A Machine-Learning Based Connectivity Model for Complex Terrain Large-Scale Low-Power Wireless Deployments

Carlos A. Oroza, Ziran Zhang, Thomas Watteyne, Steven D. Glaser

***Abstract*—We evaluate the accuracy of a machine-learning-based path loss model trained on 42,157,324 RSSI samples collected over one year from an environmental wireless sensor network using 2.4 GHz radios. The 2218 links in the network span a 2000 km^2 basin and are deployed in a complex environment, with large variations of terrain attributes and vegetation coverage. Four candidate machine-learning algorithms were evaluated in order to find the one with lowest error: Random Forest, Adaboost, Neural Networks, and K-Neareast-Neighbors. Of the candidate models, Random Forest showed the lowest error. The independent variables used in the model include path distance, canopy coverage, terrain variability, and path angle. We compare the accuracy of this model to several well-known canonical (Free Space, plane earth) and empirical propagation models (Weissberger, ITU-R, COST235). Unlike canonical models, machine-learning algorithms are not problem-specific: they rely on an extensive dataset and a flexible model architecture to make predictions. We show how this model achieves a 37% reduction in the average prediction error compared to the canonical/empirical model with the best performance. The article presents a in-depth discussion on the strengths and limitations of the proposed approach as well as opportunities for further research.**

I. Introduction

Low-power wireless technology is increasingly being used by the scientific community for remote environmental sensing. One community at the forefront is the hydrology community. Bogena *et al.* highlight the potential of low-power wireless for measuring soil water content variability [1], Pohl *et al.* do a similar analysis for understanding the snow cover [2]. Rice and Bales show how embedded sensors can be used to evaluate the water content of snow [3]. Simoni *et al.* use wireless sensor networks to model the hydrologic response of an alpine watershed [4]. Li *et al.* summarize lessons learned from deploying a wireless sensor network for soil monitoring [5]. Gutierrez *et al.* use low-power wireless to monitor water and automate irrigation [6].

Since 2013, we have been deploying and operating the American River Hydrologic Observatory (ARHO) to monitor the snowpack in the California Sierra Nevada. This observatory consists of sensors measuring snow depth, air temperature, air relative humidity, soil temperature, soil moisture, solar radiation, interconnected by 14 independent low-power wireless mesh networks, deployed in the American River basin, a 2000 km^2 area [7]. Fig. 1 shows the location of the 14 networks. These networks are deployed in locations strategically chosen to get representative spatial estimates of snow cover, soil moisture and other water-balance components. Data from these sensors are integrated with forecasting models and decision-support tools: ARHO is the core element of a new water-resource information system, and a platform for improving Hydro-Electric generation operations using real-time data. To the best of our knowledge, ARHO is the largest environmental low-power wireless sensor network in operation today.

In environmental monitoring applications, sensors are often deployed in remote regions. In the ARHO deployment, it takes hours of driving and hiking to get to a deployment site, and installing a single node takes about about an hour. When the snow season starts, the deployment sites are usually inaccessible for 3-4 months. A trial-and-error deployment, in which nodes are added/moved over the course of a couple of days to obtain the right connectivity, is hence not an option. When we install a new node, we need to be able to make an informed guess at how well that node will connect to already

C. Oroza, Z. Zhang, and S. D. Glaser are with UC Berkeley, CA, USA. (e-mail: {coroza,zhangziran,glaser}@berkeley.edu).

T. Watteyne is with Inria-Paris, EVA team, France. (e-mail: thomas.watteyne@inria.fr).

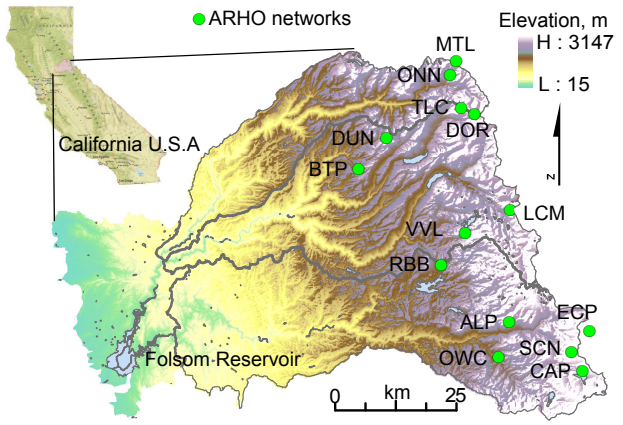

Fig. 1. The 14 low-power wireless networks deployed in the American River Hydrologic Observatory, near Sacramento, CA, USA. Each deployment site is identified by a 3-letter codename.

installed nodes. A key tool for making that guess is a connectivity model. Given different features of the deployment (the distance between the nodes, the amount of vegetation, etc), that model must produce the best possible prediction of the Receive Signal Strength (RSSI) between that node and different other nodes in the network (we use the term Receive Signal Strength (RSS) and Receive Signal Strength indicator (RSSI) interchangeably in this article).

Such a connectivity model is a basic building block for planning the physical connectivity of a deployment, regardless of the type of networking technology being used (star topology, multi-hop redundant mesh network). Even though, in the ARHO networks, we are using a particular type of low-power mesh network (see Section II), the methodology developed in this paper is not tied to that networking technology, and applies equally well to star networks. Similarly, even through we focus on low-power wireless networks deployed in forested mountainous areas, the methodology can apply to any deployment area, as long as the data used to train the model resembles the data to predict (see an in-depth discussion in Section VI-A).

Propagation models (equations) are a natural choice for predicting connectivity. Yet, canonical path-loss models are ill-suited in complex terrain due their simplifying assumptions of plane earth or free space environments. In order to address these limitations, empirical path-loss models that parametrize the effects of vegetation on path-loss are used in a variety of settings. Although these models attempt to capture the excess power loss as a function of frequency and foliage depth, they are otherwise univariate (i.e. they are only a function of distance).

