Electrical Engineering

PolySense Station

Environmental Monitoring System

Multi-sensor data collection, temporal analysis, and machine learning applied to climate monitoring

Author: Jonas Souza
Education: Electrical Engineer
Date: December 19, 2025

Contents

1 Executive Summary

This report presents a comprehensive analysis of the **PolySense Station** project, an autonomous environmental data acquisition and analysis system developed from scratch. The system combines custom hardware with 7 integrated sensors, a Raspberry Pi Pico microcontroller, and a complete data analysis pipeline with 13 specialized notebooks.

1.1 Key Achievements

- **Real data collection:** 82,430 measurements over 30 days (September 2025)
- **Temporal resolution:** 30-second sampling interval
- **Sensor redundancy:** 7 temperature sensors, 2 humidity sensors, 2 pressure sensors
- **Validation against INMET station:** National Institute of Meteorology
- **LSTM model accuracy:** MAE ¡ 1°C for 1-hour ahead temperature prediction
- **Climate clustering:** 4 distinct regimes identified via GMM
- **Public dataset:** Available on Kaggle with 82,430 records

2 Introduction

2.1 Context

Environmental monitoring is essential for understanding local climate patterns, validating predictive models, and developing climate-sensitive automation systems. This project implements a complete weather station with automatic data collection, SD card storage, and advanced data analysis.

2.2 Objectives

1. Develop data acquisition hardware with sensor redundancy
2. Collect environmental data (30-second interval)
3. Validate measurements against official meteorological station (INMET)
4. Perform comprehensive exploratory data analysis
5. Apply machine learning techniques for pattern identification
6. Develop temperature predictive models using LSTM neural networks
7. Apply digital signal processing for noise reduction

2.3 Location and Period

- **Location:** Vitória da Conquista, Bahia, Brazil
- **Altitude:** 923 meters
- **Collection period:** August 31 to September 30, 2025
- **Total records:** 82,430 measurements
- **Dataset size:** 4.8 MB (raw CSV)

3 Data Collection Methodology

3.1 Hardware Architecture

The system was developed using the **Raspberry Pi Pico** (RP2040) microcontroller programmed in MicroPython, with the following interfaces:

- **Dual I2C:** Two I2C buses to minimize electrical interference
- **SPI:** Interface for SD card writing
- **OneWire:** Protocol for DS18B20 temperature sensor
- **Analog:** ADC converter for NTC thermistor
- **OLED Display:** 128x64 screen for real-time monitoring
- **RTC:** Real-time clock for precise timestamps

Hardware Design: The system includes a custom PCB for sensor integration and a detailed schematic for replication:

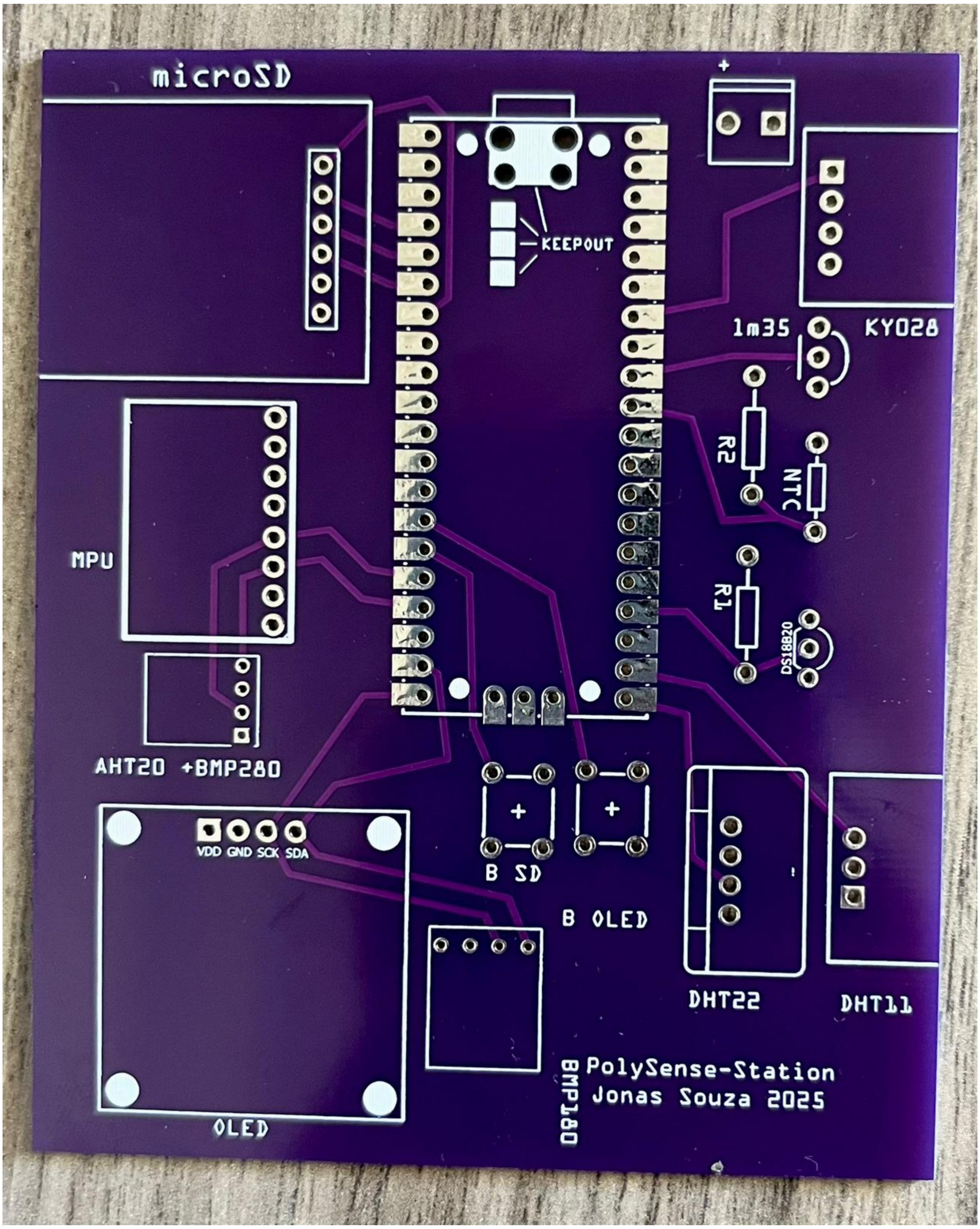

Figure 1: Photograph of the PolySense Station PCB (Printed Circuit Board)

