A Deep Hybrid Model for Weather Forecasting

Aditya Grover*
IIT Delhi
aditya.grover1@gmail.com

Ashish Kapoor
Microsoft Research
akapoor@microsoft.com

Eric Horvitz
Microsoft Research
horvitz@microsoft.com

ABSTRACT

Weather forecasting is a canonical predictive challenge that has depended primarily on model-based methods. We explore new directions with forecasting weather as a data-intensive challenge that involves inferences across space and time. We study specifically the power of making predictions via a hybrid approach that combines discriminatively trained predictive models with a deep neural network that models the joint statistics of a set of weather-related variables. We show how the base model can be enhanced with spatial interpolation that uses learned long-range spatial dependencies. We also derive an efficient learning and inference procedure that allows for large scale optimization of the model parameters. We evaluate the methods with experiments on real-world meteorological data that highlight the promise of the approach.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Machine Learning, Graphical Models, Weather Forecasting

Keywords

Gaussian Processes, Deep Learning

1. INTRODUCTION

Making inferences and predictions about weather has been an omnipresent challenge throughout human history. Challenges with accurate meteorological modeling brings to the fore difficulties with reasoning about the complex dynamics of Earth's atmospheric system. Methods have sought to define weather in terms of sets of fundamental quantities, and various characterizations have been proposed and employed in forecasting systems. We explore weather as a fundamental challenge for machine data mining and inference. We introduce methods that show promise for advancing the state of the art of weather forecasting systems.

In the U.S., the National Oceanic and Atmospheric Administration (NOAA) is responsible with providing publicly available weather forecasts, based on periodic observations. These measurements are logged in the Integrated Global Radiosonde Archive (IGRA) [4]. Forecasts for winds and temperature are accessible via NOAA's Winds Aloft program. To date, the best approaches to weather modeling rely on mathematical simulations. The methodology centers on the use of a generative model to capture atmospheric dynamics, where samples are drawn from physical simulations to make predictions [18, 12]. In contrast, we take a data-centric approach. Rather than define a generative model, we discriminatively train predictive models from the historic data, considering a historical data on a core set of variables for learning and inference about weather: atmospheric pressure, temperature, dew point, and winds. We use boosted decision trees as predictors in the studies.

Several challenges must be addressed in taking a data-centric approach to weather prediction. First, we note that the set of weather variables under consideration are tightly coupled. For example, pressure and temperature follow natural gas laws (i.e., the well-known formula, $PV = nRT$). Similarly, there is a tight relationship between relative humidity and temperature. Consequently, any model that jointly aims to predict the set of weather variables should leverage knowledge of the tight statistical couplings that are based in physics. Secondly, dependencies among the variables may have long-range influences across space and time. For instance, wind vectors across large geographic distances may follow isobaric contours. As another consideration, the weather phenomena may be affected by local geography and associated natural processes (e.g. isolated thunderstorms), as well as shifts in the large-scale structure of atmospheric phenomena (e.g. shifting of jet streams).

We aim to tackle these challenges via a representation that jointly predicts winds, temperature, pressure, and dew point across space and time. The proposed architecture combines a bottom-up predictor for each individual variable with a top-down deep belief network that models the joint statistical relationships. Another key component in the framework is a data-driven kernel, based on a similarity function that is learned automatically from the data. The kernel is used to impose long-range dependencies across space and to ensure that the inferences respect natural laws. We present an

*Research performed during an internship at Microsoft Research.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2783275.

efficient procedure for combining inferences from separate predictors of local phenomena while considering constraints imposed by the deep belief network such that the predictions respect the natural regularities expected with the large-scale phenomena.

The main contributions of this work can be summarized as follows:

1. We present a novel hybrid model with discriminative and generative components for spatiotemporal inferences about weather.

2. We design and implement a data-driven kernel function that shapes predictions in accordance with physical laws.

3. We provide an efficient inference procedure that enables optimization of the predictive model in accordance with large-scale phenomena.

4. We evaluate the methods with a set of experiments that highlight the performance and value of the methodology.

The rest of the paper is structured as follows: We next discuss background and related work. In Section 3, we describe the technical details of our approach, showing the components of a comprehensive graphical model that we call a *Deep Hybrid Model*. The learning and inference algorithms based on this model are discussed in Section 4. In Section 5, we present the results of experiments with the model on real-world data. We conclude with a brief summary and discuss future work in Section 6.

2. BACKGROUND AND RELATED WORK

The proliferation of satellites, radar, sensors, coupled with rapidly decreasing costs of storing and distributing information have catalyzed an explosion in quantities of weather data available for studies. Most work in weather forecasting to date rely on the use of generative approaches, where the weather systems are simulated via numerical methods [18, 12, 10], or rely on time-series analysis such as ARIMA models and simple classifiers based on Artificial Neural Networks [11, 10, 8, 2, 21] or Support Vector Machines [16, 19]. These statistical models often make strong assumptions such as spatial independence to overcome the curse of dimensionality, which do not hold well in practice.