The ARHO networks produce on average 10 times more network statistics than sensor measurements. Part of these statistics are RSSI measurements: each node reports the RSSI of the link to each of its neighbors, every 15 min. This dataset and gives us a unique opportunity to quantify the performance of the propagation models.

Perhaps more importantly, having this dataset allows to think about connectivity models in a radical new way. In traditional propagation models, one creates an equation which approximates the observed data, and fine-tunes the parameters in a univariate equation (multipliers, exponents, etc.) so it matches the data points best. We proposed to use a completely agnostic "big data" approach by associating to each of the RSSI measurements a set of features (distance, vegetation, terrain, etc. between the communicating nodes) and train a multivariate non-parametric model so it learns which features are most important in predicting the RSSI. The result of this machine-learning approach is a predictor: given a new set of features (e.g. a new node is added), is it able to predict the RSSI over that link. The connectivity dataset allows us to quantify the performance of this approach through cross-validation. This allows us to evaluate the accuracy of the machine learning method used, which is based on an ensemble of regression trees (Random Forest). This is a multivariate, non-parametric method in which an ensemble of decision trees are trained on existing data.

The contributions of this article are threefold:

- We present a connectivity dataset consisting of 42,157,324 RSSI measurements gathered on 2218 wireless links in the ARHO networks.
- We develop a machine-learning approach to predict link quality by training on this dataset.
- We evaluate *(i)* the suitability of traditional and this machine-learning-based model for predicting the RSSI in complex environments, *(ii)* whether features other than distance play a role in the prediction of RSSI in complex terrain, and *(iii)* the strengths and limitations of our machine learning-based methodology.

The remainder of this article is organized as follows. Section II describes the large-scale connectivity dataset. Section III discusses related connec-

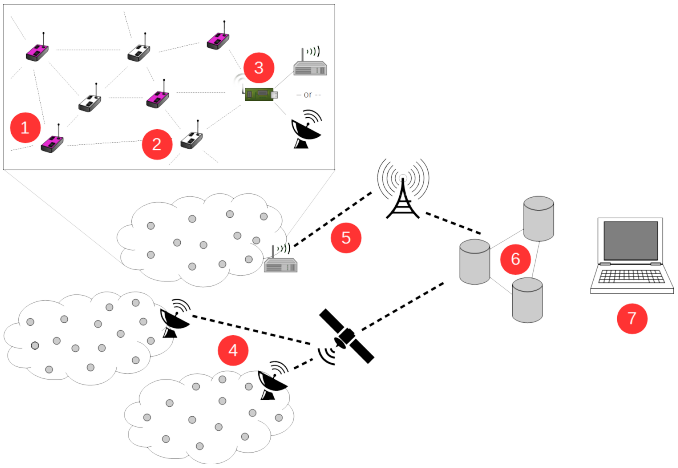

Fig. 2. Network architecture of the American River Hydrologic Observatory.

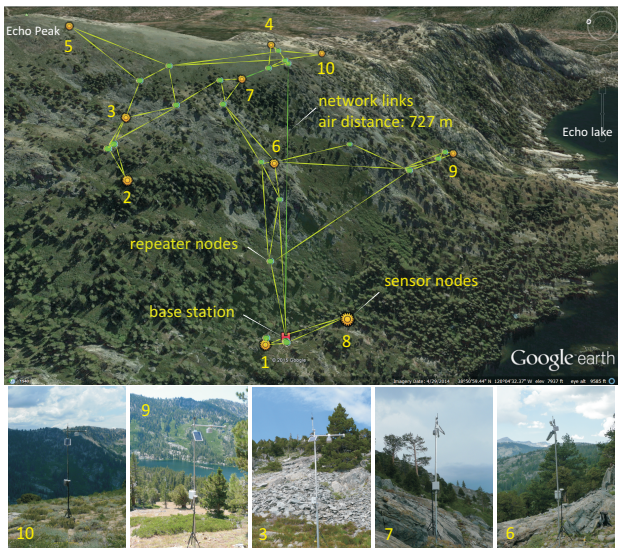

Fig. 3. The Echo Peak deployment (marked "ECP" in Fig. 1).

tivity models for forested environments. Section IV presents the machine-learning model we propose. Section V evaluates the performance of the machine learning and canonical models. Section VI summarizes the key findings, discusses the strengths and limitations of the proposed approach and presents the opportunities for further research.

II. A Large-Scale Real-World Connectivity Dataset

The connectivity data we use in this article is gathered from nodes in the ARHO. Fig. 2 shows the network architecture of the AHRO observatory. Circles numbers in the paragraph (e.g. ①) refer to the annotations in Fig. 2. Sensor stations ① are placed a hydrologically significant locations. Relay nodes ② are added to ensure redundant connectivity. The sensor data is relayed to the manager ③ node, which is connected to a Linux computer. This computer connects to the Internet through a satellite ④ or a cellular link ⑤. Seconds after the generated data is produced in the deployment site, it appears in the database ⑥ and can be visualized online ⑦.

Sensor stations are deployed in locations with diverse physiographic features of elevation, slope, aspect, and canopy coverage, across the basin. A typical network consists of 45 nodes deployed over a 1-2 km^2 area. Terrain attributes in each catchment are heterogeneous: low-elevation networks are characterized by a flat, densely-forested terrain, whereas high-elevation sites are in an Alpine environment with large variations in topography and canopy coverage. Fig. 3 shows the Echo Peak network (marked "ECP" in Fig. 1), which is representative of the other 13 networks.

