FINAL REPORT ON CHALLENGE #4: High-precision meteorological forecasting for optimizing agriculture and beyond

Name of mentor(s): Miroslav Čepek, Michal Kepka
Number of participants: 4

INTRODUCTION

Background of the Challenge

The agricultural sector is fundamentally dependent on accurate and timely weather information. Weather conditions such as temperature, dew point, wind speed, precipitation, and humidity critically influence critical operational decisions. Farmers and agricultural managers utilize forecasts to determine when to plant seeds, irrigate crops, apply fertilizers and pesticides, or schedule harvests. Even slight improvements in forecast accuracy can translate into significant economic and environmental benefits. For instance, accurate temperature and dew point forecasts can help farmers anticipate frost conditions and protect crops more effectively. Similarly, predictions of wind speed can guide irrigation and pesticide spray schedules, reducing chemical drift and wastage.

However, producing accurate local forecasts is an ongoing challenge. Global weather models, while advanced, often operate at resolutions of tens of kilometers, which is insufficient for capturing local microclimates. Traditional solutions like building proprietary weather stations for hyper-local conditions often struggle with data limitations or lack of integration with global atmospheric patterns. This results in suboptimal decision-making and missed opportunities to optimize yields, reduce environmental impacts, and enhance resilience to weather extremes.

Scope of the Effort

The project aims to combine data from local weather stations, global weather models, and historical reanalysis products to develop an advanced modeling framework capable of delivering highly accurate local forecasts. Specifically, the effort focuses on 24-hour lead-time predictions for key parameters—temperature, dew point, and wind speed—over carefully selected stations in the Czech Republic. The methods explored include a variety of machine learning (ML) models such as Multilayer Perceptrons (MLPs), Long Short-Term Memory (LSTM) networks,

CatBoost gradient boosting, and exploratory Bayesian Neural Fields. These techniques are integrated with global forecast model outputs from the Global Forecast System (GFS) and enhanced through reanalysis data sets like ERA5-Land.

By leveraging state-of-the-art ML methods and data fusion techniques, this project's objective is to push the boundaries of forecast accuracy at the local level. The implications extend beyond the Czech Republic, serving as a blueprint for other regions and microclimates. Ultimately, this integrated approach is intended to help farmers better manage their resources, reduce operational costs, increase crop resilience, and support sustainable agricultural practices.

METHODOLOGY

Team Description and Coordination

The initiative was driven by a multi-institutional consortium blending academic, industry, and applied meteorology expertise. The core team at the Faculty of Information Technology, Czech Technical University in Prague, provided the computational infrastructure, domain modeling expertise, and algorithmic innovation. Experts from CIIRC furnished the project with operational perspectives, in-field data logistics, and continuous feedback loops. The collaboration ensured that research efforts were aligned with practical needs, enabling iterative refinements that bridged the gap between theoretical modeling and real-world agricultural decision-making.

External partnerships included sporadic consultations with meteorological agencies, reanalysis data providers, and subject-matter experts in agriculture and climatology. These interactions guaranteed adherence to best practices, improved understanding of data peculiarities, and ensured that the methodology remained state-of-the-art.

Technical Background

Global-scale numerical weather prediction models (like GFS) provide a robust starting point for large-area forecasts but suffer from limited spatial resolution and inherent uncertainties. Local weather stations, while precise, only measure conditions at a single point location. The key challenge is how to scale from a coarse global grid to a highly localized prediction. Reanalysis data sets (e.g., ERA5-Land) provide a retrospective, physically consistent representation of the atmosphere and land surface, offering enhanced accuracy and resolution that can serve as a quality benchmark or training reference.

Machine learning methods excel at integrating multiple data sources and capturing complex, nonlinear relationships. Techniques range from tree-based models (like

CatBoost) to sophisticated neural networks (MLPs, LSTMs). Additionally, recent advances in probabilistic modeling—such as Bayesian Neural Fields—allow for encoding uncertainty and continuous spatial dependencies, potentially improving trust and interpretability of local forecasts.