Despite the success of machine learning in a variety of tasks, applications to the problem of weather forecasting has been limited. Exceptions include the use of Bayesian Networks for precipitation forecasts [3] and temporal modeling via Restricted Boltzmann Machines (RBM) [20, 15]. A separate thread of research has also focused on efficient representation of relational spatiotemporal data in Random Forests for prediction of severe surface-level weather processes, such as droughts and tornadoes [14, 13]. More recently, large-scale wind prediction has been presented [9] using a Bayesian framework with Gaussian Processes [17].

To date, uses of machine learning for weather prediction have been limited in several ways. First, almost all methods consider only one variable at a time and do not explore the joint spatiotemporal statistic of multiple weather phenomena. Also, to our knowledge, long-range spatiotemporal dependencies have not been modeled explicitly. Thus, models have been blind to long-range phenomena based on the laws of nature, such as winds aligning by pressure as captured by the structure and dynamics of isobars.

We introduce methods that address these limitations, via introduction of a hybrid representation. With a hybrid representation, individual predictors are discriminatively trained from historic data and local inferences from these models are combined with a deep neural network that overlays statistical constraints among key weather variables. We additionally apply a spatial interpolation scheme that respects constraints of long-range statistical dependencies. The methodology employs covariance matrix for Gaussian Process regression constructed from a large dataset. Here, the covariance matrix, also referred to as the kernel, allows us to enforce smoothness constraints over the weather variables. By ensuring that the kernel captures the dynamics of the system as informed by the training data, we are able to align estimates according to spatial constraints imposed by natural laws.

3. THE DEEP HYBRID MODEL

We seek a prediction model that respects spatiotemporal dependencies among weather variables induced by atmospheric physics. We test the framework with data drawn from a continental scale weather corpus composed of data captured via balloons. In particular, we consider the IGRA dataset consisting of balloon observations made at 60 stations across the U.S. These balloons transmit observations about wind speed and direction, temperature, geopotential height, dew point, and other weather variables. These observations are released in real time by the NOAA and later by the National Climatic Data Center following preprocessing. The data is eventually integrated into the curated IGRA dataset which is updated daily and contains historical weather data spanning decades compiled from eleven source datasets. Any data added to the archive undergoes a cycle of quality assurance to resolve potential inconsistencies among variables [4, 5].

Formally, we consider four weather variables in the model: wind velocity, $\mathbf{v}$; pressure, p; temperature, t and dew point, d. The wind observations are represented as a two-dimensional vector, $\mathbf{v} = [v^x, v^y]$ while all other weather variables are scalars. We represent weather stations (where the balloons are released) as $\mathbf{S}_L = \{\mathbf{s}_1,...,\mathbf{s}_{N_s}\}$ where N_s is the total number of weather stations. For each of these stations, we have historical weather data logged at a frequency of approximately six hours over several years.

Our approach to building the weather model was governed by the following guidelines:

1. Temporal mining: Our model should be able to identify and learn from recurring weather patterns over time.

2. Spatial interpolation: The dynamic influence of atmospheric laws on weather phenomena need to be accounted for in our predictions.

3. Inter-variable interactions: The local interdependencies between weather variables should be captured by our model.

Accordingly our model can be viewed as having three main components. The first component is a set of individual pre-

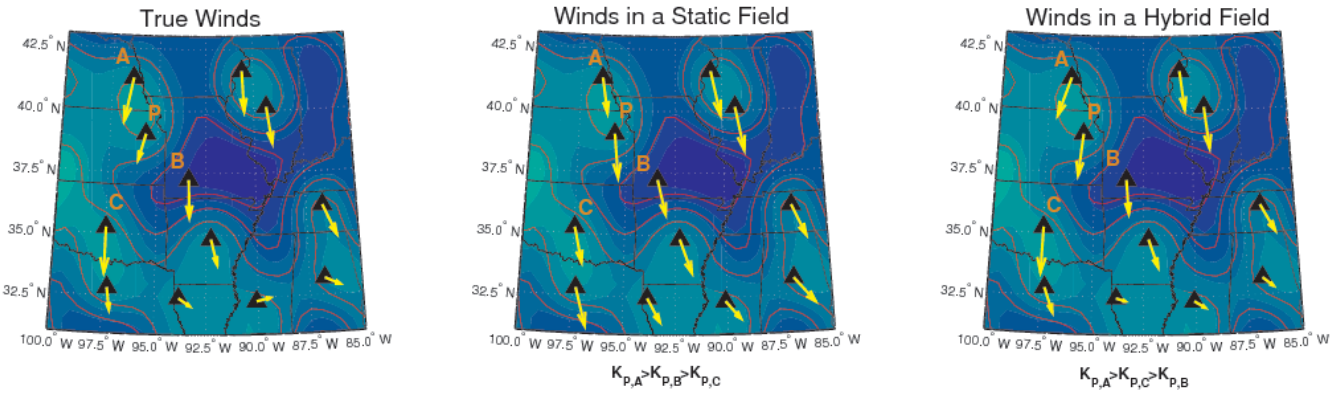

Figure 1: Spatial Interpolation of winds in a static and hybrid field. Filled contours represent temperature and isobar lines are marked in red. In the hybrid field, the interpolated wind vectors are closely aligned with the true values. However, the static field fails to account for the long-range dependencies.

dictors for the weather variables that are trained using historical data. A variety of off-the-shelf machine learning procedures can be applied to the recorded data to build these individual predictors. The second component works to refine inferences produced by the separate predictors by constraining the output to be spatially smooth and aligned with constraints imposed by physical laws. The interplay of these constraints is dynamic and hence, we develop a data-centric approach. The third component consists of a deep belief network which leads to a preference for solutions that respect the expected joint statistics of the weather variables. We describe the three key components in detail below and finally conclude this section with an integrated graphical model of our framework.