At the heart of each sensor station is a NeoMote, a low-power wireless platform commercialized by Metronome Systems, a UC Berkeley spin-off company (www.metronomesystems.com). The NeoMote is a generic sensor platform, which features a Cypress PSoC micro-controller and a SmartMesh IP low-power wireless mote, in a hardened weather-proof design. We chose to use two different types of omni-antennas with different gain level to adapt with the terrain challenges. Pairs of 6-dB gain antennas were used on flat areas with less than 10-degree slopes. When the slope between to radios are greater than 10-degree, we switch to pairs of 4-dB antennas. The lower-gain antennas has a wider beam width in the vertical direction, hence the performance is improved on severe slopes. The pins on the micro-controller are programmable, and allow a user to connect any sensor and actuator to the board (www.cypress.com). The SmartMesh IP mote is the element that communicates wirelessly, and is commercialized by the Dust Networks product group at Linear Technology (www.linear.com/dust). The low-power wireless mesh network created between the devices offers >99.999% end-to-end reliability and a over decade of battery lifetime [8].

Besides generating sensor measurements, the network continuously produces network statistics for an operator to be able to assess its "health". The network generates approximately 10 times more

network statistics than sensor measurements. Each node in the network produce a "Health Report" every 15 min which contains the list of neighbors it is communicating with, and – among other things – the average RSSI of the packets it received from that neighbor over the past 15 min. Over the course of 1 year, we have collected 42,157,324 such RSSI measurements from the 2218 wireless links that make up the ARHO deployments.

Every entry in the dataset of RSSI values is annotated with a set of "features" to characterize the topographic and canopy structures between the two nodes which have exchanged the data used to make that RSSI measurements. The locations of the deployed nodes were logged the Magellan Explorist 710 handheld GPS unit. Features at the associated locations are extracted from two digital raster maps. We use a 30 m resolution digital elevation model (DEM) from the National Elevation Dataset (www.nationalmap.gov/elevation.html). The DEM stores elevations above sea level in meters for the latitude and longitude coordinates of each pixel. Each pixel is 30 m wide in the North-South direction. The spacing varies in the east-west direction, depending on latitude. The clusters are separated with distances from 10 to 50 km. Features associated to canopy density are extracted from a percent-tree canopy cover raster developed by the National Land Cover Database (www.mrlc.gov/nlcd2011.php) (NLCD) with 30 m resolution. The NLCD map gives a relative canopy density value ranging from 0 to 100, representing the percentage of tree canopy cover for each pixels [9]. It should be noted that 30 m is a large area compared to the vegetation size. Given the large variability of tree height, species, etc. the NCLD values may not precisely correspond to the vegetation density between nodes. It should also be noted that for 3.4% of the data, the distance is less than 30m. In this situation, the density is marked as the density of the grid cell that contains both nodes.

The features we annotate each RSSI measurement with are:

1) **Path ground distance** (a number in meters): the distance between the two radios communicating for this RSSI measurement, calculated from their GPS locations and elevations.
2) **Mean percent tree canopy cover** (a number between 0% and 100%): the average pixel value from the NLCD vegetation map along the line-of-sight path between the two communicating nodes.
3) **Terrain complexity**: the standard deviation of the raster values from the DEM along the line-of-sight path between the two communicating nodes.
4) **Vegetation variability**: the standard deviation of the raster values of the NLCD vegetation map along the line-of-sight path between the two communicating nodes.
5) **Path angle**: the angle between the line-of-sight path between the two communicating nodes and horizontal.
6) **Source canopy coverage**: the bi-linear interpolated values of the NLCD vegetation map pixels at the source and receiver locations, a number between 0% and 100%.
7) **Receiver canopy coverage**: same calculation as the "Source canopy coverage" feature, but at the receiver node.

III. Related Connectivity Models for Forested Environments

The aim of a propagation model is to predict the expected connectivity. Models based on the physics of the diffusion of electromagnetic waves in an ideal medium (Friis propagation) provide a first-order approximation of the expected connectivity. More complex models, such as "plane earth" account for constructive/destructive interference based on the height of the transmitter and receiver nodes, and the assumption of an ideal flat and empty environment. These propagation models are described in Section III-A.

In forested environments, signal strength is attenuated by vegetation. Related work has focused on empirically modeling the excess signal strength loss due to canopy, based on field measurements. These models are described in Section III-B.

A. Canonical Propagation Models

The simplest path-loss model is "free space" propagation. It assumes unobstructed, line-of-sight decay of an electromagnetic wave, based on the Friis transmission equation, see (1) (and equation (1) in [10] "free space").

$$\frac{P_r}{P_t} = \frac{G_t G_r \lambda^2}{(4\pi)^2 d^2 L} \tag{1}$$

In (1), P_t is power transmitted by the transmitter, P_r is power received by the receiver, G_t and G_r are the gains of the antennas at the transmitter and receiver, respectively, λ is the signal wavelength (m), d is the distance between the transmitter and receiver (m), L is the system loss factor (equal to 1 for free space, but modified in the empirical models detailed Section III-B).

"Free space" propagation does not capture the effect of ground reflection. The "plane earth" is a second canonical model which takes into account the effect of ground reflection, under the assumption of an infinite ground plane, see (2) (or (2a) in [10]).

$$\frac{P_r}{P_t} = 2\frac{G_t G_r}{L}(\frac{\lambda}{4\pi d})^2[1 - cos(k_\omega \frac{2h_t h_r}{d})] \quad (2)$$

In (2), h_r and h_t are the heights of the transmitter and receiver (m), respectively, k_w is the wave number (m^{-1}).

