Connor
Charles River Analytics
625 Mt. Auburn St.
Cambridge, MA 02138

David F. Boutt
University of Massachusetts
611 North Pleasant Street
Amherst, MA 01003

Joe Gorman
Charles River Analytics.
625 Mt. Auburn St.
Cambridge, MA 02138

Abstract

The US Army faces a significant burden in planning sustainment operations. Currently, logistics planners must manually evaluate potential emplacement sites to determine their terrain suitability. Sites subject to rainfall-runoff responses such as flooding are ill-suited for emplacements, but evaluating the likelihood of such responses requires significant time and expertise. To reduce the time and to ease the difficulty of logistics site selection we demonstrated a series of Terrain Impact Decision Extensions (TIDE) for use in logistics planning tools and processes. TIDE performs data-fusion over a variety of terrain and weather data sets using probabilistic relational models (PRMS), providing a high-performance alternative to physics-based hydrologic models.

1. INTRODUCTION

Maintaining a constant supply of water and fuel is critical to sustaining the US Army's forces in the field. (Department of the Army, 2008). To provide access to these critical resources, logistics planners must deliver those resources with minimal failure to establish and maintain emplacements (e.g., tanks, fuel lines) capable of storing these crucial commodities. Water and fuel supplies must not be susceptible to disruption, damage, and contamination by water due to rainfall-runoff responses such as flooding, overland flow, and ponding (i.e., the temporary accumulation of surface water).

Currently, the risks posed by rainfall-runoff responses to potential emplacement sites are manually evaluated, and require considerable expertise and time. Site evaluation is further complicated for areas lacking detailed data that describe terrain, soil properties, and subsurface conditions (e.g., the presence of aquifers). This occurs due to the significant role played by terrain, soil, and subsurface factors in the effects of rainfall run-off on terrain. A decision aid capable of automatically evaluating the suitability of emplacement sites would reduce the time needed for evaluation by logistics planners and improve the quality of sites selected.

To reduce the time and the difficulty of logistics site selection we designed and demonstrated a series of Terrain Impact Decision Extensions (TIDE) for logistics planning tools and processes. TIDE performs data-fusion over a variety of terrain and weather data sets, and uses probabilistic relational models (PRMs) to reason with uncertainty to evaluate the suitability of potential logistics sites against a series of expert rules for a variety of emplacement systems. By using PRMs to rank the severity of potential rainfall-runoff responses, TIDE was able to site determine suitability much faster than by rigorous physical simulation. Additionally, PRMs can reason with incomplete data (e.g., a lack of detailed soil information), making them useful even when evaluating data-poor regions.

1.1 PROBLEM DESCRIPTION

The rainfall-runoff response of landscapes is a fundamental problem in the field of hydrology (Singh, 1988). The accumulation of water at a particular time-space location on the Earth's surface (i.e., terrain ponding) is the result of the confluence of many climatologic, hydrologic, and physical factors and parameters. During a liquid precipitation event (e.g., rain), water is transported in three main ways: water can run-off/on in the form of overland flow; infiltrate into the soil and become ground water; or be transferred back into the atmosphere via evapotranspiration. Overland flow can in turn lead to the accumulation of surface water (e.g., flooding), which poses a risk to US Army emplacements.

The terrain assessment model must account for multiple aspects of the area of interest. First, overall climatic conditions (i.e., arid, semi-arid, humid) have an important influence over the relative distribution of water in the three pathways. The model must account for uncertainty in weather predictions and climatologic predictions. Second, the model must account for local factors within the area of interest that influence the rainfall-runoff response. There are many such factors, including rainfall intensity and duration, slope of the land, land use and land cover characteristics (i.e., vegetation and impervious surfaces), soil and air temperature, soil hydraulic properties, and soil moisture conditions. The data sources for these factors may be incomplete or inaccurate, introducing additional uncertainty.

The prediction of where, when, and how long water will accumulate on the land surface is reliant on constraining parameters that describe the above processes and conditions. Fortunately, hydrologists have been developing tools to both quantify these factors and develop quantitative models for predicting rainfall-runoff response to precipitation events.

These models are often based on solving complex equations that govern the physics of surface and subsurface water (Abbott, Bathurst, Cunge et al., 1986; Panday & Huyakorn, 2004) or assign statistical values to terrain based on observation (Yoram, 2003). These models are not practical for US Army planning because they require complete data sets, are extremely time-consuming to compute, and do not scale to the levels of detail and scope required by US Army logistics planners.

2. APPROACH

Given the potential incompleteness of input parameters (including terrain, soil and subsurface data), our approach uses a probability-based method to track the inferences made about data through the model. For TIDE to be useful, the system must infer terrain characteristics, soil properties, and subsurface conditions from limited data. While terrain elevation data is available for most of the world at varying levels of detail, soil data is less prevalent. Land use, land cover (e.g., vegetation), as well as the soil's hydrologic properties and moisture conditions are all factors in predicting rainfall-runoff response. When this information is not directly available, it needs to be estimated or inferred. For example, soil properties for a given region within the United States may be well-known and stored in a Geographic Information System (GIS) database, but this data may be unknown for many rural regions around the world. An exhaustive geological survey of potential sites within that region is not possible given time and personnel constraints. Even when terrain, soil, and subsurface data are present, it may not be at resolutions high enough to be relevant to the emplacements (e.g., a map with soil data at a resolution of 500m is of limited use when selecting a site for a fuel line less than a meter across). In cases where data describing terrain characteristics, soil properties, and subsurface conditions are absent, purely rule-based approaches are insufficient, as rules alone are poorly-suited to handling incomplete data. The system must be capable of reasoning with limited or incomplete data before executing any impact assessment rules. The PRM model developed under the TIDE effort is capable of reasoning with incomplete data and inferring data that may be absent. Additionally, while our initial model is very simple, further work may expand the model to be very complex. The object-oriented PRM approach is well-suited for such complexity.