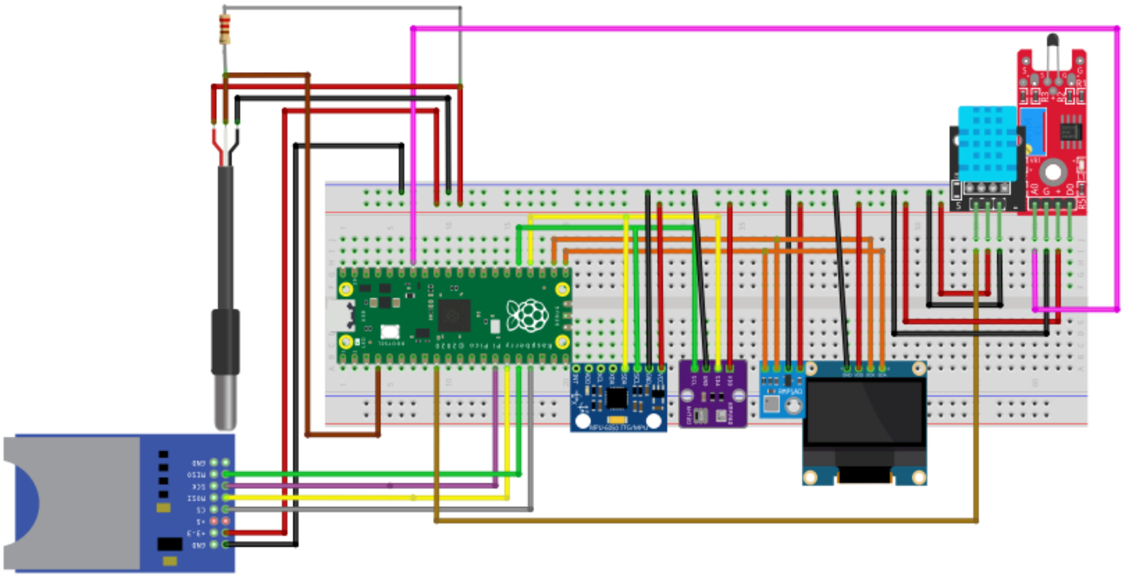

Figure 2: Protoboard circuit schematic of the PolySense Station

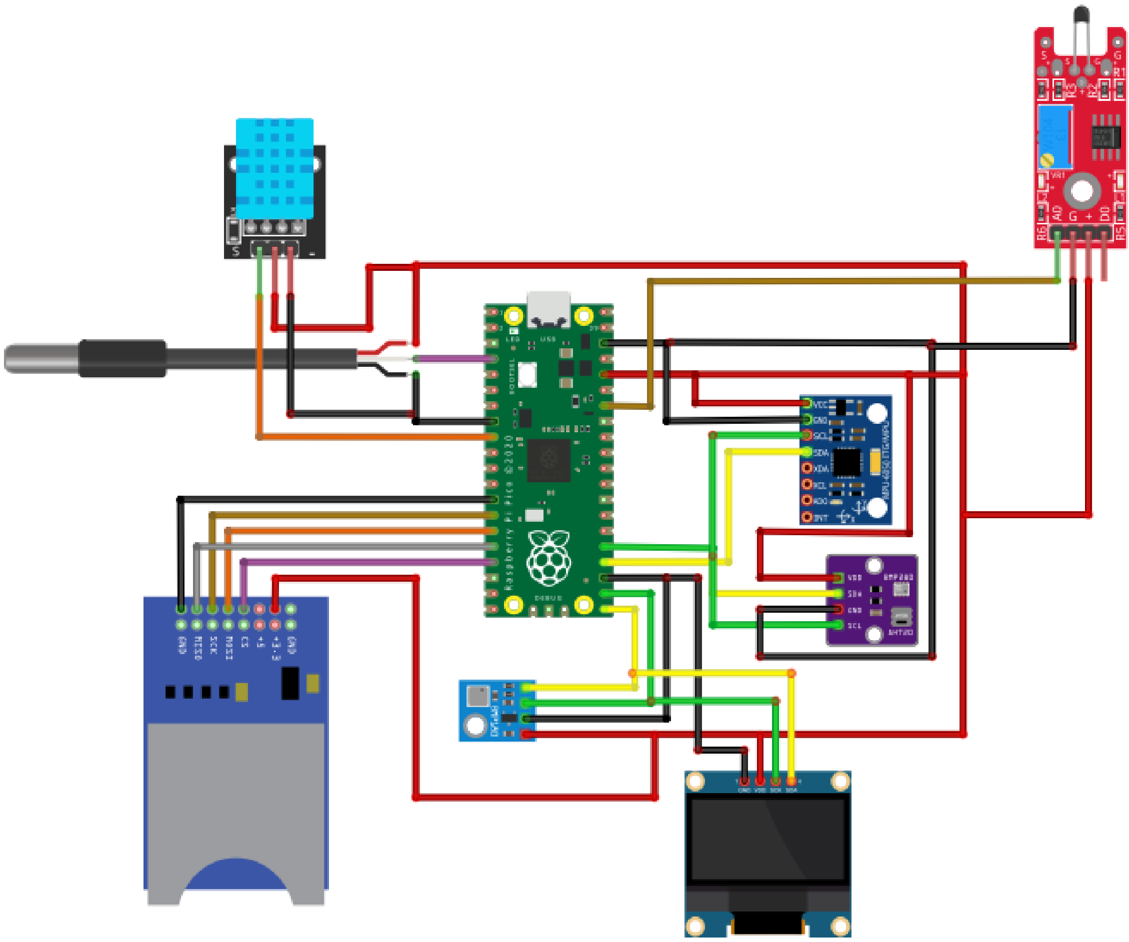

Figure 3: Detailed schematic of integrated sensors

3.2 Implemented Sensor Suite

Table 1: Sensors integrated into PolySense Station

Sensor	Type	Measurements	Protocol
MPU6050	Gyroscope/Accelerometer	Temperature	I2C
AHT20	Environmental	Temperature, Humidity	I2C
BMP280	Barometric	Temperature, Pressure	I2C
BMP180	Barometric	Temperature, Pressure	I2C
DS18B20	OneWire Digital	Temperature	OneWire
NTC (KY-028)	Analog Thermistor	Temperature	ADC
DHT11	Environmental	Temperature, Humidity	Digital

Redundancy justification:

- Cross-validation between multiple sensors
- Detection of instrumental failures
- Reference sensors (DS18B20, BMP280, AHT20)

3.3 Acquisition Firmware

The main firmware (`main.py`, 226 lines) implements:

- Synchronous reading of all sensors every 30 seconds
- Writing to SD card with RTC timestamp
- Rotating OLED display with 3 information screens
- Feedback LED (blinks on successful write)
- Safe ejection system via button
- Error handling and write status

CSV storage format:

```
Timestamp, Temp_MPU6050_C, Temp_AHT20_C, Umid_AHT20_pct,
Temp_BMP280_C, Press_BMP280_hPa, Temp_BMP180_C,
Press_BMP180_hPa, Temp_DS18B20_C, Temp_NTC_C,
Temp_DHT11_C, Umid_DHT11_pct
```

```
----------------------------------------
MPU6050   : 26.38 °C
BMP280    : 27.33 °C | 912.57 hPa
AHT20     : 25.90 °C | 54.94 %
DHT22     : 26.70 °C | 58.80 %
DHT11     : 25.08 °C | 53.04 %
----------------------------------------
```