B. Empirical Propagation Models in Forested Environments

While canonical plane earth and free space path loss models provide a first-order approximation of path loss, their simplifying assumptions make them unrealistic in the general case. We are looking for a model representative of our deployments in forested environments. Models specific to this environment model the excess loss induced by the foliage characteristics between each link. They modify the loss factor (L) from the canonical models. Well-known propagation models for forested environments include Weissberger's modified exponential decay model [11], the ITU Recommendation (ITU-R) model [12] and the COST235 model [13].

The Weissberger [11] model assumes that propagation occurs through a dense body of dry trees (see (4) in [10]). It assumes that propagation only occurs through the trees, and is not diffracted over the top of the trees. The loss factor is given by (3).

$$L_{Weiss} = \begin{cases} 1.33 f^{0.284} d^{0.558} & 14 \text{ m} < \text{d} < 400 \text{ m} \\ 0.45 f^{0.284} d & 0 \text{ m} < \text{d} < 14 \text{ m} \end{cases} \quad (3)$$

In (3), L_{Weiss} is the loss due to foliage, f is the transmission frequency (GHz), d is the distance between transmitter and receiver (m).

Another common propagation model is ITU-R [12]. Like Weissberger, it assumes that the majority of the signal propagates through a body of trees (See (5) in [10]). The measurements for ITU-R were primarily made in the Ultra High Frequency (UHF) range. The loss factor is given by (4).

$$L_{ITU-R} = 0.2 f^{0.3} d^{0.6} \quad (4)$$

In (4), f is the transmission frequency (MHz), d is the distance between transmitter and receiver (m).

The Weissberger and ITU-R models do not account for seasonality (i.e. differing amounts of vegetation when trees have leaves or are bare). In order to account for this effect, the COST235 model [13] was developed based on measurements carried out over two seasons when trees are "in-leaf" and "out-of-leaf". Measurements were made in the millimeter-wave frequencies (9.6 to 57.6 GHz). The resulting parametrization is given by (5).

$$L_{COST235} = \begin{cases} 26.6 f^{-0.2} d^{0.5} & \text{out-of-leaf} \\ 15.6 f^{-0.009} d^{0.26} & \text{in-leaf} \end{cases} \quad (5)$$

In (5), f is the transmission frequency (MHz), d is the distance between transmitter and receiver nodes (m).

Several other models have been developed to take into account specific characteristics. In [14], the authors derive path loss as a function of trunk height gain k for a pine tree environment. The authors in [15] model RSSI inside a forest, based on factors such as the average density of trees, or the average trunk diameter. They find that the path loss coefficient decreases linearly with the average tree density multiplied by the trunk diameter. The authors in [16] use high-resolution LiDAR data to calibrate a log-normal path loss-model.

C. Comparison to Real Data

All of the canonical and empirical propagation models in Sections III-A and III-B are equations. Starting from ideal physics in the canonical models, the empirical propagation models fine-tune parameters for each equation (e.g. the system loss factor L) so the resulting equation better matches experimentally gathered data. These models use the distance between nodes as the only variable.

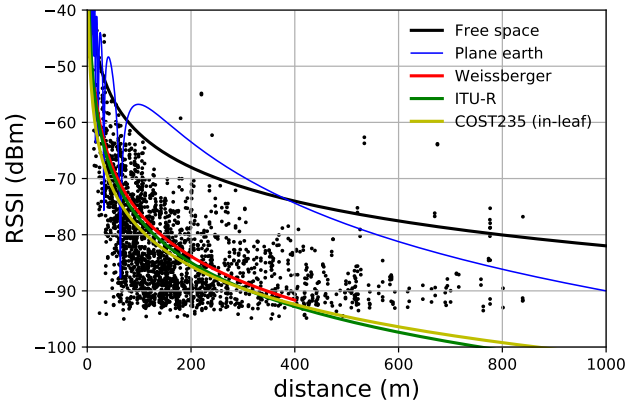

Fig. 4. Comparing the different canonical and empirical propagation models (lines) against the measurements gathered on the 2218 wireless links of the ARHO networks (cloudpoint). Canonical and empirical propagation models as details in Sections III-A and III-B, respectively; measurements gathered in Section II.

Since we have collected a large number of real-world RSSI measurements on the 2218 wireless links from the ARHO networks, we are interested in seeing how well the models surveyed in Section III are able to match the measurements. We plot in Fig. 4 the RSSI as a function of distance predicted by the different models, and overlay our measurements. In the models, we set the parameters to match that of our deployment: G_t = 4 dBi, G_r = 4 dBi, λ = 12.5 cm (for 2.4 GHz), $h_t = h_r$ = 5 m. We also use the "in-leaf" variant of the COST235 model, as it corresponds best to the deployment environment. Fig. 4 suggests the models do not match our empirical data well, and over-estimate the measured RSSI. Moreover, the shape of the signal strength decay is not exponential: at short distances, there is much greater signal strength variability than predicted by the models. The 1% sensitivity level of the LTC5800 chip used in the deployment is -93 dBm; which is the reason why there are no points below -95 dBm.

To quantify the error between the models and the measurements, we use each of the models on each of the 2218 wireless links in the dataset. We record the difference between the RSSI predicted by the models, and the average RSSI of that link over the year of measurement. Fig. 5 shows the results as a histogram. Table I contains the average and standard deviation of the prediction error for the distributions shown in Fig. 5.