Description of the Process of Solution

1. **Data Integration and Preprocessing:**
    - **Local Data (HadISD):** Hourly records from 27 selected stations in the Czech Republic provided ground-truth local measurements. Variables included temperature, dew point, wind speed, precipitation totals, cloud cover, and pressure.
    - **Global Data (GFS):** Global Forecast System data were aligned with station coordinates. For each station, GFS predictions relevant to the next 24 hours were extracted. This provided a suite of coarse predictors representing large-scale atmospheric conditions.
    - **Reanalysis Data (ERA5-Land):** While ERA5-Land is not available in real-time, it represents an accurate, high-resolution baseline. ERA5-Land served as a target to train a superresolution mapping from GFS to ERA5-like fields, enhancing the fidelity of input features.
2. **Feature Engineering:**
   After mapping GFS grids to each station, the local measurements and GFS forecasts were merged into a single training set. Key parameters driving forecast accuracy were identified, including helicity, surface temperature, precipitable water, and local station dew point. The inclusion of additional atmospheric variables ensured a rich feature space.
3. **Model Development: Baseline Approaches:**
    - **Persistence:** Using the last measured station value as a 24-hour forecast.
    - **Direct GFS:** Using GFS forecast values directly without ML post-processing.
4. **Machine Learning Models:**
    - **CatBoost:** A gradient boosting method on decision trees, chosen for its robustness, high accuracy, and ability to handle a large feature set efficiently. CatBoost also provided insights into feature importance through SHAP values, guiding further optimization.
    - **Multilayer Perceptron (MLP):** A fully connected neural network capturing complex input-output relationships in a feedforward manner.
    - **LSTM Networks:** Exploiting temporal sequences, LSTMs captured memory of recent station conditions, improving the understanding of evolving atmospheric patterns.

5. **Bayesian Neural Fields (Exploratory):**
   Bayesian Neural Fields treat spatial forecasting as learning a continuous function over space and time, with uncertainty quantification built-in. By defining a prior over function space and updating beliefs with observed data, these fields potentially provide better uncertainty estimates. The approach allows the model to better handle areas with sparse station coverage or rapidly changing conditions. Although still under exploration, Bayesian Neural Fields might refine forecasts where deterministic methods struggle.
6. **Incorporation of Estimated ERA5-Land Data:** A U-Net model was trained to transform GFS forecasts into ERA5-Land-like estimates. The U-Net, with its encoder-decoder architecture and skip connections, effectively learned superresolution mapping. By feeding these U-Net-derived ERA5-Land approximations into the ML models, the training data became more spatially coherent and potentially closer to ground truth conditions. This step aimed to leverage ERA5's high-resolution realism to enhance local forecasts without waiting for actual ERA5-Land data.

Data & Equipment List

- **Data:**
  - HadISD station data (historical hourly measurements)
  - GFS operational forecasts
  - ERA5-Land reanalysis data (2015-2021 for U-Net training)
- **Equipment and Software:**
  - GPU-accelerated servers for deep learning (e.g., NVIDIA Tesla V100 or A100)
  - Python ecosystem (NumPy, Pandas, xarray, TensorFlow, PyTorch)
  - CatBoost library for gradient boosting
  - Docker or Singularity containers for reproducible computing environments
  - Git-based version control and CI/CD pipelines for continuous model integration and testing
- **External Services:**
  - Institutional HPC clusters for parallel processing
  - Secure cloud storage for data and model artifacts
  - Visualization and BI dashboards to share results with stakeholders

Detailed Implementation Plan

- **Phase 1 (2-3 Months):**
  Data acquisition, cleaning, and synchronization. Implement scripts to align station data with corresponding GFS forecasts. Establish ERA5-Land and

GFS interpolation routines. Begin preliminary training of U-Net for superresolution mapping.

- **Phase 2 (3-4 Months):**
  Train and refine ML models (CatBoost, MLP, LSTM) on GFS+station data. Conduct hyperparameter tuning and validate results on a held-out portion of 2022 data. Begin integrating U-Net output into the ML frameworks. Initiate experiments with Bayesian Neural Fields to incorporate uncertainty modeling.
- **Phase 3 (1-2 Months):**
  Evaluate final model configurations on 2023 validation data. Compare the baseline, ML-only, and ML+ERA5-Land approaches. Quantify the incremental benefits of Bayesian Neural Fields in terms of both accuracy and uncertainty representation.
- **Phase 4 (Ongoing):**
  Operationalize the best-performing model, create APIs for external stakeholders (e.g., vineyard managers), and maintain a continuous improvement cycle. Monitor forecast accuracy in real-time, re-train models as new data accumulate, and iteratively incorporate feedback from users.