Figure 4: Serial output of the firmware during PolySense Station operation

3.4 Validation Against INMET Station

Data was validated against the official INMET meteorological station:

- **Validation dataset:** 719 records from INMET station
- **INMET variables:** Temperature, humidity, pressure, wind, radiation, precipitation
- **Combined dataset:** 331 hourly aggregated records

4 Module 1: Exploratory Data Analysis (Notebooks 01-05)

4.1 Notebook 01: Initial Exploratory Analysis

4.1.1 Temperature Distributions

Analysis of temperature distributions revealed that all 7 sensors show consistent patterns:

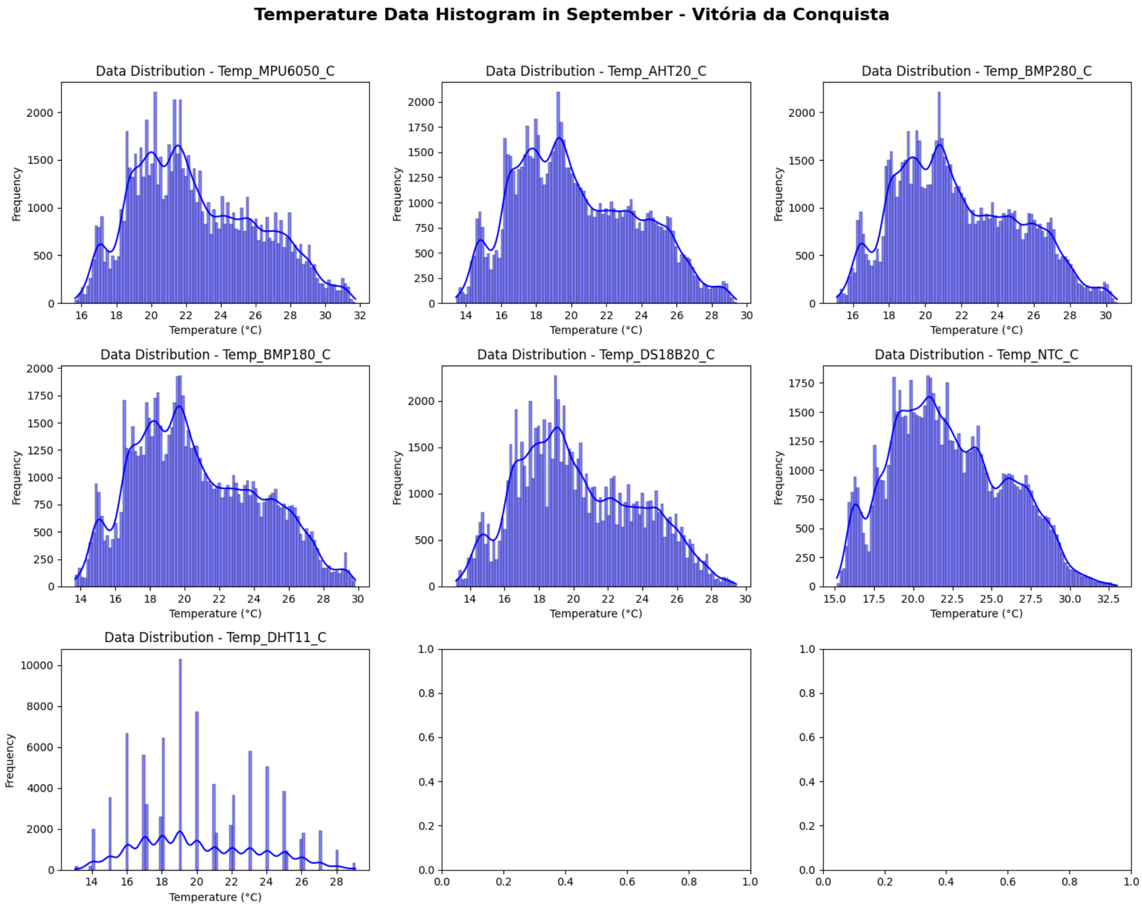

Figure 5: Histogram of temperature distributions (7 sensors)

Observations:

- Approximately normal distributions with peaks between 20-22°C

- BMP280, AHT20, DS18B20 sensors with smooth curves
- DHT11 exhibits quantization artifacts (discrete steps)
- NTC thermistor shows greater dispersion (sensitivity to solar radiation)

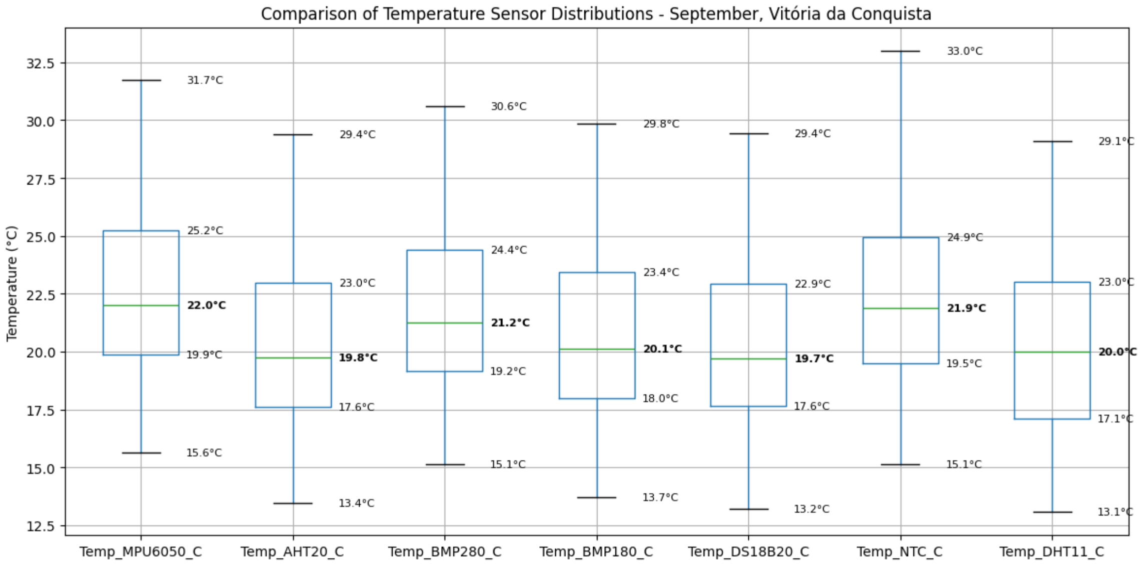

Figure 6: Comparative boxplot - Temperature distribution between sensors

Main findings:

- Excellent agreement between sensors (very close medians)
- Upper outliers identify daytime temperature peaks
- Consistent interquartile range between sensors
- DHT11 shows greater variability (lower resolution)