It is clear from Fig. 4 and Table I that the

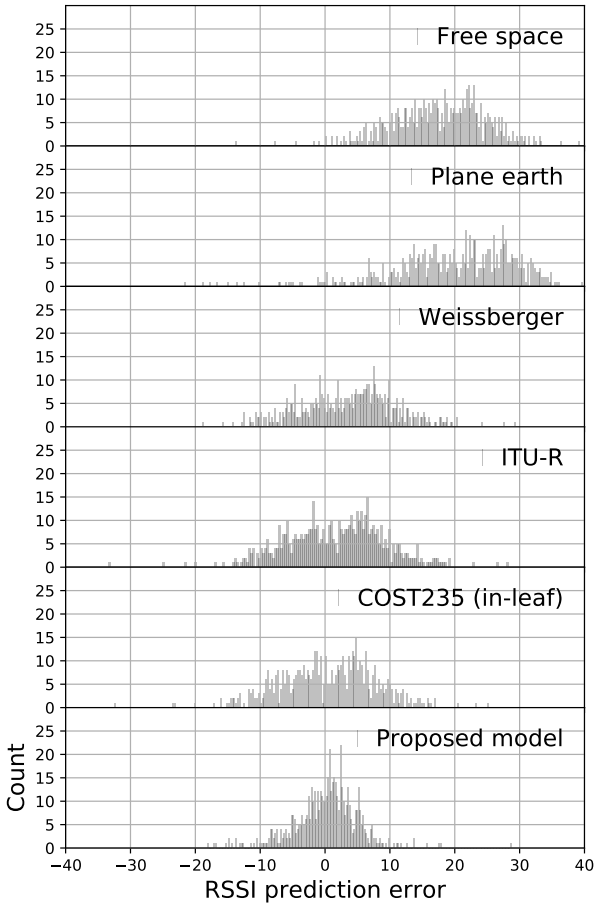

Fig. 5. Distribution of errors under canonical and empirical models (top panels), compared to proposed model (bottom panel) for year-averaged RSSI data.

TABLE I
PERFORMANCE EVALUATION OF THE PROPAGATION MODELS TO THE ARHO DATASET.

Model Name	Avg. Error (dBm)	SD Error (dBm)
Free space [17]	20.5	8.60
Plane earth [17]	17.8	6.92
Weissberger [11]	6.65	4.70
ITU-R [12]	6.37	4.64
COST235 [13]	5.91	4.37
Proposed model	3.72	3.41

models cannot be used as-is to accurately predict the RSSI between two nodes deployed in the field. Other phenomena besides distance and vegetation affect signal strength, including specifics about the environment we deploy in (e.g. terrain), or the hardware we use (e.g. transmission power, antenna matching, antenna alignment, radiation pattern). We could produce another model, in which we take into account these phenomena, and create an equation that best matches our data. We believe that such matching makes the model more specific to a particular set of empirically-gathered data, and less generally useable. We question whether using a simple equation is the right approach.

Our conclusion is that trying to model every physical phenomenon in an equation is a non-starter. Minute changes to the environment (e.g. node position [18], antenna alignment, the quality of the antenna connector) can cause the RSSI to vary by over 10 dB. Understanding, measuring and modeling each of these phenomena is unfeasible for any real deployment. Our intuition is that we should instead *learn from the wireless links deployed now and predict the performance of wireless links installed in the future*. Section IV develops this idea, and presents a model based on the Random Forest algorithm (a common algorithm used in machine learning). We show how this model achieves a 37% reduction in the average prediction error compared to the canonical/empirical model with the best performance.

IV. A Machine Learning Model

Given the limitations of the canonical and empirical models, we want to determine whether a multivariate connectivity model could be trained on the RSSI measurements and associated features described in Section II, and used to predict RSSI at un-instrumented locations. We describe a number of potential machine learning algorithms in Section IV-A, and the proposed model in Section IV-B.

A. Overview of Machine Learning

Machine-learning algorithms are trained to identify patterns in historical data. This is very different from the canonical and empirical models surveyed in Section III, which fit a specific function "a priori". Patterns learned by these algorithms can be non-linear, multivariate, and can be used both for predicting which category a piece of data belongs to (classification), and predicting continuously-valued outputs from a set of inputs (regression). Algorithms in machine learning are broadly divided into two categories: "supervised" and "unsupervised." Supervised algorithms require observations of the output to learn patterns. Unsupervised algorithms learn patterns in the space of independent variables without observations of the dependent variable (e.g. clustering). The field of machine learning has seen a number of recent applications to low-power wireless networking, including for localization and routing [19]. In this study, we have observations of the output (the RSSI), and try to predict a continuously valued function. We hence design a "supervised regression" machine learning solution.

A number of algorithms can be used to solve supervised regression problems, including: Support Vector Machines, Neural Networks, Nearest Neighbors, Gaussian Processes, and Random Forest. In, this study we evaluate four algorithms: Random Forest, Adaboost, Neural Networks, and K-Neareast-Neighbors. Decision Trees are considered to be one of the best off-the-shelf algorithms because they are not sensitive to independent variable scaling or the inclusion of irrelevant variables. [20] Also, decision trees are not "black-box" models: each split in the decision tree can be inspected once the model is trained. A single decision tree is known overfit data [21]. To address this, Random Forests combine estimates from multiple trees using a random selection of features to arrive to a consensus of the true output [22]. This process prevents the model from over-fitting the data (i.e. fitting the noise rather than the trend). The accuracy of the algorithm is affected by parameters of the estimator such as the maximum tree depth and the size of the ensemble. Decision tree depth controls the maximum depth of the decision tree (i.e. how many splits on the independent variables are made). The size of the ensemble is the number of decision trees the outputs are averaged over. In general, a small ensemble with deep decision trees has a greater tendency to overfit than a shallow ensemble of many decision trees. These parameters must be tuned for the RSSI model, which is discussed in Section IV-B.

Once the model is trained, its indicates which features are more important. This is calculated either by computing the out-of-bag sampling error (MSE) during training, then permuting each predictor vari-

able and computing the difference in sampling errors [22], or by computing how frequently a given feature is used to perform splits in the estimator.