The PRM output is used to generate maps showing the likelihood for flow accumulation at a given location for a certain amount of time. We based our models on the Hortonian Infiltration and Runoff/On (HIRO2) model, which was originally developed for the USDA (Meng, Green, Salas et al., 2008). This model predicts rainfall-runoff responses, including runoff channels (in which surface water flows) and the time until ponding occurs. The HIRO2 model performs well, but operates at larger scales than are useful to emplacement selection, generally being most accurate at scales of hundreds of meters. The HIRO2 model served as the basis for our model, but was modified to operate at higher levels of fidelity without significantly compromising performance.

Bayesian modeling techniques have been used in the field of hydrology for decades (Vicens, Rodriguez-Iturbe, Schaake et al., 1975), but the majority of this work has different goals than TIDE. Bayesian modelling approaches generally take existing models that use direct measurements as inputs (e.g., rainfall) and predict specific hydrologic response values (e.g., runoff rate, groundwater level). Bayesian techniques are then used to calibrate the models parameters to improve their accuracy (Beven & Binley, 1992; Thiemann, Trosset, Gupta et al., 2001; Vrugt, Ter Braak, Clark et al., 2008).

TIDE differs from past Bayesian hydrologic models in two fundamental ways. First, our model attempts to predict the impact of rainfall-runoff responses, not their precise values. Generally speaking, US Army logistics planners are not concerned about predicting the exact amount of surface water that may accumulate, but are instead primarily concerned about the impact on the mission. For example, the difference between 1.2 meters of standing water or 2.4 meters is irrelevant if either makes the mission impossible to complete.

Second, the TIDE model must perform with reasonable accuracy in regions of the world that have little, if any, hydrologic data observations (e.g., hourly flow rates for a stream) that can be used to train or calibrate a model. Instrumenting and measuring rainfall-runoff responses in these areas may be too costly, logistically infeasible, or dangerous. As a consequence the TIDE model must rely on generally available data (e.g., elevation, land cover, weather).

2.1 PROBABALISTIC RELATIONAL MODELS

To represent our terrain and hydrologic models in a probabilistic form that allows us to determine the suitability of an area of interest, we designed a probabilistic relational model (PRM) (Koller & Pfeffer, 1998; Pfeffer, Koller, Milch et al., 1999; Friedman, Getoor, Koller et al., 1999). PRMs describe the world in terms of classes of objects, instances of those classes and relationships between them. Serving as a powerful extension of Bayesian Networks (BNs), PRMs use object-oriented semantics that capture attribute, structural, and class uncertainty to overcome computational and storage complexity challenges faced by BNs.

The design of PRMs has proven to be useful in representing a wide range of complex domains that involve uncertainty and require flexibility and reusability. In regard to complexity, PRMs capture the logical and relational structure of a domain. For example, PRMs specify how one attribute influences the value of another attribute. In our PRM, the value of the attribute, RankDrainageCapacity, is dependent on the values of attributes, LandCoverType and SoilType, from two other classes. Therefore, the model uses the values of RankDrainageCapacity's dependent attributes, LandCoverType and SoilType, to infer the value of RankDrainageCapacity.

To handle uncertainty, PRMs use probability distributions encoded in the model to determine values of unknown variables. The value of LandCoverType and SoilType for a location are retrieved from data sources outside the model and then posted to the model. Therefore, there is little uncertainty in regard to these two attributes. Conversely, the value of RankDrainageCapacity is inferred inside the model using probability distributions. To overcome the uncertainty involved with this attribute, encoded in the model is a map of possible combinations of land cover and soil types to appropriate probability distributions. Relying on the team's hydrologic expertise, we created initial distributions for each possible pair of land cover and soil types, as well as each land cover and soil type provided the other attribute was unknown. Similarly, using domain knowledge, we supplied distributions for each individual slope ranking, flow ranking, and drainage ranking assuming that the other two attributes were unknown. Given the increased number of combinations for RankRunoffPotential, to obtain distributions in the case that two or three attributes were known, we multiplied the probabilities of the known attributes for each of the five possible RankRunoffPotential values. The distributions for Suitability were much simpler to encode, as only five probability distributions that required no further calculations were necessary. While these initial distributions pass face validation, future work is needed to adjust the distributions to meet higher accuracy needs.

To support flexibility and reusability, PRMs allow the reuse of the same class probability models for all instances of a class. New probabilistic models do not have to be constructed for each new situation. Instances of classes can be configured in any way desired for a given situation. The relationships that hold between these instances are captured by the PRM. For example, our PRM contains a class SiteModel that has one attribute, Suitability. The value of Suitability depends on the instance of the Runoff class' attribute, RankRunoffPotential. To reason over this model, one must create instances of both the SiteModel and Runoff classes.