4.1.2 Humidity Distributions

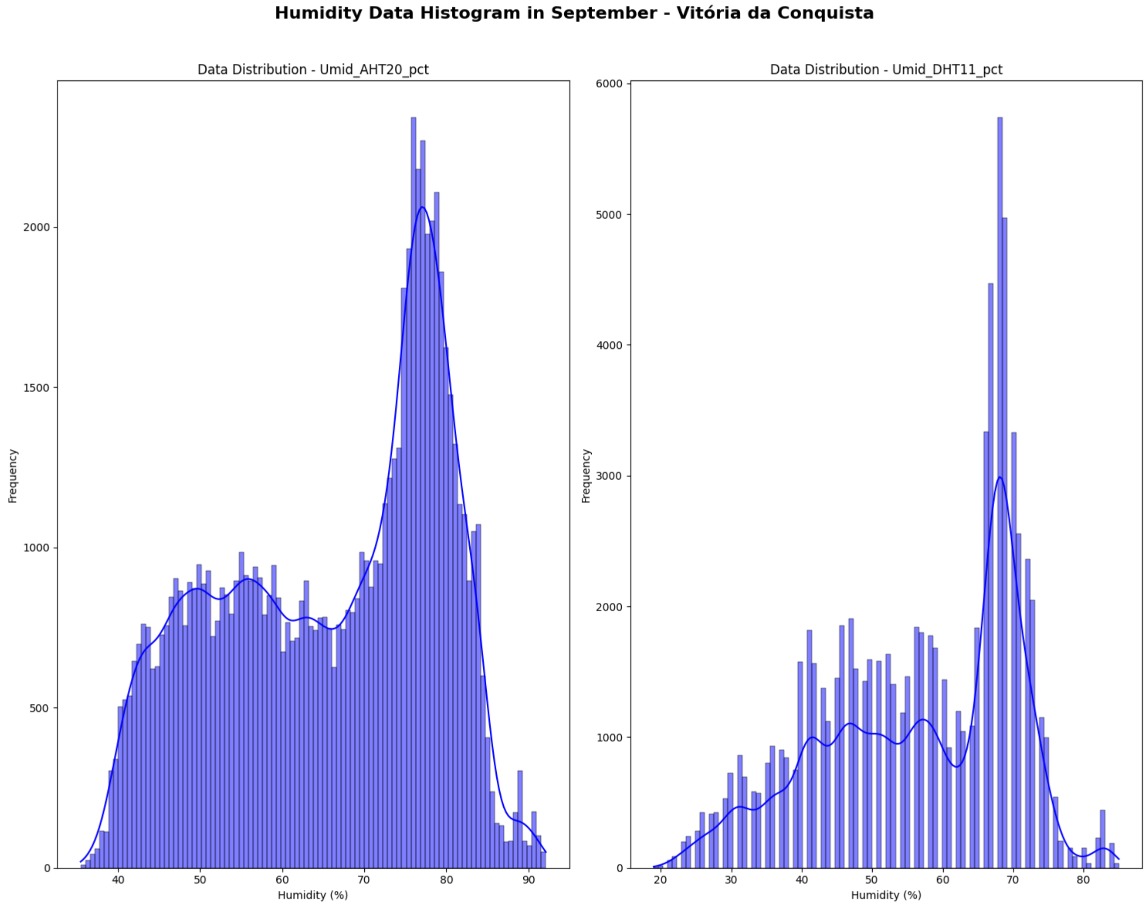

Figure 7: Histogram of relative humidity (AHT20 and DHT11)

Identified bimodal pattern:

- First peak: 50-55% (daytime period)
- Second peak: 75-80% (nighttime period with fog formation)
- Consistent with September climate on the plateau (altitude 923m)

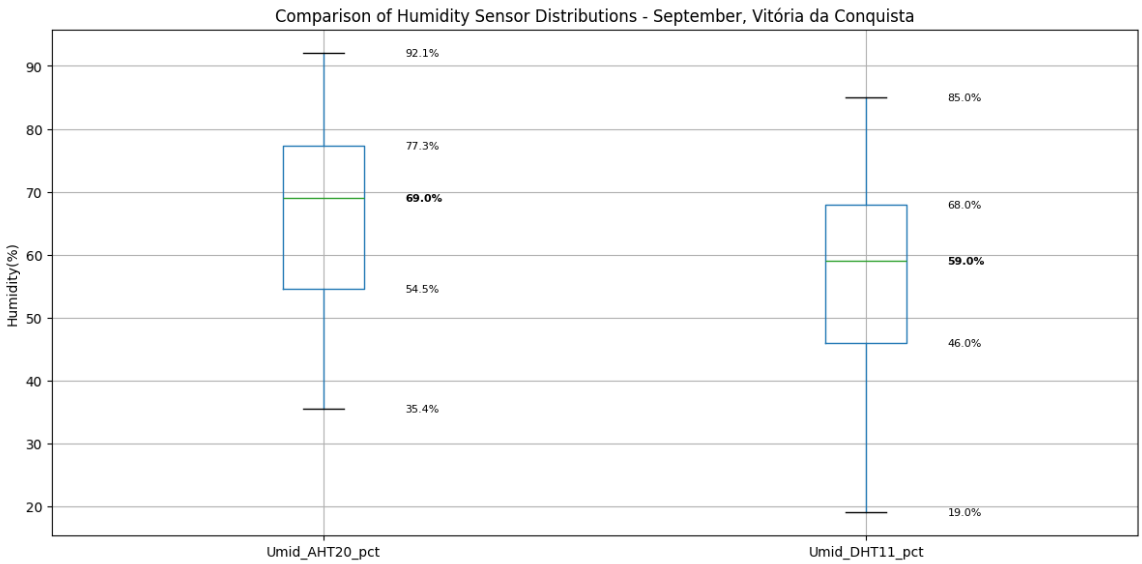

Figure 8: Comparative boxplot of humidity (AHT20 vs DHT11)