B. Model Implementation and Parameter Tuning

We develop the connectivity model using Scikit-Learn version 18.1, an open-source machine learning package implemented in Python (http://scikit-learn.org/). We first divide the average annual RSSI data along the 2218 links into three subsets using randomized sub-sampling ("test train split" in Scikit-Learn). We use standard splitting ratios: training (50%), cross-validation (25%), testing (25%). Training and cross-validation sets are used to determine the optimal parameters for the models. The input features were scaled to values between 0 and 1 as K-Nearest-Neighbors and Neural Networks are sensitive to the scale of the input space. We use a grid search cross-validation scheme with 3 folds to determine the optimal parameters for each model.

For Random Forest we evaluate the following parameters: maximum tree depths between 5 and 50 (at 5-unit intervals), number of ensemble members between 10 and 130 (at 20-unit intervals), and between 1 and 7 features for each split. Splits are axis-alined. Mean squared error is used as the information gain criteria. All other values are set to defaults. For Adaboost, we evaluate linear, square, and exponential loss functions, and evaluate between 2 and 30 estimators (at 2-unit intervals). In the Neural Network, we use 1 hidden layer with between 2 and 20 neurons, between 20,000 and 100,000 maximum iterations (at 20,000-unit spacing). Four activation functions are considered: identity, logistic, tanh, and rectified linear (relu). For K-Nearest-Neighbors, we evaluate between 2 and 50 neighbors (at 1-unit intervals).

V. Model Validation and Results

To validate each model, we use data not used during the training process. We train the model using the best parameters determined in Section IV-B, and evaluate the model's accuracy on the cross-validation dataset selected at random from the available data. Table II shows the chosen parameters, along with the mean absolute error for each method. We discuss the accuracy of the chosen model on the cross-validation dataset in Section V-A. We then discuss the relative contribution of each independent variable in Section V-B.

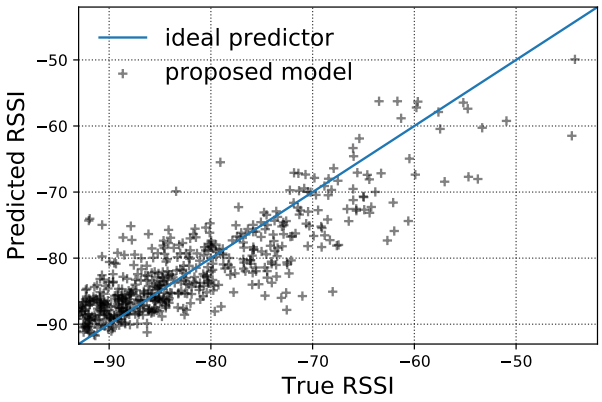

Fig. 6. RSSI predictor accuracy compared to an ideal predictor (blue line) on the testing dataset of 555 RSSI measurements.

A. RSSI Prediction Accuracy

The accuracy of the predictor on the 555-sample cross-validation dataset is depicted in Fig. 6. The blue line represents an ideal predictor; black points is the predicted data for the 555-sample cross-validation dataset. The R^2 score of the predictor is 0.51. The predictor exhibits a slight positive bias at very low values of RSSI (less than -85 dBm) and a slight negative bias at values greater than -85 dBm. We attribute the bias at low RSSI values to the fact that frames cannot be received with an RSSI well below the sensitivity; which causes the predicted RSSI to exhibit a slight positive bias. Overall, the predictor exhibits near-zero bias (0.18 dBm).

Fig. 5 shows how this predictor compares to the predictions done with the canonical and empirical RSSI models from Section III. All of the canonical and empirical propagation models exhibit positive bias on the testing dataset (the bias is 19.5, 17.7, 3.61, 1.75, and 0.49 dBm for plane earth, free space, Weissberger, ITU-R, and COST235, respectively). It should be noted that the comparison is limited to 2.4 GHz data. The models also show a higher mean absolute error and higher error variability (error standard deviation) than the proposed model (Table I). Of the existing models, the COST235 (in-leaf) model shows the highest accuracy. The proposed model, however, exhibits an average prediction error 37% lower than that of the COST235 (in-leaf) model.

TABLE II
OPTIMAL PARAMETERS AND ACCURACY FOR EACH MODEL CONSIDERED IN THE CURRENT STUDY.

Algorithm	**Optimal hyper-parameters**	**MAE (dBm)**
Random Forest	max depth: 20, max features: 4, number of estimators: 70	3.72
K-Nearest-Neighbors	number of neighbors: 4	5.10
Neural Network	activation: tanh, number of neurons: 4, 100,000 iterations	5.15
AdaBoost	loss function: linear, number of estimators: 12	5.55

TABLE III
INDEPENDENT VARIABLE IMPORTANCES INFERRED FROM THE ENSEMBLE REGRESSION (NORMALIZED).

Feature	**Mean importance**	**SD importance**
Path ground distance	0.47	0.10
Terrain complexity	0.15	0.08
Vegetation variability	0.10	0.04
Mean percent canopy	0.09	0.02
Path angle	0.08	0.03
Source canopy	0.05	0.01
Receiver canopy	0.05	0.01

B. Feature Contribution

One important aspect of this study, which is made possible by the machine learning approach, is to determine which additional independent variable ("feature") is important to predict the RSSI. Random Forests provide a natural ranking of features in the model, based on the degree to which splits on each variable improve the split quality criterion (mean squared error – MSE – in the current study). In Scikit-Learn, this value is computed based on how frequently a feature is used to split in the ensemble in order to determine a "mean importance" for each feature. Table III shows the normalized contribution of each independent variable used in the model. Path ground distance is the most important feature in the model, followed by terrain complexity, vegetation variability, and mean percent tree canopy. Attributes related to the local characteristics of each node (source/receiver canopy coverage) exhibit lower significance. Given that the model is trained only on 2.4 GHz data, it should be noted that the feature importances may change for different frequencies.