3.1 Base-Level Predictors

The base-level predictors are individual regression functions that are trained using historical data at different temporal granularities. The intuition is that long-term historical records of weather should provide insights about the weather at particular locations, given sets of observations in the immediate past. In general, the weather conditions change gradually over time and also exhibit cyclicity through seasons, consequently enabling some success in predicting the signals. We need to train different predictors for each station and range of altitudes considered as weather conditions change significantly across the vertical profile.

The performance of the local regressions depends critically on evidential features. We consider features over short- and long-term spans of time. For the short-term features, we consider the values of weather variables over the last seven days. As observations come at twelve-hour intervals, we consider shorter-term features by time of the day. Such short-term segmentation can be useful because winds, temperature, and other weather variables may differ significantly over day and night due to the influence of solar heating. We consider separate short-term features for day and night rather than averaging over the daily variation. Features spanning longer periods of time incorporate average seasonal data on weather variables. The long-term features are computed for several years in the past to reduce the influence of atypical weather phenomena. Given the set of engineered features, we use an ensemble of boosted decision-tree learners to make predictions.

3.2 Data-Centric Kernel for Spatial Interpolation

The individual predictors provide predictions only for particular locations (the weather stations), and we need to interpolate the results across larger spatial regions. To extend predictions in a smooth manner beyond the weather stations, we rely on smoothness constraints induced via the GP prior.

The covariance or kernel matrix K captures the notion of similarity among data points that are close in space and time and is the key in determining the accuracy of spatial interpolation. While static Radial Basis Function (RBF) kernels based on distance give reasonable estimates, they fail to capture the dynamics of the system. For instance, predictions about wind velocity at location $\mathbf{s}^*$, are not necessarily influenced similarly by weather at equidistant stations, per factors such as regional turbulence. We need to have an ability to capture a preferential bias towards classes of functions that respect certain physical constraints among the weather variables. The physical constraints include long-range spatial dependencies, such as wind vectors aligning with isobars[1] and modeling the direct relationship between pressure, temperature, and dew point due to natural gas laws.

We use a novel kernel defining a GP prior. For any pair of locations i and j, if the current pressure, temperature, and the wind direction are denoted as p, t and θ respectively, then we define our kernel as:

$$K_{i,j} = K^D_{i,j} \cdot K^\theta_{i,j} \cdot (\epsilon K^p_{i,j} + (1-\epsilon) K^t_{i,j}). \quad (1)$$

Here, $K^D_{i,j}, K^\theta_{i,j}, K^p_{i,j}$ and $K^t_{i,j}$ are RBF kernels over geographic distance, the angle of the wind, pressure and temperature respectively and ϵ is a tunable parameter such that $0 \leq \epsilon \leq 1$. The resulting kernel matrix would be positive semi-definite as the proposed kernel function is a linear combination and Hadamard product of kernels.

[1]Since we are not doing surface-level predictions, the effect of friction is negligible.

Multiple kernels are commonly used to integrate similarity notions from different sources [6]. In our case, the similarity between any two sites is a function of the geographic distance as well as the similarity in the weather variables. We note that the kernel K^θ over the wind direction plays a critical role in inducing long-range dependencies. As an example, consider two stations A and B with wind vectors $[a_x, a_y]$ and $[b_x, b_y]$, respectively. We are performing interpolation separately in X and Y directions. Hence, for any station, e.g., station A, we can assume independence in the two directions such that a neighboring station B can only induce an air flow change in a_x through b_x and similarly, a_y is only influenced by b_y. K^θ captures this intuition by defining an RBF over the angles made by the wind vectors with the corresponding axis for which the kernel matrix is defined. The balance between the pressure gradient force and Coriolis effect (geostrophic force) causes the winds to follow isobars. This implies that stations in close vicinity having similar pressure will have winds aligned in the same direction, justifying the contribution of K^p in computing the similarity.

3.3 Joint Modeling of Weather Variables

Weather variables are influenced heavily by the interaction of several factors. At the most fundamental level, these dependencies are based in the natural laws of thermodynamics. Approaches to inferences about weather relying on numerical simulation seek to characterize these dependencies analytically. However these interdependencies are complex and unpredictable, which explains the limited success of analytical techniques. At the same time, discriminative statistical analysis beyond temporal and spatial techniques described above, does not generalize well for domains with the dynamism of weather phenomena. For weather, it is natural to consider architectures that can automatically learn rich representations from raw data. Hence, we model the joint distribution between weather variables through a deep belief network (DBN).