The flexibility and reusability of PRMs grant us the ability to reason over millions of locations. For each location, the relevant set of known facts about specific attributes – the land cover type, soil type, rank of slope and rank of flow –must be provided to the instances of classes. As we transition to discuss our PRM in greater detail, it will become more evident that these four key features of PRMs – complexity, uncertainty, flexibility, and reusability – are crucial to obtaining successful results. Bayesian Networks could also apply to this problem, as the relational structure is fixed for every instance. Nevertheless, the object-oriented representation of PRMs were quite helpful in designing the model.

2.2 PRM EDITOR

We developed a PRM Editor that provides an intuitive graphical user interface (GUI) that allows users to create complex PRMs by defining classes of objects, adding attributes to those class definitions, creating instances of the classes, and specifying the relationships between them.

Upon launching the PRM Editor, the user can navigate between three views: the global view, class view, and instance view. While these three views are initially blank, the panels become populated with information and graphical representations of the model. The global view allows a user to view the PRM as a whole in a folder format. Its top-level folder, named after the PRM, can be expanded to display three other folders, enums, classes, and instances. The enums and instances folders can further be expanded to show all enumerations and instances of classes in the model. Within the classes folder are additional folders for each class that can be expanded to view the attributes in that class.

Unlike the global view, the class view displays a graphical representation of the PRM. Each class is represented by a box labeled with the class name. If applicable, arrows are automatically drawn between boxes indicating super and subclasses (parent-child relationships). The instance view also displays boxes that, rather than represent classes, represent the instances of classes in the model.

To begin utilizing these three views, the user has the option to either load an existing model into the GUI or create a new PRM. After loading or initializing the model, the user can begin building the model by adding classes. When creating a class, the user must specify the name and parent class of the new class. In the case of our model, we created six classes, none of which had parents, so this field remained blank.

Adding an attribute requires more detail than adding a class. A user must specify the attribute name, type, and resolution. The user has the option to assign single or multi as the attribute's type, as well as choose from a list of possible types. Possible types contained in this list include integer, real number, Boolean, type, nothing, as well as all of the classes and enumerations created by the user. If the attribute is of type enumeration, the user must have previously defined the appropriate enumeration. For example, in our model, the SiteModel class' attribute, Suitability, is of type enumeration. The possible values for Suitability are VeryPoor, Poor, Medium, Well and VeryWell. Therefore, we created an enumeration called RankPoor, to represent an attribute with these five possible values. Before defining an attribute's resolution, the user must create instances of other classes. By creating instances of classes, attributes in other classes can depend on the attributes of these instances. Figure 1 shows the global and class relationship views after the six classes and their six respective instances have been created.

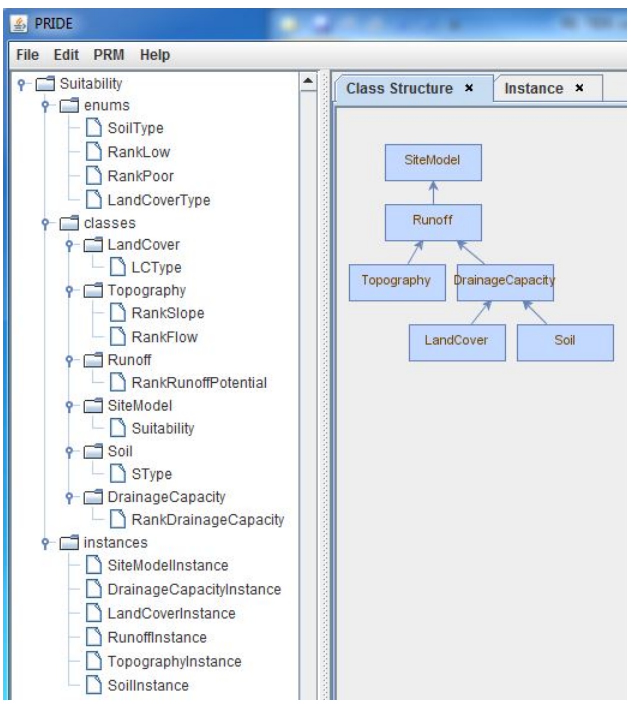

Figure 1: Global View (left), Class Relationship View (right)

To complete the implementation of the previously discussed attribute, Suitability, its resolution must be defined. The resolution of an attribute can be assigned as nothing, assignment, or dependency. Upon creating the attribute, the default choice is nothing. If the user updates the resolution to assignment, the user must enter the exact value of the attribute or its reference. For example, if the attribute were an integer, the user could indicate that value was 10. Alternatively, the reference could be set to an attribute of another class that was also an integer. The appropriate resolution for Suitability is dependency. Therefore, the user must specify the influencer, the attribute that Suitability depends on, as well as the conditions and their respective distributions. The conditions are the possible values of the influencer. Each possible value of the influencer is paired with a CPD indicating the likelihood of each possible value of the attribute. Suitability depends on the instance of the Runoff class', RunoffInstance, attribute RankRunoffPotential. RankRunoffPotential has five possible values – VeryLow, Low, Medium, High, and VeryHigh. Therefore, Suitability will have five conditions and five distributions that indicate the probability of each of Suitability's five possible values occurring given the value of RankRunoffPotential.

Having defined four enumerations, six classes, seven attributes, and six instances in our model using the PRM Editor, the model was saved to as a .prm file that could be used by the TIDE system.