VI. DISCUSSION AND OPPORTUNITIES

This paper introduces a radical new way of thinking about wireless connectivity models. Instead of relying on an *expert understanding* of the physics of wireless propagation, we propose a *agnostic computational* approach in which patterns are identified in recorded data. The machine learning tools used are generic: they are not aware that the data is related to wireless connectivity.

The main result, which is counter-intuitive, is that this agnostic approach yields better results than the expert approach. The proposed approach reduces the average prediction error by 37%, when compared the expert model with the best performance. The quantitative study is based on a large connectivity dataset of 42,157,324 measurements gathered for one year on 2218 wireless links: we have confidence in the results. These findings open up many new possibilities for understanding, planning and diagnosing wireless networks.

A. Discussion

On top of the numerical results, we want to conclude this article with a discussion about the strengths and limitations of the proposed approach.

Being able to compute which features are important is a key benefit of the methodology developed. First, it confirms the importance of distance as a key feature for predicting the wireless connectivity between devices, but it also highlights that other features are important.

An immediate drawback is that these extra features are not common to all deployments. While the mean percent tree canopy cover is an important feature for our networks, it doesn't apply to for example a smart factory application. This means the model created during the learning phase only applies to the particular environment it was created in. That being said, the *methodology* can be used in all environments, and the same remark applies to canonical and empirical models.

Perhaps the main drawback of the approach is that one needs a lot of data to train the model. This has two main implications. First, one needs training data, which leads to a chicken-and-egg situation in new deployment environments. When deploying in a completely new environment, one needs to build

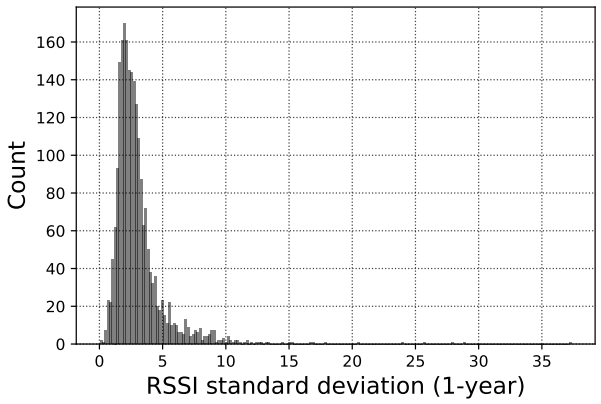

Fig. 7. Standard deviation of RSSI over all links over a one-year period.

up a dataset by measuring the connectivity between nodes that were deployed without assistance from a model. As the dataset grows, and as more and more nodes/networks are deployed, the model can be refined to start helping with the deployment. Second, one needs the dataset to be gathered in an environment that has similar distributions of independent variables as the environment used train the model. The 14 low-power wireless networks from the American River Hydrologic Observatory are similar in that they are deployed outdoor in a mountainous forested areas, and composed of the same devices and radio technology. Fig. 7 quantifies this similarity. It shows the standards deviation of the RSSI of the different links over the one-year period of the dataset. It is less than 5 dB for 90% of the of links. The model would be less accurate if the different networks would be deployed in very different areas.

Another limitation of the proposed approach is that the feature selection needs to be done well. Table III shows the 7 features the training data is annotated with. Even though the machine-learning approach is agnostic to the independent variable selection and can discard irrelevant variables, it still takes expert knowledge and "intuition" to select which feature to use.

Further complicating the problem, the importance of the feature can evolve depending on the setup. For example, the "path angle" feature has an mean importance of 0.08 (see Table III), which is low. This, in part, can be explained by the choice of using a lower-gain 4 dBi antennas for nodes on the slope. If a 8 dBi antenna were used, the spread of the vertical radiation pattern would be narrower, probably leading to an increased mean importance of the "path angle" feature.

To conclude the discussion, the model developed in this article achieves very good results, with a 37% reduction in the average prediction error compared to the canonical/empirical model with the best performance. That being said, the methodology is applicable only to cases where (1) training connectivity data have already been gathered for similar deployments and (2) there is some physical intuition about which features are important, and hence which features to annotate the data with.

B. Opportunities for Further Research

The methodology presented in this article is a radical new way of predicting connectivity in wireless networks. It opens up numerous opportunities of further research, including to understand/minimize the drawbacks highlighted in Section VI-A:

First, develop a methodology to guide feature selection. Such a methodology would guide feature selection in a systematic way, and reduce the amount of expert guidance needed. The goal would be limit the possibility of "missing" important features.

Second, employ next-generation remote sensing tools (e.g., airborne laser scanning, LIDAR) to address the limitations of the NLCD maps discussed in Section II. LIDAR data enables sub-meter modeling of the canopy structure. This would better represent the true density between each node, and would likely improve the accuracy of the machine-learning algorithm.

Third, evaluate how much training data is required to build a model with good enough accuracy, and how that accuracy evolves as the size of the data set increases. The ultimate goal is to be able to assess how the model behaves if the training data is built-up as more and more nodes are deployed. This study could be done with the same dataset, by considering data from an increasing number of wireless links.

Fourth, build a prediction placement tool. This study would use the model proposed in this article to optimize network topologies in complex terrain over the set of feasible signal repeater placements. This would facilitate the automated deployment of new networks to ensure they are robust to path-loss.

Fifth, apply the same methodology in different environments. This study would gather a similar connectivity dataset in a different environment (e.g. a smart factory), verify that the methodology applies equally, and quantify the difference in connectivity with the model presented in the present article.

Finally, expand the number of machine-learning models and parameters evaluated. This study offers a preliminary analysis of the most appropriate methods, but the model can be further refined by evaluating more algorithms and parameter optimization methods on the dataset provided with this study.