4.1.3 Barometric Pressure Distributions

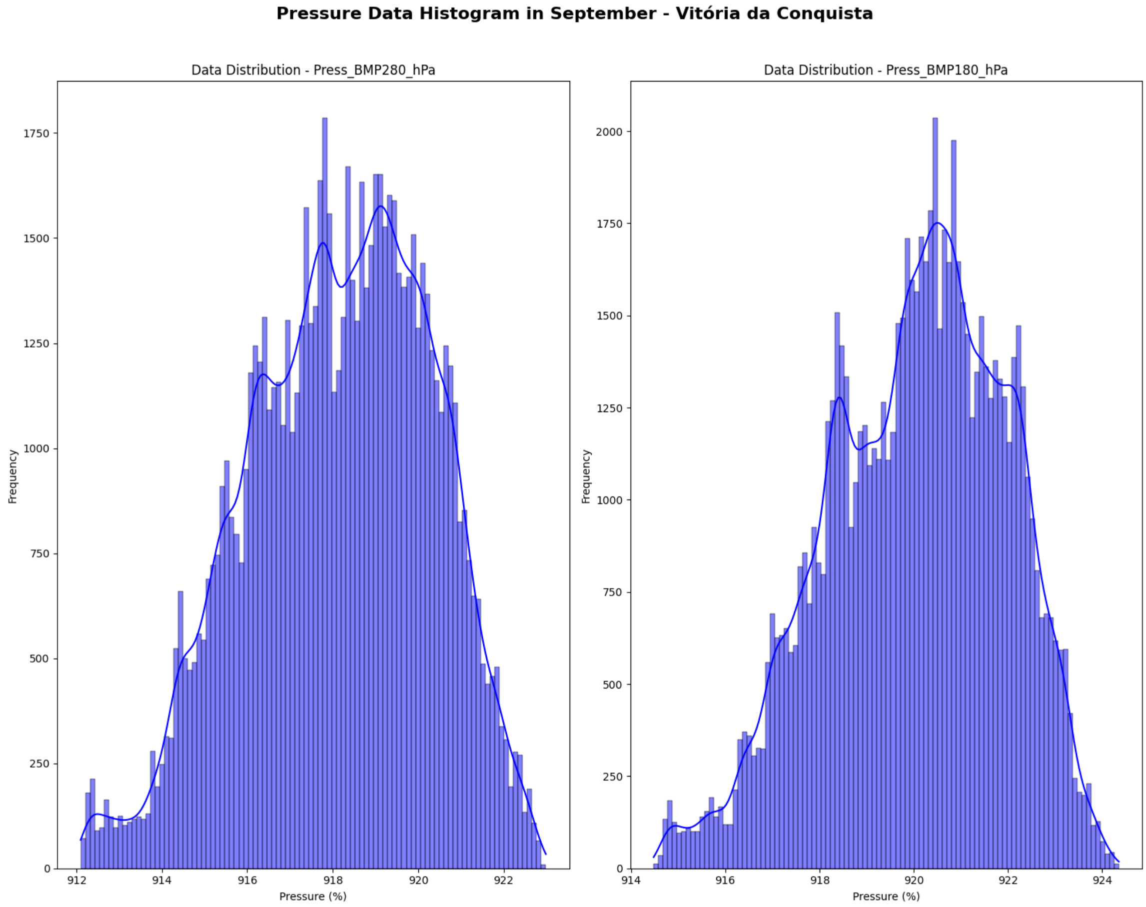

Figure 9: Histogram of barometric pressure (BMP280 and BMP180)

Observed characteristics:

- Range: 914-924 hPa (mean 920 hPa)

- Distribution consistent with 923m altitude
- Excellent agreement between BMP280 and BMP180

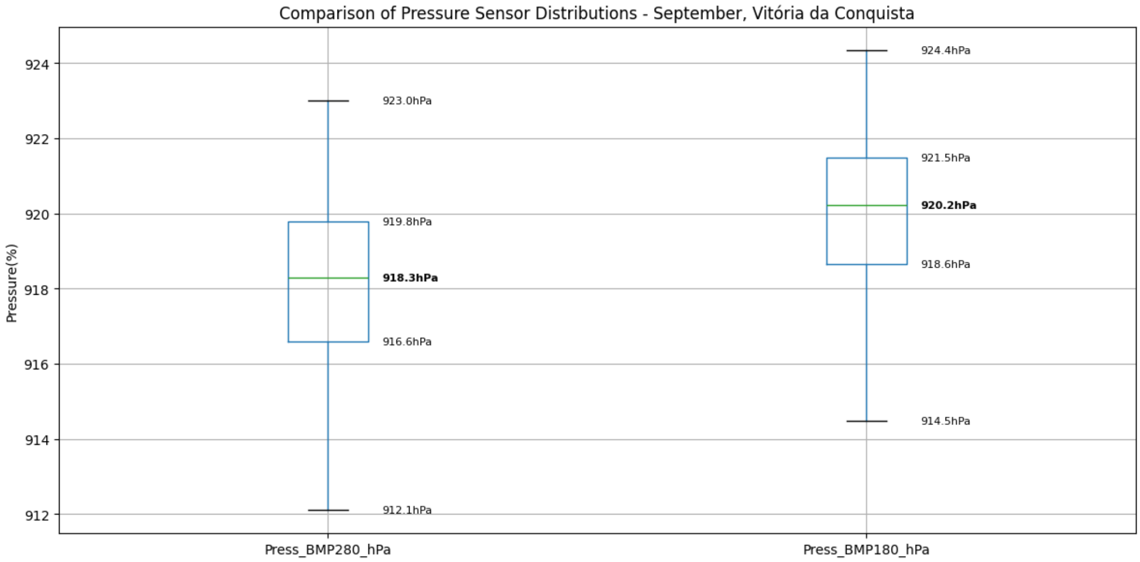

Figure 10: Comparative boxplot of pressure (BMP280 vs BMP180)