2.3 HYRDOLOGIC MODEL

A PRM consists of a set of class probability models. The final version of our PRM (Figure 2) contains six classes – SiteModel, Runoff, Topography, DrainageCapacity, LandCover, and Soil. Each class has a set of attributes. Attributes are either simple or complex. Simple attributes are random variables that represent direct properties of an object, such as the type of land cover or type of soil, whereas complex attributes represent relationships to other objects. The attributes in our model are all simple.

Logical relationships can be described between classes. The lines in Figure 2 represent these relationships. Assuming we have an instance of every class, an instance of LandCover is related to an instance of the DrainageCapacity class by the LandCoverType attribute.

Each simple attribute is associated with a set of parents and a CPD. The parents are determined by the attributes that the attribute depends on. Attributes can depend on either other simple attributes of the same object or of related objects. An example of an attribute of an object depending on an attribute of a related object is the dependence of the RankDrainageCapacity on LandCoverType.

Attributes of related objects are specified via attribute slot chains, such as the slot chain LandCoverInstance LandCoverType. This slot chain begins with the object representing the land cover of a location, and accesses the simple attribute indicating the type of land cover at this location. The model specifies that the RankDrainageCapacity attribute of the DrainageCapacity class has this slot chain as a parent. To reiterate, this indicates that the RankDrainageCapacity depends probabilistically on the LandCoverType. The other slot chain parent of RankDrainageCapacity is SoilInstance.SoilType. It is important to emphasize that these parent relationships are general, meaning that the land cover or soil type may vary from scenario to scenario, but the probabilistic relationships hold for all scenarios (e.g., when a new area is investigated).

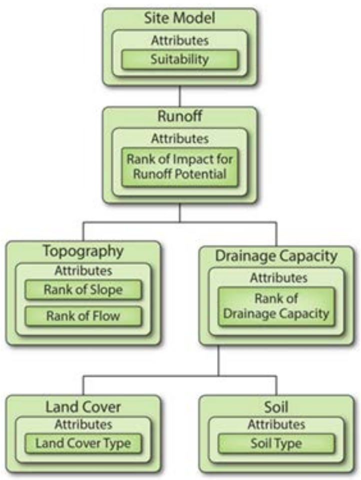

Figure 2: Hydrologic PRM

In our tree-structured PRM, the attributes in the classes directly below another class are parents to the attributes in the class above them. Therefore, the attributes in the three leaf classes, LandCover, Soil and Topography, do not have parents. The values of these attributes are derived outside the model and posted as evidence to the model. Conversely, the values of the attributes in the remaining classes, SiteModel, Runoff, and DrainageCapacity, are inferred from the data available within the model.

Recall that the other information associated with a simple attribute is a CPD that specifies a distribution over values of an attribute given the values of its parents. In the case of Suitability, its parent is Runoff.RankRunoffPotential. Table 1 shows the code for the implementation of the SiteModel class, complete with its attribute, Suitability, specification of its parent, RankRunoffPotential, and CPD for every possible value of RankRunoffPotential. The bolded line specifies, in plain terms, that if the RankRunoffPotential is VeryLow then there is 10% likelihood Suitability is VeryPoor, 15% likelihood Suitability is Poor, 50% likelihood Suitability is Medium, 15% likelihood Suitability is Well and 10% likelihood Suitability is VeryWell.

Similar to how parent relationships are defined, the assigned CPD is general; it holds no matter what the specific related objects are.

Table 1: SiteModel Class Implementation

```
class SiteModel = {
     Suitability:       single       RankPoor       depends       on
[RunoffInstance.RankRunoffPotential]
     case [VeryLow] =>
           (0.1 -> VeryPoor, 0.15 -> Poor, 0.5 -> Med, 0.15 -> Well,
0.1 -> VeryWell)
     case [Low] =>
           (0.7 -> VeryPoor, 0.1 -> Poor, 0.1 -> Med, 0.05 -> Well,
0.05 -> VeryWell)
     case [Med] =>
           (0.05 -> VeryPoor, 0.1 -> Poor, 0.7 -> Med, 0.1 -> Well,
0.05 -> VeryWell)
     case [High] =>
           (0.05 -> VeryPoor, 0.05 -> Poor, 0.1 -> Med, 0.7 -> Well,
0.1 -> VeryWell)
     case [VeryHigh] =>
           (0.05 -> VeryPoor, 0.05 -> Poor, 0.1 -> Med, 0.1 -> Well,
0.7 -> VeryWell)
     case [_] =>
           (0.2 -> VeryPoor, 0.2 -> Poor, 0.2 -> Med, 0.2 -> Well, 0.2
-> VeryWell)
             }
```

With a clear understanding of how relationships and CPDs are specified in the model, we can discuss how inference ultimately determines if a location is suitable. The basic order of how the model performs inference is: once the values of an attribute's parents are known, the value of that attribute can be inferred. Therefore, the process begins by posting evidence to the leaf classes. First, the LandCoverType, SoilType, RankSlope, and RankFlow evidence is posted to the model. Next, the model can infer the value of RankDrainageCapacity from the land cover and soil data. For example, if the LandCoverType is Shrub and the SoilType is Vertisols, the probability distribution encoded in the model for RankDrainageCapacity given this evidence is: case [Shrub,Vertisols] => (0.7 -> VeryPoor, 0.3 -> Poor, 0.0 -> Med, 0.0 -> Well, 0.0 -> VeryWell). Again, this distribution can be interpreted as: If LandCoverType is Shrub and SoilType is Vertisols, there is 70% likelihood the RankDrainageCapacity is VeryPoor and 30% likelihood the rank of drainage capacity is Poor. Once the CPD of RankDrainageCapacity is determined, the distribution can be used in conjunction with the Topography evidence to infer the value of the RankRunoffPotential. This process propagates up the model, as RankRunoffPotential influences the value of Suitability.