The DBN consists of layers of stacked Restricted Boltzmann Machines (RBM) where the connections between any two layers of a RBM form a bipartite graph. The top layer of the DBN consists of five units corresponding to the normalized values of the latent weather variables (two units for representing 2D vector winds). We assume a Gaussian prior over these variables, such that $W_i \sim \mathrm{N}(m_i, d_i)$, each unit having a bias a_i. The primary level interactions between the variables give rise to a secondary set of features represented as the layer 2 hidden units, $\mathbf{H}$. Similarly, we can have another RBM below to capture the interactions between $\mathbf{H}$ and the layer 3 units, $\mathbf{G}$. The hidden units follow a Bernoulli distribution and have a biases b_j and c_k. The weights between the first two layers are $\mathbf{U}= [u_{ij}]$ and the next two layers are $\mathbf{V} = [v_{jk}]$. Several structural and tunable design parameters are involved, which we discuss in the next section.

3.4 Probabilistic Graphical Model

The graphical model for the proposed approach is shown in Figure 2. The matrix $\mathbf{W}$ is the collection of all the weather variables $\mathbf{w}_i$ denoting the true value at each location i. The observations $\mathbf{z}_i = \{v_i^x, v_i^y, p_i, t_i, d_i\}$ recorded at any of the sites is simply a noisier version of this true value. We use plate notation to show observations at N_s number of weather

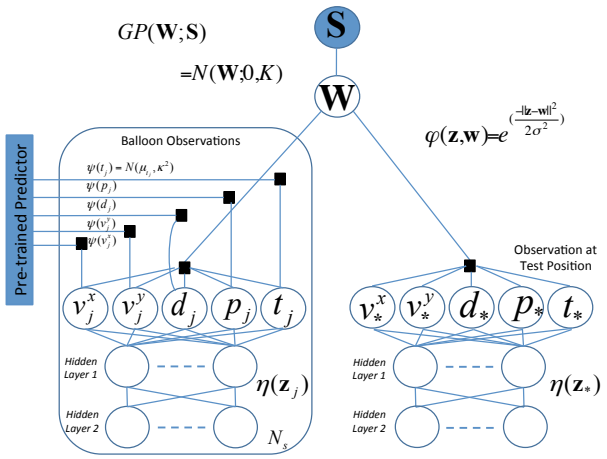

Figure 2: Deep hybrid model. Probabilistic graphical model for weather prediction where weather stations denoted by S, induce a Gaussian process (GP) prior over the true values of the weather variables W. Only noisier versions ($\mathbf{z}_i$) of the true values are observed at all the sites and are related via ϕ. The forecasts given by the pre-trained predictor are related to the future observations via the potential $\psi(\cdot)$. The joint distribution of the true weather variables is further constrained via a deep belief network ($\eta(\cdot)$). All potentials arise at a test site s^*, except that there is no pre-trained predictor.

stations. Each weather station has random variables for wind velocity, dew point, pressure and temperature. Each of these random variables are constrained via (a) an individual predictor that is trained on the historical data ($\psi(\cdot)$), (b) a Gaussian process prior ($GP(\cdot)$) and the Gaussian likelihood function $\phi(\cdot)$ that use data dependent prior to impose spatial and functional smoothness and (c) a deep belief network that encourages solutions that respect the joint statistics observed in historical data and that are also aligned with physical laws.

Similarly, at a test site s^*, we use the GP prior and the likelihood to first interpolate observations made at the weather stations. These interpolations are then further constrained via the deep belief network to impose the joint statistical distribution of the weather variables. Formally, we have the following distribution corresponding to the graphical model:

$$p(\mathbf{W}, \mathbf{Z}|\mathbf{S}) \propto GP(\mathbf{W}; \mathbf{S}) \prod_{i \in L \cup *} \phi(\mathbf{w}_i, \mathbf{z}_i) \eta(\mathbf{z}_i) \prod_{i \in L} \psi(\mathbf{z}_i).$$

Here, $\mathbf{Z}$ is the collection of random variables representing the observations at any of the locations. The terms $GP(\cdot)$ and $\phi(\cdot)$ enforce the smoothness constraints as defined by the data dependent kernel (described in section 3.2). The potential term $\eta(\cdot)$ arises due to the deep belief network component (Section 3.3). Finally, the term $\psi(\cdot)$ applies only to the weather station sites and enforces consistency with the prediction of the pre-trained regression functions. In particular, the observations are related to the output of an individual predictor via a simple Gaussian function: e.g. $\psi(p) = N(\mu_p, \kappa^2)$, where μ_p is the individual prediction for the pressure variable.