ACKNOWLEDGMENT

This work is partially supported by Inria through the REALMS associate team, the European Commission through the H2020 F-Interop and H2020 ARMOUR projects, the National Science Foundation (NSF) through a Major Research Instrumentation Grant (EAR-1126887), the Southern Sierra Critical Zone Observatory (EAR-0725097), California Department of Water Resources (Task Order UC10-3).

REFERENCES

[1] H. R. Bogena, M. Herbst, J. Huisman, U. Rosenbaum, A. Weuthen, and H. Vereecken, "Potential of Wireless Sensor Networks for Measuring Soil Water Content Variability," *Vadose Zone Journal*, vol. 9, no. 4, pp. 1002–1013, 2010.

[2] S. Pohl, J. Garvelmann, , J. Wawerla, and M. Weiler, "Potential of a Low-cost Sensor Network to Understand the Spatial and Temporal Dynamics of a Mountain Snow Cover," *Water Resources Research*, vol. 50, no. 3, pp. 2533–2550, March 2014.

[3] R. Rice and R. C. Bales, "Embedded-Sensor Network Design for Snow Cover Measurements around Snow Pillow and Snow Course Sites in the Sierra Nevada of California," *Water Resources Research*, vol. 46, pp. 1–13, 2010.

[4] S. Simoni, S. Padoan, D. Nadeau, M. Diebold, A. Porporato, G. Barrenetxea, F. Ingelrest, M. Vetterli, and M. Parlange, "Hydrologic Response of an Alpine Watershed: Application of a Meteorological Wireless Sensor Network to Understand Streamflow Generation," *Water Resources Research*, vol. 47, no. 10, pp. 1–16, 2011.

[5] Z. Li, N. Wang, A. Franzen, P. Taher, C. Godsey, H. Zhang, and X. Li, "Practical Deployment of an In-Field Soil Property Wireless Sensor Network," *Computer Standards & Interfaces*, vol. 36, no. 2, pp. 278–287, February 2014.

[6] J. Gutierrez, J. F. Villa-Medina, A. Nieto-Garibay, and M. A. Porta-Gandara, "Automated Irrigation System Using a Wireless Sensor Network and GPRS Module," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 1, pp. 166–176, 2014.

[7] Z. Zhang, S. Glaser, T. Watteyne, and S. Malek, "American River Hydrological Observatory: Monitoring the Sierra Nevada Snowpack," *IEEE Internet of Things Journal*, 2016, [under review].

[8] T. Watteyne, J. Weiss, L. Doherty, and J. Simon, "Industrial IEEE802.15.4e Networks: Performance and Trade-offs," in *IEEE International Conference on Communications (ICC)*, London, UK, 8-12 June 2015.

[9] C. G. Homer, J. A. Dewitz, L. Yang, S. Jin, P. Danielson, G. Xian, J. Coulston, N. D. Herold, J. D. Wickham, and K. Megown, "Completion of the 2011 National Land Cover Database for the Conterminous United States-Representing a Decade of Land Cover Change Information," *Photogrammetric Engineering and Remote Sensing*, vol. 81, no. 5, pp. 345–354, 2015.

[10] O. Kurnaz and S. Helhela, "Near Ground Propagation Model for Pine Tree Forest Environment," *International Journal of Electronics and Communications*, vol. 68, no. 1, pp. 944–950, 2014.

[11] M. A. Weissberger, "An Initial Critical Summary of Models for Predicting the Attenuation of Radio Waves by Trees," Department of Defense, Electromagnetic Compatibility Analysis Center, Annapolis, Maryland, USA, Tech. Rep. EST-TR-81-101, July 1982.

[12] I. T. U. (ITU), "Influence of Terrain Irregularities and Vegetation on Tropospheric Propagation," in *CCIR XVth Plenary Assembly*, vol. V: Propagation in Non-Ionised Media, no. ITU-R Report 236-6, Dubrovnik, Croatia, 1986.

[13] M. Hall, "Radiowave Propagation Effects on Next-generation Fixed-services Terrestrial Telecommunication Systems," European Cooperation in Science and Technology (COST), Tech. Rep. ICT COST Action 235, 1996.

[14] X.-M. Guo, X.-T. Yang, M.-X. Chen, M. Li, and Y.-A. Wang, "A Model with Leaf Area Index and Apple Size Parameters for 2.4 GHz Radio Propagation in Apple Orchards," *Precision Agriculture*, vol. 16, no. 1, pp. 180–200, 13 September 2015.

[15] J. A. R. Azevedo and F. E. S. Santos, "An Empirical Propagation Model for Forest Environments at Tree Trunk Level," *IEEE Transactions on Antennas and Propagation*, vol. 59, no. 6, pp. 2357–2367, 2 June 2011.

[16] S. Demetri, G. P. Picco, and L. Bruzzone, "Estimating Low-power Radio Signal Attenuation in Forests: A LiDAR-based Approach," in *International Conference on Distributed Computing in Sensor Systems (DCOSS)*. Fortaleza, Brazil: IEEE, 10-12 June 2015.

[17] Wiley-Blackwell, Ed., *Antennas and Propagation for Wireless Communication Systems*, 2nd ed. Wiley-Blackwell, 2007.

[18] T. Watteyne, S. Lanzisera, A. Mehta, and K. Pister, "Mitigating Multipath Fading Through Channel Hopping in Wireless Sensor Networks," in *International Conference on Communications (ICC)*. Cape Town, South Africa: IEEE, 23-27 May 2010.

[19] M. Abu Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications," *IEEE Communication Surveys & Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014.

[20] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Berlin, Germany: Springer Series in Statistics, 2001, vol. 1.

[21] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[22] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.