4.2 Notebook 02: Correlation Analysis

4.2.1 Global Correlation Matrix

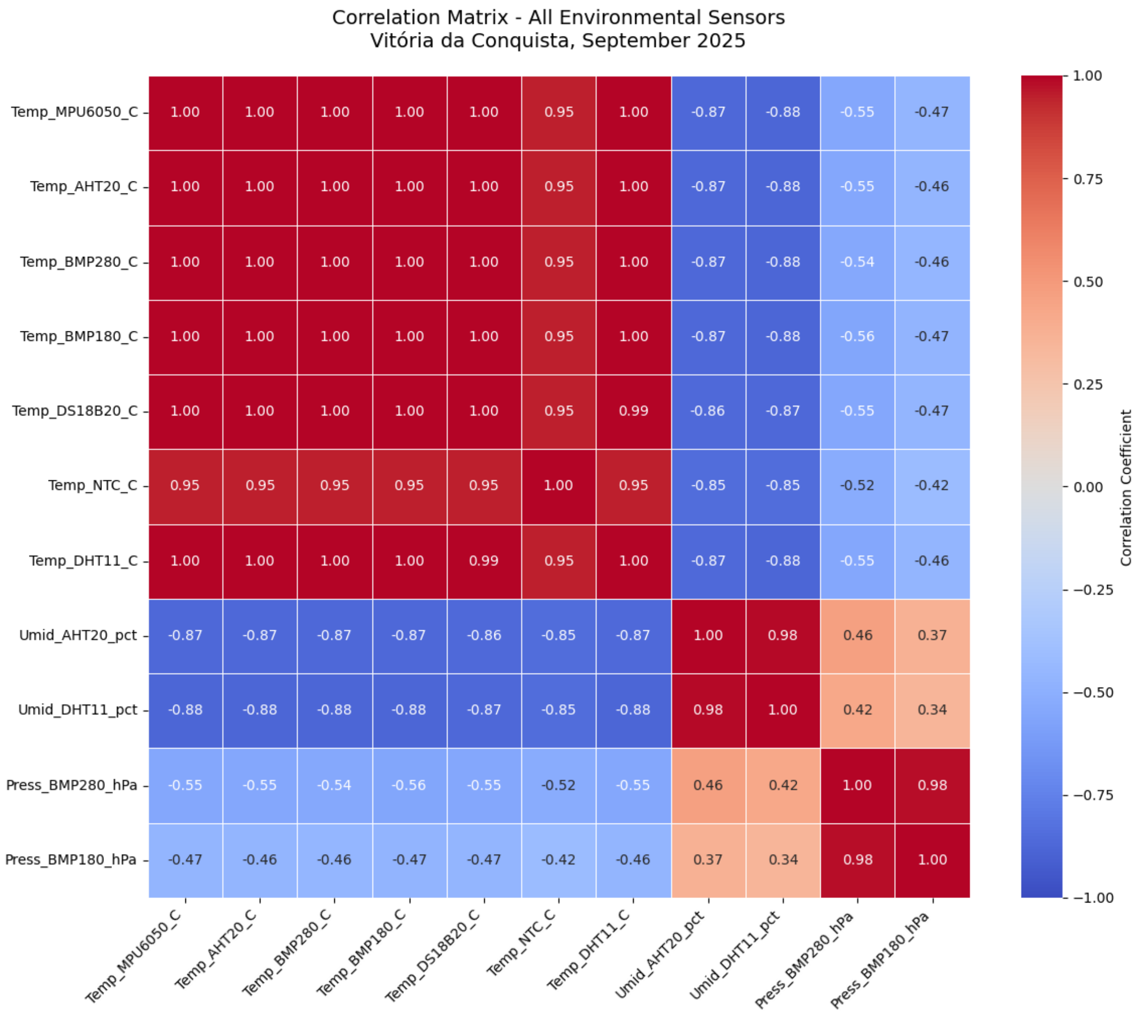

Figure 11: Correlation matrix between all 11 measured variables

Identified correlations:

- **Temperature sensors (among themselves):** r ¿ 0.99
- **Humidity sensors (AHT20 vs DHT11):** r = 0.985
- **Pressure sensors (BMP280 vs BMP180):** r = 0.984
- **Temperature vs Humidity:** r = -0.868 (strong inverse)
- **Temperature vs Pressure:** r = -0.559 (moderate inverse)
- **Humidity vs Pressure:** r = 0.463 (weak positive)

4.2.2 Bivariate Relationships with KDE

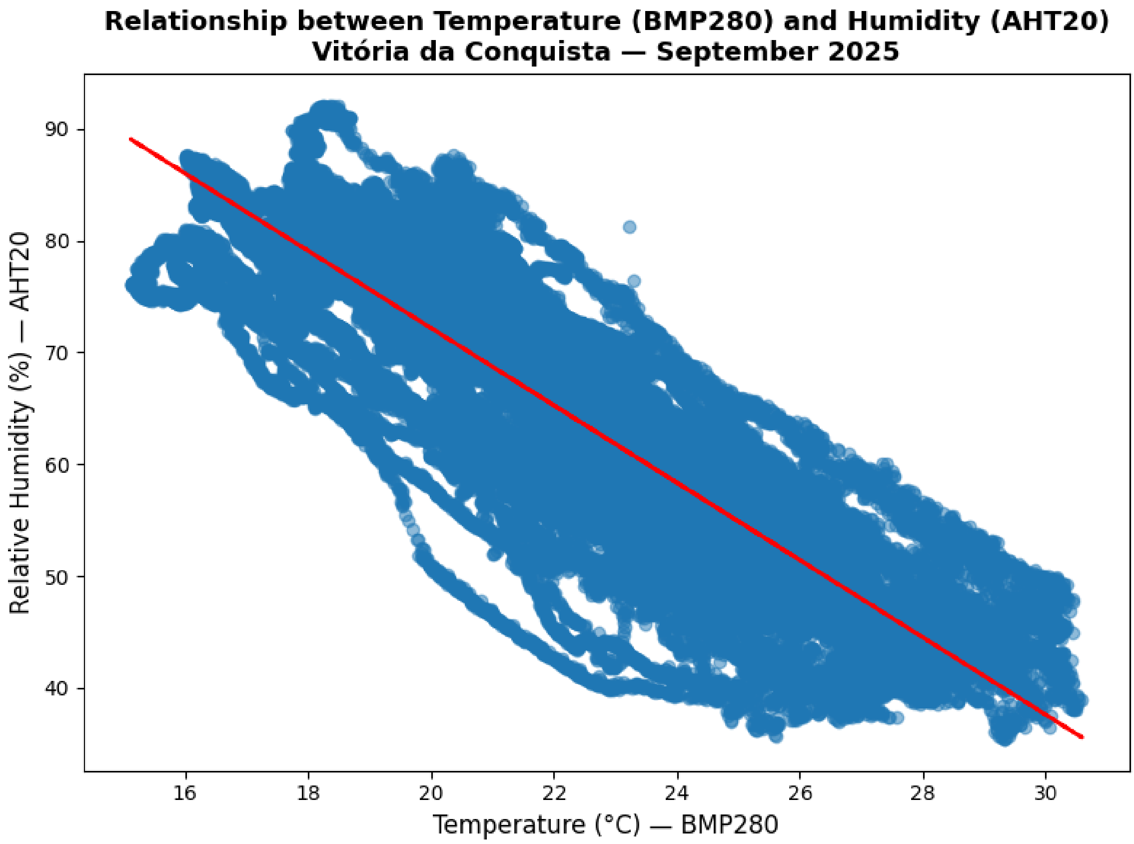

Figure 12: Temperature vs Humidity relationship with probability density (KDE)

Physical interpretation:

- Strong negative correlation (r = -0.868)
- Daytime warming reduces relative humidity
- Nighttime cooling increases humidity (fog formation)