Algorithm 1 Deep hybrid model learning.

procedure TRAINWEATHERMODELS

▷ Boosted Decision Trees for Every Location, Variable
  for all $x \in \{\mathbf{v}, p, t, d\}$ **do**
    for all $\mathbf{s} \in \mathbf{S}$ **do**
      $trainData \leftarrow$ getTrainData(x, getHistData($\mathbf{s}$))
      $param \leftarrow$ getBestParam($trainData$)
      $BstDecTree[x, \mathbf{s}] \leftarrow$ TrainBDTree($param$)
    end for
  end for

▷ GP Hyperparameters for Every Weather Variable
  for all $x \in \{\mathbf{v}, p, t, d\}$ **do**
    $hyParam \leftarrow$ getBestHParam(x, getAllHistData())
  end for

▷ DBN joint model training through CD
  $DBNmodel \leftarrow$ ContDivergence(getAllHistData())

end procedure

Algorithm 2 Deep hybrid model inference.

procedure FORECASTWEATHERVARIABLE(x, $\mathbf{s}^*, \mathbf{Z}$)
▷ Prediction Variable: x, Test Site: $\mathbf{s}^*$, Observations: $\mathbf{Z}$

  if $\mathbf{s}^* \in \mathbf{S}$ **then**
    $tmp^* \leftarrow$ getBDTreePred($x, \mathbf{s}^*, \mathbf{Z}$)
   ▷ uses corresponding BstDecTree model from Alg. 1

  else
    for all $\mathbf{s_i} \in \mathbf{S}$ **do**
      $tmp_i \leftarrow$ getBDTreePred($x, \mathbf{s_i}, \mathbf{Z}$)

      $w_i \leftarrow$ DBNinference(x, tmp_i)
     ▷ uses DBN model from Alg. 1

      Append w_i to $\mathbf{w}$
    end for

    $tmp^* \leftarrow$ GPinterpolate($x, \mathbf{s}^*, \mathbf{w}, \mathbf{Z}$)
  end if

  $w^* \leftarrow$ DBNinference(x, tmp^*)
  return w^*

end procedure

4. ALGORITHMIC DETAILS

To make the *Deep Hybrid Model* work in practice, we need to learn several parameters pertaining to the three components and design an efficient inference procedure for testing. Here we note that since we operate our model in batch mode, we can afford to have an elaborate learning procedure. Specifically, the deep belief network component indeed has high training time requirements. On the other hand, since our forecasts are made in real time, inference at test time needs to be extremely efficient. We now discuss the learning and inference algorithms which achieve these objectives.

4.1 Learning

Given the historical observations at various weather stations, we train the various components of our model in order to get the best predictive capability. We perform piecewise training of individual components, where the individual predictors, the parameters of the DBN and the kernel hyperparameters of the GP kernel are estimated. A simplified workflow for the training procedure is given in Algorithm 1.

In particular, we trained Boosted Tree-based Learners using the set of short- and long-term features described previously and used the best models that were obtained for each weather station in the U.S. for a range of altitudes from 3000 feet up to 39000 feet with an interval of 3000 feet. The optimal parameters with regard to the number of leaves, number of iterations, and the learning rate, were obtained through analysis with a 10-fold cross-validation study.

Similarly, the hyperparameters of the data driven kernel were set via 10-fold cross validation and the final values of the kernel bandwidths were set to 150, 0.1, 0.05 and 1 for distance, wind angle, temperature and pressure respectively and ϵ was set to 0.2. Finally, the DBN component of our model was trained via a standard contrastive divergence procedure [7]. We explored the following parametric ranges: the learning rate (0.1-0.01), the number of greedy iterations of convergence divergence (1-100) and the batch size (10-1000). The structural properties of the neural net such as the number of hidden layers (1-3) and the number of neurons in each hidden layer (50-500) were also experimented with. The configurations yielding best cross validation results comprised of two stacked RBMs consisting of 50 and 150 hidden neurons, trained with a learning rate close to 0.1, batch size of 100 and 20 greedy iterations. The limitation to these set of parameters is purely because of the high computational requirements and engineering effort in training deep networks, and indeed, the gains could be potentially more significant if the deep belief network is trained over a richer range of parameters.

4.2 Inference

Given the trained components, we seek to determine the posterior distribution over the set of observations $\mathbf{z}^*$ at the test site $\mathbf{s}^*$. Exact inference in the proposed model is hard due to the presence of potential functions $\eta(\cdot)$ induced via the deep belief network. We apply piecewise approximate inference as illustrated in Algorithm 2. For the trivial case, when the prediction needs to be made at a weather station site, we invoke the pre-trained predictor models to provide a forecast which is then refined using the deep belief network. If, however, we need to make a prediction at an arbitrary test site, the refined estimates computed for all weather stations are interpolated. These refined estimates are then interpolated to the test site via the Gaussian Process component. We note that, since the kernel function is data driven, we use simple interpolated values of the weather variable at the test site in order to compute the kernel. Given the interpolated values at the test site, we then carry out a last refinement of prediction in order to resolve the estimates with the joint statistical constraints imposed by the deep model.

At the heart of the inference scheme, we employ an iterative procedure that aligns the predicted estimates with the potential induced via the deep model. We use a variational approximation to resolve and refine the posterior distribution over the observations $\mathbf{z}$. Formally, we denote the approximation of the refined posterior by $q(\mathbf{z}_i) \sim \mathrm{N}(\mu_i, \sigma_i)$.