Analysis of Needs of Stakeholder Groups

- **Farmers and Agricultural Managers:**
  Require precise, location-specific forecasts to optimize labor and resource inputs. Smaller, family-run farms depend heavily on short-term forecasts to decide on delicate operations like frost protection or targeted irrigation.
- **Agricultural Advisories and Cooperatives:**
  Need reliable predictions to guide member farms, offer best practice recommendations, and ensure synchronized harvesting and storage to prevent crop spoilage.
- **Policy Makers, Insurers, and Agricultural Economists:**
  More accurate forecasts can inform policy decisions, risk assessments, and insurance premium calculations. They facilitate planning for weather-related disasters and support mitigation strategies.
- **Meteorological Services and Data Providers:**
  Gain insights on how station-level data can refine global forecasts, potentially improving national and regional weather services.

Experimental Results

When tested on 2023 data, the advanced ML models significantly outperformed both the persistence and direct GFS baselines. For temperature forecasting, incorporating GFS and local station data reduced the mean absolute error (MAE) from over 2°C (GFS alone) to about 1.07°C (CatBoost). Similar improvements were observed for

dew point and wind speed. The introduction of ERA5-Land-like features from the U-Net provided marginal improvements, especially in stable atmospheric conditions.

Initial trials with Bayesian Neural Fields revealed promising avenues for capturing forecast uncertainty and spatial variability. Although not yet achieving substantial accuracy improvements beyond the deterministic CatBoost or MLP models, Bayesian Neural Fields displayed more robust performance in uncertain scenarios and offered interpretable uncertainty estimates. This probabilistic approach may prove invaluable as the operational context expands or when data quality varies.

FINDINGS & CONCLUSION

Discussion of the Results and Findings

The key finding is that machine learning frameworks, leveraging both local station measurements and large-scale forecasts, can dramatically enhance local forecast accuracy. The CatBoost model proved particularly strong, likely due to its ability to handle complex interactions in a large feature space. Neural network approaches (MLP, LSTM) also provided improvements, especially in capturing temporal dynamics with LSTMs. The superresolution step using U-Net to incorporate ERA5-Land features improved model fidelity slightly, indicating that data enhancement via reanalysis information is a valuable line of investigation.

Bayesian Neural Fields offered an insight into the future of spatial weather modeling—moving beyond point forecasts into fields of predictions with quantified uncertainties. While not yet surpassing deterministic methods in raw accuracy, these probabilistic models open opportunities for risk-based decision-making. Stakeholders could, for example, assess not just the most likely forecast scenario but also the confidence or likelihood of extreme conditions.

Further Improvements

Future work could explore several enhancements:

- **Expanded Feature Sets:** Incorporating soil moisture indices, vegetation indices (from satellite imagery), or topographical data might refine local predictions further.
- **Longer Lead Times and Multi-Horizon Forecasts:** Extending beyond 24-hour forecasts to multiple days while maintaining accuracy and employing uncertainty quantification could greatly expand the usefulness for farm planning.
-

- **Real-time ERA5-Land Proxies:** Investigating alternate reanalysis products or improved nowcasting methodologies could yield better stand-ins for ERA5-Land and reduce dependency on latent data.
- **Refinement of Bayesian Neural Fields:** Further research into hyperparameter tuning, prior selection, and optimization strategies may uncover the full potential of Bayesian Neural Fields, making them competitive in both accuracy and robustness.
- **Integration with Operational Decision Systems:** Embedding the improved forecasts into decision-support tools, mobile applications, or farm management software could facilitate immediate practical application, ensuring that the project's benefits are realized at scale.

In conclusion, this project demonstrated that combining local weather data, global forecasts, reanalysis-inspired features, and advanced ML or probabilistic models can significantly improve local weather predictions. The improvements support more informed agricultural decisions, ultimately strengthening resilience and promoting sustainable practices. The exploration of Bayesian Neural Fields stands as an innovative step towards more sophisticated, uncertainty-aware local weather forecasting, with the potential to further refine and enhance future predictive frameworks.