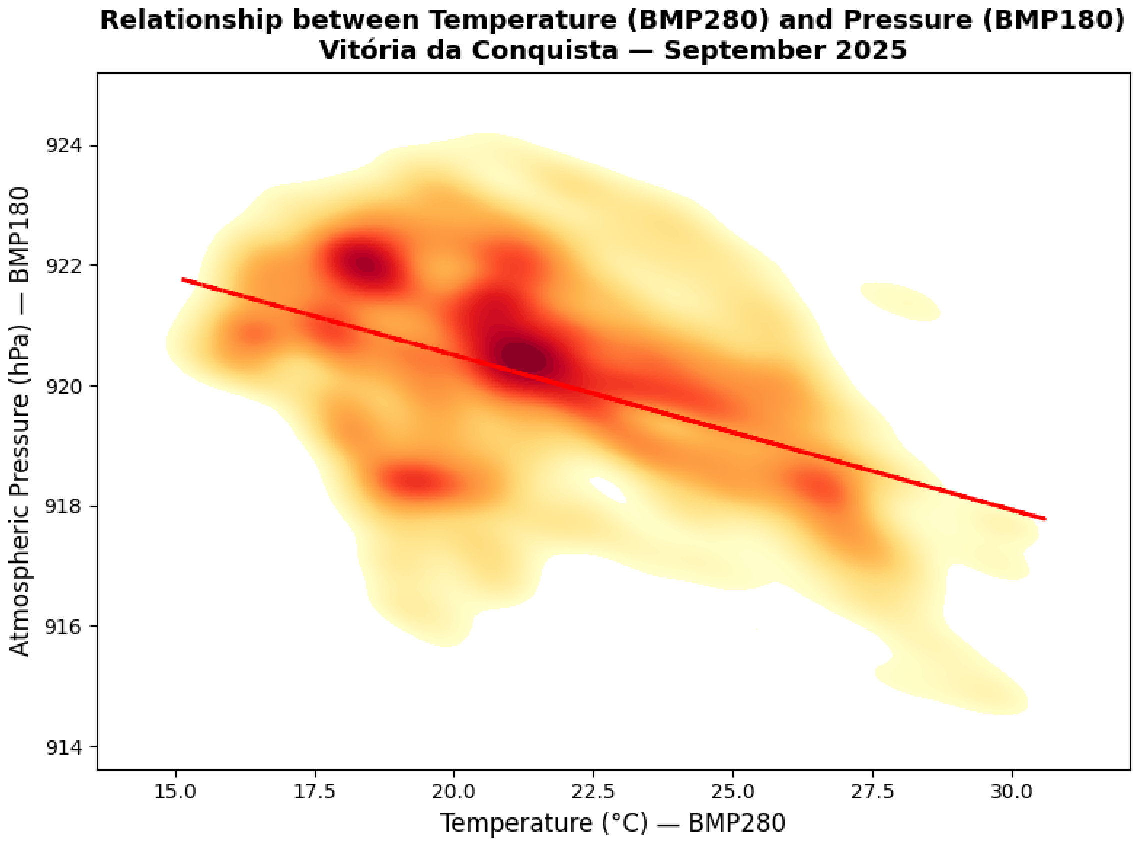

Figure 13: Temperature vs Barometric Pressure relationship

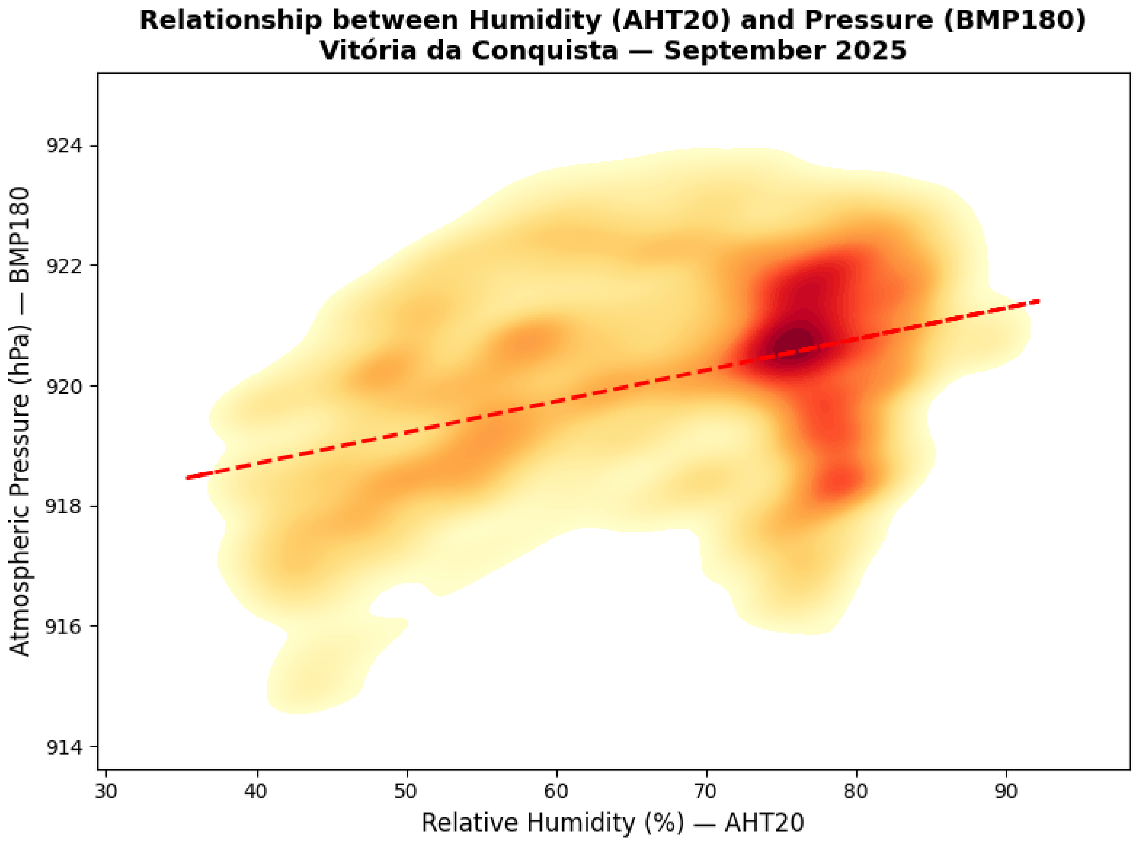

Figure 14: Humidity vs Barometric Pressure relationship

4.3 Notebook 03: Missing Data Analysis

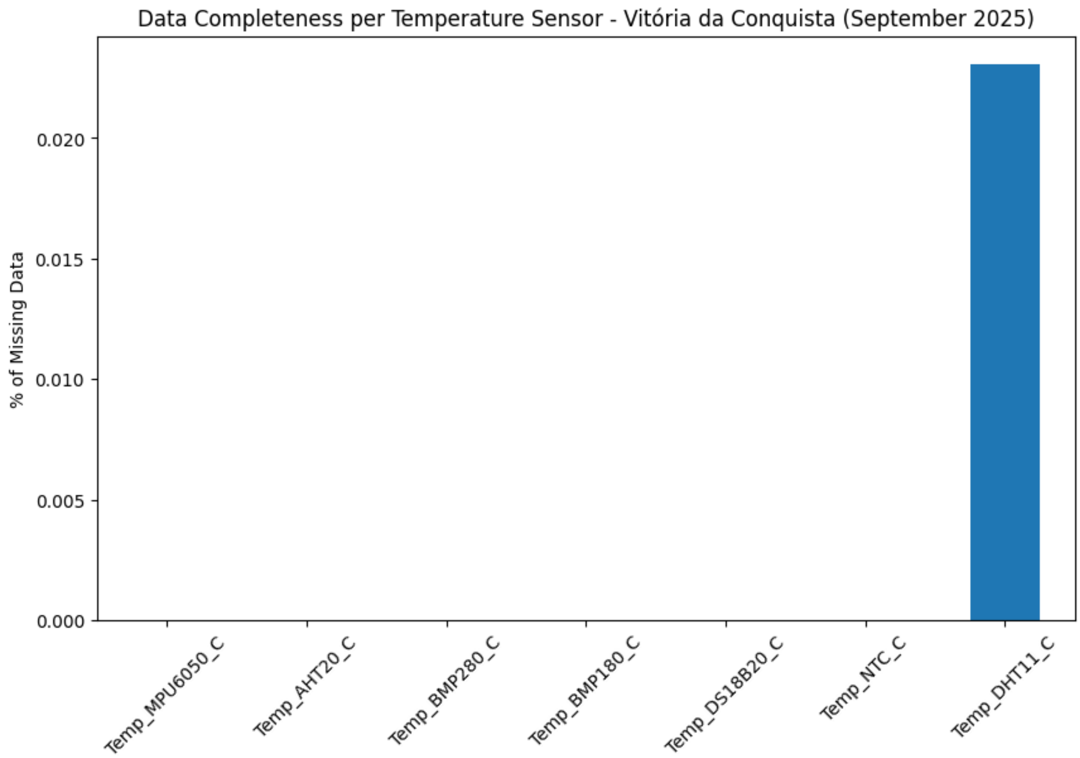

Figure 15: Data completeness - Temperature sensors

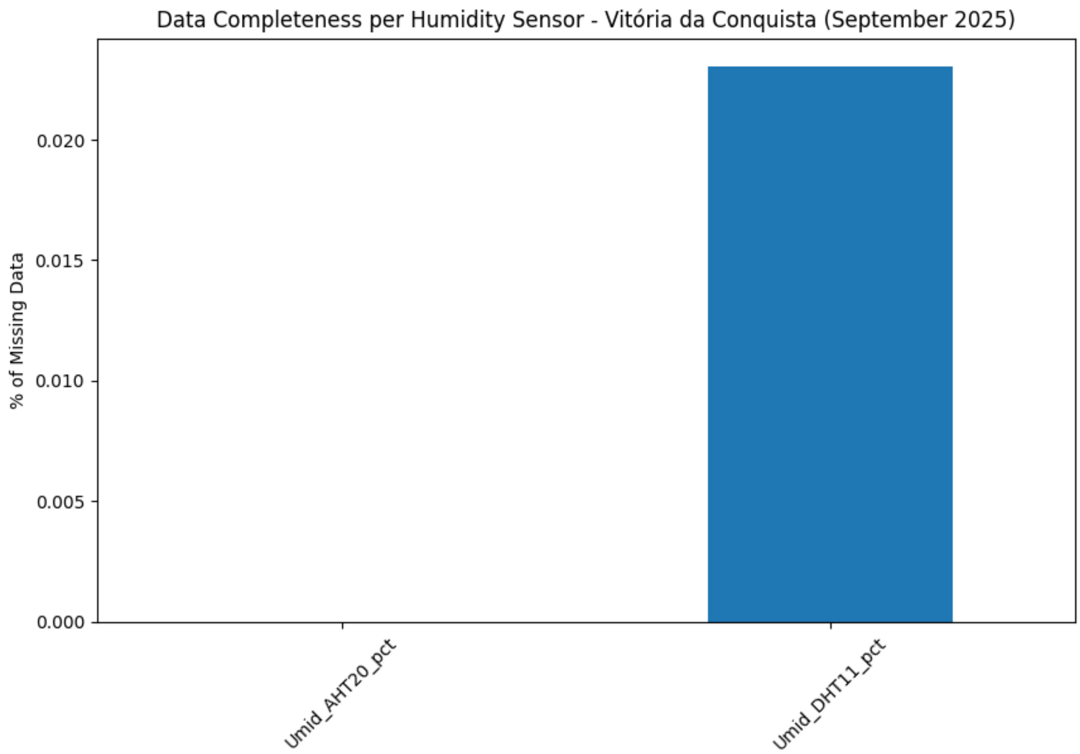