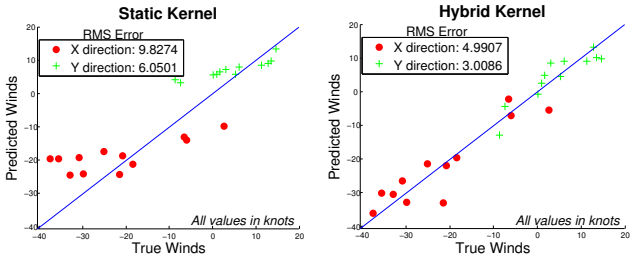

Figure 3: True versus interpolated wind plots for static and hybrid kernel. Static interpolation shows high deviations from true winds. Atmospheric dynamics are more effectively captured with use of a hybrid data-centric kernel.

Additionally, we also approximate the posterior over the latent variables and $q(H_j) \sim \text{Bern}(\gamma_j)$ and $q(G_k) \sim \text{Bern}(\beta_k)$ for the two hidden layers respectively. The following variational updates are then used to estimate the parameters of the distribution (l^{th} component):

$$\text{Layer 1 to 2:} \quad 1/\gamma_l^{t+1} = 1 + e^{-[b_l + \sum_i \mu_i^t u_{il}]}$$

$$\text{Layer 2 to 3:} \quad 1/\beta_l^{t+1} = 1 + e^{-[c_l + \sum_j \gamma_i^{t+1} v_{jl}]}$$

$$\text{Layer 3 to 2:} \quad 1/\gamma_l^{t+1'} = 1 + \frac{1-\gamma_l^{t+1}}{\gamma_l^{t+1}} e^{-[b_l + \sum_k \beta_k^{t+1} v_{lk}]}$$

$$\text{Layer 2 to 1:} \quad \mu_l^{t+1} = (m_l + d_l^2 a_l + d_l^2 \sum_j \gamma_j^{t+1'} u_{lj})/(1+d_l^2)$$

$$\sigma_l^{t+1} = d_l/\sqrt{1+d_l^2}$$

The mean parameters μ_l are initialized to the estimates m_l, while γ and β are initialized randomly. The parameter d_l corresponds to the variance in the initial estimates and signifies our confidence in those predictions. We set these variances via a cross-validation procedure over historical data. The derivation for the above update equations follows from the application of prior work in variational inference [1] to deep belief networks.

5. EXPERIMENTAL EVALUATION

We performed a set of experiments to evaluate the proposed methodology. In the experiments, we explored three main questions. First, we compare and highlight the advantage of the spatial interpolation procedure that relies on a data-centric dynamic kernel matrix to the more commonly used static kernel matrix. Second, we seek to compare the proposed model with a baseline approach. Third, we explore the importance of modeling the joint statistics of predictive variables via the deep belief network. Finally we compare the wind forecast results with those of state-of-the-art systems.

The experiments were based on five years of historical data, from 2009 to present, extracted from the IGRA dataset. The data consists of balloon observations recorded at 60 locations across the continental US.

5.1 Interpolation in a Hybrid Field

To illustrate the efficacy of a hybrid kernel in handling long-range spatial dependencies among weather variables, we considered a cluster of stations spread across the central U.S. (states demarcated by black lines in Fig. 1) and interpolated winds via Gaussian Process regression at these stations, considering the wind measurements from the rest of the U.S. stations. Thus, each station served as an independent test point, whose value is interpolated using a GPR model and compared against the true winds shown in Fig. 1 (a). Fig. 1 (b) shows the interpolated wind vectors when a static kernel matrix is used. Here an entry $K_{i,j}$ in the matrix is simply a decreasing exponential in the geographical distance between two stations i and j. In contrast, the hybrid approach, as illustrated in Fig. 1 (c), captures the similarity between each pair of stations such that every entry $K_{i,j}$ of the matrix is computed dynamically at training time using the formula given in Eq. 1. The pressure, temperature, and angle θ between the wind vectors are the known values for the current time step.

Now consider the following stations: Topeka (Kansas), Omaha (Nebraska), Springfield (Missouri) and Norman (Oklahama), referred to in Fig. 1 as stations P, A, B and C, respectively. For a static interpolation of winds at P, a higher contribution would come from B than C, as B is geographically closer. However, the temperature and pressure conditions at C are closely aligned to that of P and end up contributing more in the hybrid approach. We observe that $K_{P,A}$ is maximum in both cases. Hence, the hybrid kernel does not ignore distance as a similarity criteria. However, in cases involving a tradeoff between distance and other weather variables, their combined contribution might alter the relative importance of a particular neighboring station, as in the aforementioned case. The quantitative gains in predication accuracy are displayed in the RMS plots in Fig. 3 (a, b).

Weather Variable	RMS Error Reduction (in %age)		
	X	Y	Overall
6 hours	2.17	2.05	2.11
12 hours	1.05	1.01	1.03
24 hours	1.05	0.97	1.01

Table 1: Improvement in performance obtained using the deep belief network. The final step of refinement uses the DBN results to further improve prediction accuracy.

5.2 Dynamic Prediction and Deep Learning

In another experiment, we evaluate the performance gains due to the final refinement step of the deep belief network. The percentage reduction in error for wind forecasts for three time steps in the future are shown in Table 1. We see that the DBN leads to an additional 1-2% error reduction and clearly, modeling the joint statistics of the weather variables helps in making better predictions. We observed a performance improvement of similar magnitude for the other weather variables as well.