To determine a site's suitability, the model uses all available data. While accuracy increases with amount of available data, our model is capable of reasoning with incomplete or no data. In the case that data is unavailable or unknown for the four inputs – LandCoverType, SoilType, RankSlope, and RankFlow – the probability is evenly distributed over all possible values.

Our PRM Editor utilizes the open source Figaro probabilistic programming language (PPL) (www.cra.com/figaro) to perform inference. PPLs provide a powerful and flexible way to represent probabilistic models using the power of programming languages. In addition, PPLs offer general-purpose reasoning algorithms for inference and machine learning. Our implementation utilizes the Metropolis-Hastings reasoning algorithm, capped with a runtime of 5,000 milliseconds per inference.

2.4 INTEGRATION

Our PRM used data from the following data sources.

2.4.1 SRTMVF2

The Shuttle Radar Topography Mission (SRTM) was a joint project between NASA and the National Geospatial-Intelligence Agency (NGA) to create high-resolution land surface data for much of the world (roughly 80% of the Earth's land surface is covered). The SRTM Void-Filled 2 (SRTMVF2) data set is at 1-arc-second (approximately 30-meter) resolution data, with many gaps in data void-filled using interpolation techniques (Dowding, Kuuskivi, Li et al., 2004). The SRTMVF2 dataset serves as our primary elevation data source, as our hydrologic model is heavily dependent on accurate, high-resolution elevation data. However, we have identified that there are gaps within the SRTMVF2 elevation data. In areas where no SRTMVF2 data can be found, we can fall back to lower resolution DTED data, including the SRTMVF1 and SRTMVF0 data sets. Elevation is used to determine inputs to our model, slope and water flow. Slope and flow implicitly capture the spatial relationships of each DTED point with its neighbors, allowing the PRM to reason about each point's data independently.

Slope

Slope is determined using the elevation dataset. For the initial effort, we used a simple algorithm that iterates across each elevation point. For each point, the relative change, dE, in elevation is calculated for each adjacent point (excluding diagonally adjacent points.) The dE value with the greatest magnitude is selected, and the distance between points (1 arc-second in the case of the SRTMVF2 dataset) is used to calculate the angle of the terrain's surface, Θ. This value can range from 0 degrees (i.e., perfectly flat) to 90 degrees (which would be a perfectly vertical surface.) While there are more elaborate methods for determining slope that provide more accurate results, this technique can process millions of points in a matter of minutes, and yields sufficient accuracy for the needs of the terrain assessment model.

Once the slope angles have been calculated using the algorithm described above, they are translated from a continuum of [0, 90) to five discrete values, which are used as inputs for the terrain model. Table 2 shows how angle ranges are mapped to model inputs.

Table 2: Mapping terrain slope angles to model inputs

Angle Range	Rank of Slope
$0 \le \Theta \le 10$	Very Low
$10 < \Theta \le 20$	Low
$20 < \Theta \le 30$	Medium
$30 < \Theta \le 60$	High
$\Theta > 60$	Very High

Flow

The elevation dataset is used to predict flow channels – that is, paths that surface water is likely to take in the event of rainfall. A greater amount of flow indicates a risk of surface water accumulation. To calculate flow, we relied on the TopoToolbox (Schwanghart & Kuhn, 2010). The toolbox includes techniques for predicting flow estimation. The flow values predicted can vary wildly. In the case of our AOI, estimated flow varied between 0 and over 3,300. To normalize the dataset, we first transformed the flows to a logarithmic scale (changing the range from 1 to ~9.7) and then normalized the results to [0, 1].

Table 3: Mapping Flow to Model Inputs

Flow Range	Rank of Flow
flow is exactly 0	Very Low
$0 < flow \le 0.1$	Low
$0.1 < flow \le 0.2$	Medium
$0.2 < flow \le 0.5$	High
$flow > 0.5$	Very High

As shown in Table 3, these values are then translated into five discrete inputs for the terrain assessment model (same as the slope). As with the slope values, the process of mapping flows to discrete ranking values is independent from the flow calculations. This means that calculating the flows (a process that took roughly two hours for the Demonstration Scenario's AOI) need only be run once per AOI, even if we adjust model values or how flow values are mapped to model inputs.

2.4.2 GeoCover

Earth Satellite Corporation (EarthSat) developed the GeoGover data set, a global landcover database. The GeoCover dataset consists of 13 land cover classes and is available for much of the world (Cunningham, Melican, Wemmelmann et al., 2002). Classes of land cover include grasslands, agriculture areas (i.e., farmland), wetlands, and water/ice. This data will serve as additional inputs to our terrain models so we can more accurately assess rainfall-runoff response. The GeoCover dataset will also enable TIDE to identify bodies of water.

2.4.3 Harmonized World Soil Database

The Harmonized World Soil Database (HWSD) was produced by the European Union's European Commission Joint Research Centre (more specifically, the Land Management Unit of the Institute for Environment and Sustainability.) The HWSD is a 30 arc-second (approximately 90-meter) resolution that contains detailed information about the top soil and subsoil properties. It was created by merging data from four different soil databases (Nachtergaele, Van Velthuizen, Verelst et al., 2008).