After establishing the superiority of the data-centric kernel and the DBN independently, we evaluate the prediction accuracy of the full deep hybrid model for each weather variable[2], aggregated over all stations in the continental U.S., where current and historical data is available.

[2]The IGRA dataset provides the geopotential height and dew point at roughly constant pressures. These quantities,

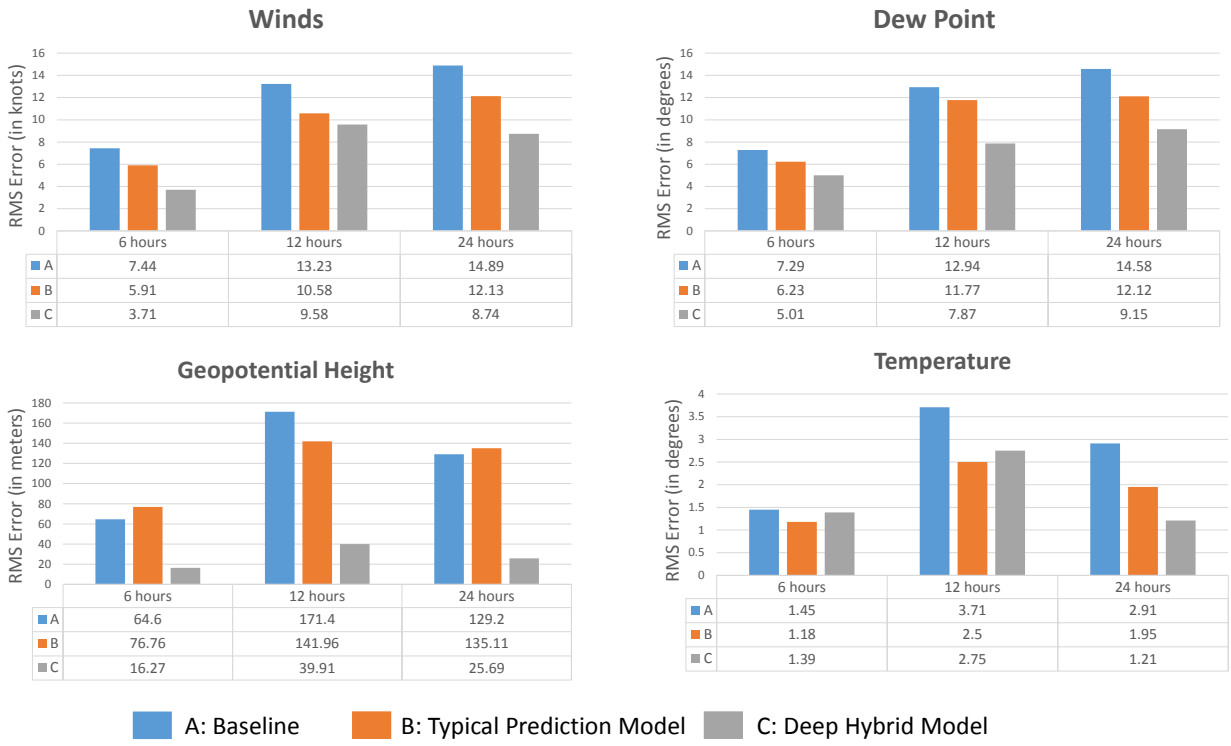

Figure 4: Results on predicting weather variables for different approaches. The temporal predictors that employ a hybrid data-centric scheme for interpolation and use DBNs for modeling the joint relationship among weather variables show significant improvements over the baselines.

The accuracy of the proposed model is compared with two baseline models in Fig. 4 for three future time steps as before. The baseline prediction, marked as **A** in Fig. 4, uses the current values as estimates for the future; the intuition with using current values to predict the future is that weather conditions typically do not change greatly over a day. For the second baseline (marked **B**), we construct a typical spatiotemporal prediction model, where baseline boosted decision tree predictors are augmented with a static interpolation scheme. We observe that the deep hybrid model (marked **C**) comprising of a dynamic data-driven interpolation scheme and a DBN in addition to the boosted decision tree predictors used in **B**, significantly outperforms both the baselines. In a couple of cases involving short-term temperature forecasts, **B** marginally outperformed the DHM, suggesting limited interdependences between temperature and other variables for short-term predictions.

5.3 Comparison with State of the art

Apart from winds, forecasts for other variables across the vertical atmospheric profile are not available for comparative analyses. We compare the wind predictions of the proposed model against two forecast systems. The first one proposed by [9] makes predictions using a static GPR interpolation scheme, coupled with relative velocity data obtained through airplanes. Our second set of comparisons is with the Winds Aloft forecast, released by NOAA every six hours, for three time steps into the future: 6, 12 and 24 hours. Table 2 show the accuracy of the two forecast systems for the Seattle station. The results summarize the predictions made for the weather stations in the state of Washington for a period of one month. We observe that, while Kapoor et al. 2014 achieve better performance than NOAA, the proposed method shows significantly better performance than both of the competitors.

under reasonable assumptions, serve as proxies for pressure and specific humidity, respectively.

Time Step	Model	RMS Error (in knots)		
		X	Y	Overall
6 hours	Deep Hybrid Model	**2.29**	**1.33**	**1.81**
	Kapoor et al. 2014	3.94	2.16	3.05
	NOAA	3.18	3.44	3.31
12 hours	Deep Hybrid Model	**4.44**	**2.59**	**3.56**
	Kapoor et al. 2014	5.03	3.93	4.48
	NOAA	5.13	4.34	4.88
24 hours	Deep Hybrid Model	**6.57**	**3.82**	**5.19**
	Kapoor et al. 2014	8.93	5.24	7.08
	NOAA	8.79	6.37	7.58

Table 2: Comparison of the proposed methodology with state of the art in wind prediction. Results summarized here are for weather stations in Washington for a period of one month. We observe that the new model results in significantly lower errors than competitive models. The best performance is indicated in bold.

6. CONCLUSION AND FUTURE WORK

We presented a weather forecasting model that makes predictions via considerations of the joint influence of key weather variables. We introduced a data-centric kernel and showed how using GPR with such a kernel can effectively interpolate over space, taking into account weather phenomena such as turbulence. We performed temporal analysis using short- and longer-term features within a gradient-tree based learner. We augmented the system with a deep belief network and tuned the parameters to model the dependencies among weather variables. A set of experiments on real-world data shows that the new methodology can provide better results than NOAA benchmarks, as well as recent research that had demonstrated improvements over the benchmarks.

Future work includes projecting weather predictions to more distant times into the future. We are also interested in exploring the use of computations of the value of information to guide sensing at weather stations. We note that airplanes in flight can serve as sensors of wind speeds, as explored in [9]. We wish to investigate the boosts in predictive power that might be achieved via integrating such additional data into the hybrid model.

Acknowledgments

We are grateful to Imke Durre for answering our queries concerning the IGRA dataset. The first author would like to thank Microsoft Research, Redmond for conducting the Worldwide Internship Program that made this research possible.

7. REFERENCES

[1] M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.

[2] L. Chen and X. Lai. Comparison between ARIMA and ANN models used in short-term wind speed forecasting. In *Power and Energy Engineering Conference (APPEEC), 2011 Asia-Pacific*, pages 1–4. IEEE, 2011.

[3] A. S. Cofino, R. Cano, C. Sordo, and J. M. Gutierrez. Bayesian networks for probabilistic weather prediction. In *15th Eureopean Conference on Artificial Intelligence (ECAI)*, 2002.

[4] I. Durre, R. S. Vose, and D. B. Wuertz. Overview of the Integrated Global Radiosonde Archive. *Journal of Climate*, 19(1):53–68, 2006.

[5] I. Durre, R. S. Vose, and D. B. Wuertz. Robust automated quality assurance of radiosonde temperatures. *Journal of Applied Meteorology and Climatology*, 47(8):2081–2095, 2008.

[6] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[7] G. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[8] I. Horenko, R. Klein, S. Dolaptchiev, and C. Schütte. Automated Generation of Reduced Stochastic Weather Models I: simultaneous dimension and model reduction for time series analysis. *Multiscale Modeling & Simulation*, 6(4):1125–1145, 2008.

[9] A. Kapoor, Z. Horvitz, S. Laube, and E. Horvitz. Airplanes aloft as a sensor network for wind forecasting. In *Proceedings of the 13th international symposium on Information Processing in Sensor Networks (IPSN)*, pages 25–34. IEEE Press, 2014.

[10] V. M. Krasnopolsky and M. S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks*, 19(2):122–134, 2006.

[11] R. J. Kuligowski and A. P. Barros. Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Weather and Forecasting*, 13(4):1194–1204, 1998.

[12] G. Marchuk. *Numerical methods in weather prediction*. Elsevier, 2012.

[13] A. McGovern, D. John Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams. Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(4):407–429, 2011.

[14] A. McGovern, T. Supinie, I. Gagne, M. Collier, R. Brown, J. Basara, and J. Williams. Understanding severe weather processes through spatiotemporal relational random forests. In *2010 NASA conference on intelligent data understanding*, 2010.

[15] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee. Structured Recurrent Temporal Restricted Boltzmann Machines. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1647–1655, 2014.

[16] Y. Radhika and M. Shashi. Atmospheric temperature prediction using support vector machines. *International Journal of Computer Theory and Engineering*, 1(1):1793–8201, 2009.

[17] C. E. Rasmussen. Gaussian processes for machine learning. 2006.

[18] L. F. Richardson. *Weather prediction by numerical process*. Cambridge University Press, 2007.

[19] N. I. Sapankevych and R. Sankar. Time series prediction using support vector machines: a survey. *Computational Intelligence Magazine, IEEE*, 4(2):24–38, 2009.

[20] I. Sutskever, G. E. Hinton, and G. W. Taylor. The Recurrent Temporal Restricted Boltzmann Machine. In *Advances in Neural Information Processing Systems*, pages 1601–1608, 2009.

[21] C. Voyant, M. Muselli, C. Paoli, and M.-L. Nivet. Numerical Weather Prediction (NWP) and hybrid ARMA/ANN model to predict global radiation. *Energy*, 39(1):341–355, 2012.