This data allows the model to more accurately predict how terrain will respond to surface water (for example, how quickly water will be absorbed into the soil.) This dataset's low resolution means that some terrain boundaries (such as coasts) and geographical features (such as bodies of water) are of low accuracy compared to the other data sets.

2.5 MISSION DECISION RULES

The system must provide a set of logistic system-specific terrain assessment rules for a variety of systems and purposes (e.g., Tactical Water Distribution System, Assault Hose Line System). Terrain suitability may vary from system to system—for example, a suspended hose may be unaffected by some types of standing water while a ground-level hose could be at risk for contamination. Rule sets for individual systems will need to account for these differences, allowing planners to choose the appropriate system given the characteristics of a prospective emplacement site. Additionally, logistics planners must be able to easily modify and expand these rules as new systems are introduced, and as mission requirements change. (For example, different rules would be used for route planning than well placement.)

In our initial effort, we have implemented some basic rules that filter terrain suitability for a hypothetical fuel line. The fuel line has two requirements: (1) it must be installed on flat land (so the pumps can function properly); and (2) the fuel lines cannot be placed in standing water (to prevent contamination), which includes bodies of water (such as lakes) and areas that are prone to flooding.

To determine suitability for the fuel lines, we take slope, land coverage, and hydrologic suitability as inputs. We then apply a set of rules as described in Table 4. The rules transform the hydrologic suitability into mission suitability. These rules favor flat land over sloped land.

Table 4: Mission Rules

Condition	**Effect**
If the point is a body of water	Mission suitability is Very Poor
If slope is ranked as "Low" or "Very Low" and hydrologic suitability is "Medium"	Mission suitability is High
If slope is ranked as "Low" or "Very Low" and hydrologic suitability is *not* "Medium"	Mission suitability is equal to hydrologic suitability
If slope is ranked as "Medium" and hydrologic suitability is "High" or "Very High"	Mission suitability is Medium
If slope is ranked as "Medium" and hydrologic suitability is *not* "High" or "Very High"	Mission suitability is equal to hydrologic suitability
If slope is "High" or "Very High"	Mission suitability is Very Poor

Currently, rules are distinct from the PRM model, so that custom rules can be written for different operational needs while using the same PRM model. For example, while the PRM output is constant, the mission requirements for a long fuel pipeline may be very different than the mission requirements for a convoy. The fuel line would have a very low tolerance for changes in elevation (as the pumps cannot handle the increased workload) and would be susceptible to contamination from standing water. The convoy, while still limited by severe terrain or flooding, would be much more resilient to water and slopes.

A standard rules engine and associated rules language, such as that provided by JBoss, would allow mission experts to author rules for the TIDE system without requiring them to understand the PRM or hydrology.

2.6 VISUALIZATION

Outputs of the PRM and the rules engine, as well as the data sources themselves, were rendered within NASA WorldWind. WorldWind can accept a variety of GIS data formats and is easily customizable. Using the open-source the Geospatial Data Abstraction Library (GDAL), we wrote custom modules to render the HWSD and GeoCover data sets, while the SRTMVF2 data was loaded in using built-in WorldWind methods. Model output and rules output were rendered as textures which were then projected onto the WorldWind globe at the appropriate coordinates, but the data could easily be written to a variety of GIS formats.

The hydrologic suitability is presented as belief values in five categories: {Very Poor, Poor, Medium, High, Very High}. Related military impact assessment (e.g., weather impact assessment) is done at three intervals (e.g., low risk, medium risk, and high risk.) We expanded our model to use five intervals instead of three to present additional granularity in the model's output. Further work is needed to determine the best number of intervals and their thresholds.

For the initial effort, the category with the highest belief value is selected as the 'correct' suitability value. These categories are then color-mapped for visualization: {Red, Orange, Yellow, Green, Blue}. The same categories and colors are used for the rules output.

3. DISCUSSION

For our area of interest and the 1 arc-second SRTMFV2 set, there are 25,934,402 points to process. Executing the entire PRM for each point would be unnecessarily complex – instead, we store each unique combination of {soil type, land cover, rank slope, rank flow} and store the associated beliefs. This means we can simply look up the correct PRM output for each unique combination of inputs, which need only be run through the PRM once. As a result, we are able to process all 25.9 million points in only two hours. (Further updates to the Figaro library should increase runtime performance as well.) In a full-scale TIDE system, the PRM values for all combinations could be calculated once and only once, and then stored in a database for quick reference. This database would only need to be updated when the PRM is updated.

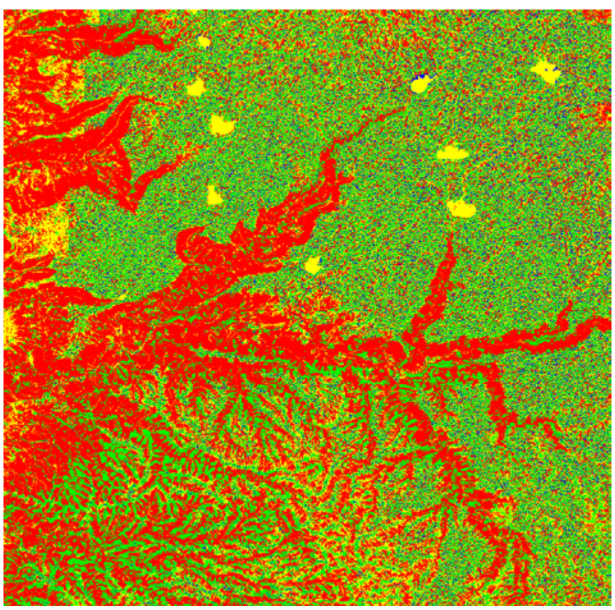

Figure 3: PRM Model Output

Figure 3 shows the output of the model. (Figures 3 and 4 are best viewed in color.) The output of the model is very grainy as each point in the elevation set can have a distinct rank. Of note are the red regions running across the central region of the image – these are riverbeds and their surrounding valleys, which were detected despite those bodies of water not being explicitly present within our GeoCover or HWSD data sets.

Figure 4 shows the output of our rules engine (and a region slightly larger than the figure above). While these rules are very simple, they demonstrate how rules can transform the high-density output of the models (Figure 3, above). The model output scores each point in the elevation grid (approximately a 30 by 30 meter square when using the SRTMVF2 data set), producing a very dense output. Rules can be used to simplify the models' output into easier-to-interpret regions. With these simple rules, we were able to execute rules across the entire region in five minutes. Figure 4 shows the same riverbeds as in Figure 3, but the view is expanded to show a large lake to the west, which has been appropriately flagged as having very poor mission suitability. Unlike the riverbeds (which were predicted by the PRM), this body of water is present within both the GeoCover and HWSD data sets (at differing levels of precision).

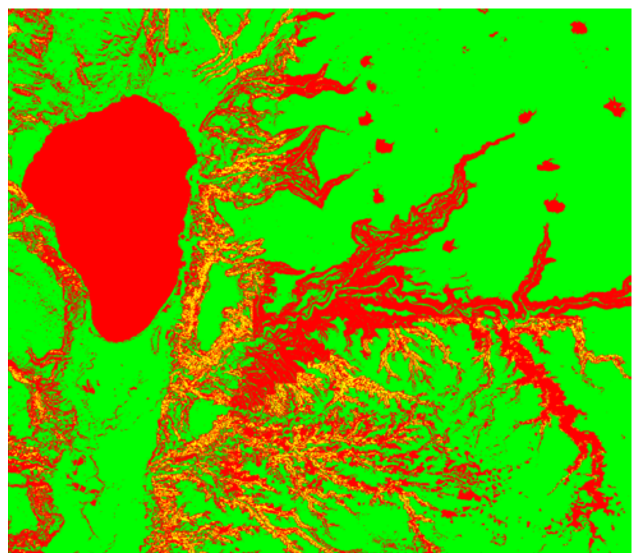

Figure 4: Rules Output

4. FUTURE WORK

4.1 TERRAIN AND HYDROLOGIC MODELS

Future improvements to the model begin by incorporating more data. The more information captured by the model, the more accurate the inferences will be. The next data source to integrate is precipitation data. Depending on the duration of the mission, weather or climate data would be used. For example, missions spanning from zero to four months would heavily rely on weather information, missions spanning from four to eight months would integrate both weather and climate data, and missions lasting longer than eight months would incorporate climate data. We will also work to quantitatively evaluate the performance and applicability of our models.

Our approach to testing and verifying the accuracy of our models is two-fold. First, we will compare the outputs of our models to those of existing, alternative hydrologic models. These models are often based on solving complex equations that govern the physics of surface and subsurface water (Abbott et al., 1986; Panday & Huyakorn, 2004) or assign statistical values to terrain based on observation (Yoram, 2003). These models are not practical for US Army planning because they require complete data sets, are extremely time-consuming to compute, and do not scale to the levels of detail and scope required by US Army logistics planners. However, their outputs have been validated when tested on carefully monitored and measured regions of terrain, typically within the US. By running the TIDE models on the same regions and comparing its output to that of the established models, we can confirm that the TIDE models are functioning correctly.

Second, we will gather existing data sources of rainfall-runoff responses. Several regions within the United States have had their rainfall-runoff responses measured at various degrees of fidelity. For example, the Leaf River basin in Mississippi has over forty years of time series data that includes precipitation and runoff (Yapo, Gupta, Sorooshian et al., 1996). Additional data sources could be built from flood records and high water level records. These data sets will serve to validate the PRM models used by TIDE. They may also serve as training data to calibrate the model to more accurately predict the severity of rainfall run-off responses (e.g., flooding).

4.2 DATA FUSION MODEL

Our basic solution for handling cases of limited or missing data assumes that each value is equally likely if no evidence is posted to the model. Under this assumption, the accuracy of our inferences declines with limited or no data. The inferences are only as strong as the data known and evidence provided.

Future improvements for how to reason with incomplete or no data involve adjusting the prior distributions. Although the prior distributions in our current model assume that all values of an attribute are equally likely if no data is available, one would argue this is not representative of the real world. We plan to explore the possibilities of more representative prior distributions. For example, the prior distribution for land cover type could reflect that fact that over 70% of the earth's surface is covered in water, making it the most likely of the seven values.

This being said, the most dramatic mitigation of consequences due to incomplete data or unknown values will result from future improvements to the model itself rather than the dependencies. As we integrate more data sources into the model, the number of attributes and dependencies will increase, resulting in more accurate inferences. Existing data can also be used to infer missing data. For example, using higher-resolution data (such as elevation data or land cover data) we can easily determine that the HSWD fails to cover the coastlines. We can then predict the missing values using spatial relationships. Ambiguous areas could be assigned multiple values with different confidence values. Figure 5 shows how the two HSWD regions could be used to infer the values for the missing regions.

Point A, to the north, would be assigned a high probability of having luvisols as the dominate soil type. Point B would be assigned near equal probabilities of being either luvisols or vertisols. Point C, to the south, would be assigned a high probability of vertisols as the dominate soil type. The inference used for point B could be assigned to any region near the boundaries of low-resolution data sets – for example, point D could also be assigned a probability of being either vertisols or luvisols; even though the data set classifies it as vertisols, the resolution is low enough that the point could be a misclassification. The assigned probabilities, along with the soil types themselves, would serve as inputs to the PRM models.. For example, the soil type input to our PRMs for Point D could be "{Vertisols-50%, Luvisols-50%} instead of simply {Vertisols}.

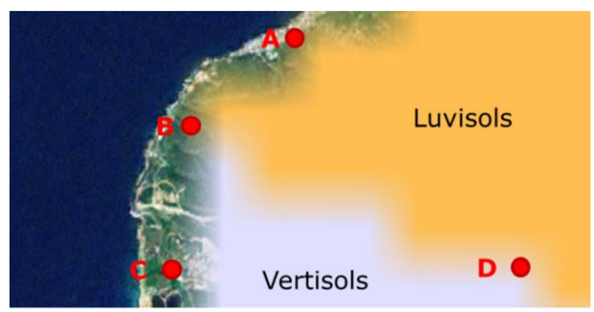

Figure 5: Reasoning about incomplete data

5. CONCLUSIONS

Flooding, and other terrain rainfall-runoff responses, pose significant risk and cost to US Army operations. Assessing the magnitude of flood risk and the impact it will have on a mission requires both time and expertise that may not always be available. An automated system for predicting the likelihood and impact of flooding and surface water accumulation would be of great benefit to logistics planners and the US Army at large.

During our initial effort, we demonstrated the feasibility of Terrain Impact Decision Extensions to predict rainfall-runoff response. We have identified key data sources required for predicting flooding and have developed an initial set of models that are capable of identifying regions that are at high risk of flooding. These models are capable of processing millions of data points per hour, allowing them to process thousands of square kilometers. We feel these models and their performance indicate our approach is sound, and future work will refine and validate the models' performance.

Acknowledgements

This work was performed under US Army Research Lab contract number W911QX-13-C-0111. The authors would like to thank Mr. Peter Grazaitis for his significant technical support and eager engagement on this project. This work was funded in its entirety by ARL. We would also like to thank Ms. Yvonne Fuller and Ms. Jill Oliver for their assistance in preparing this paper.

References:

Abbott, M. B., Bathurst, J. C., Cunge, J. A., O'Connell, P. E., and Rasmussen, J. (1986). An introduction to the European Hydrological System—Systeme Hydrologique Europeen,"SHE", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of hydrology*, 87, 45-59.

Beven, K.& Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, 6, 279-298.

Cunningham, D., Melican, J., Wemmelmann, E., and Jones, T. (2002). GeoCover LC-A moderate resolution global land cover database. In *ESRI International User Conference*.

Dowding, S., Kuuskivi, T., and Li, X. (2004). Void fill of SRTM elevation data–principles, processes and performance. In *Images to Decisions: Remote Sensing Foundations for GIS Applications, ASPRS, Fall Conf., Sep*, 12-16.

Friedman, N., Getoor, L., Koller, D., and Pfeffer, A. (1999). Learning probabilistic relational models. In *Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)*.

Koller, D.& Pfeffer, A. (1998). Probabilistic frame-based systems. In *Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 580-587.

Meng, H., Green, T. R., Salas, J. D., and Ahuja, L. R. (2008). Development and testing of a terrain-based hydrologic model for spatial Hortonian Infiltration and Runoff/On. *Environmental Modelling & Software*, 23, 794-812.

Nachtergaele, F., Van Velthuizen, H., Verelst, L., Batjes, N., Dijkshoorn, K., Van Engelen, V., Fischer, G., Jones, A., Montanarella, L., and Petri, M. (2008). Harmonized world soil database. *Food and Agriculture Organization of the United Nations*.

Panday, S.& Huyakorn, P. S. (2004). A fully coupled physically-based spatially-distributed model for evaluating surface/subsurface flow. *Advances in water Resources*, 27, 361-382.

Pfeffer, A., Koller, D., Milch, B., and Takusagawa, K. T. (1999). SPOOK: A system for probabilistic object-oriented knowledge expression. In *14th Annual Conference on Uncertainty in AI (UAI)*.

Schwanghart, W.& Kuhn, N. J. (2010). TopoToolbox: A set of Matlab functions for topographic analysis. *Environmental Modelling & Software*, 25, 770-781.

Singh, V. P. (1988). Hydrologic systems. Volume I: Rainfall-runoff modeling. *Prentice Hall, Englewood Cliffs New Jersey. 1988. 480*.

Thiemann, M., Trosset, M., Gupta, H., and Sorooshian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research*, 37, 2521-2535.

Vicens, G. J., Rodriguez-Iturbe, I., and Schaake, J. C. (1975). A Bayesian framework for the use of regional information in hydrology. *Water Resources Research*, 11, 405-414.

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., and Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research*, 44.

Yapo, P. O., Gupta, H. V., and Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. *Journal of Hydrology*, 181, 23-48.

Yoram, R. (2003). *Applied stochastic hydrogeology*.: Oxford University Press.