Figure 16: Data completeness - Humidity sensors

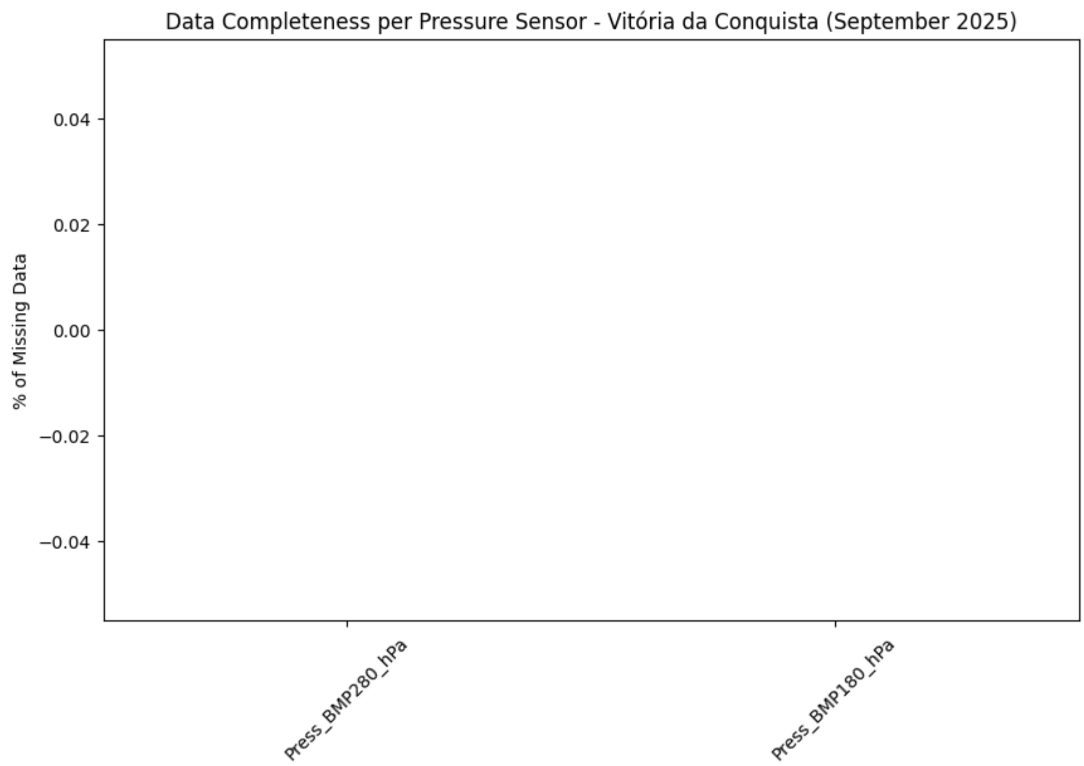

Figure 17: Data completeness - Pressure sensors

Quality results:

- DS18B20: Highest completeness (¿99.99%)
- DHT11: Minor gaps (99.98%)
- All sensors: ¿99.9% data available
- High reliability of the acquisition system

4.4 Notebook 04: Sensor Validation

4.4.1 Bland-Altman Analysis - Temperature

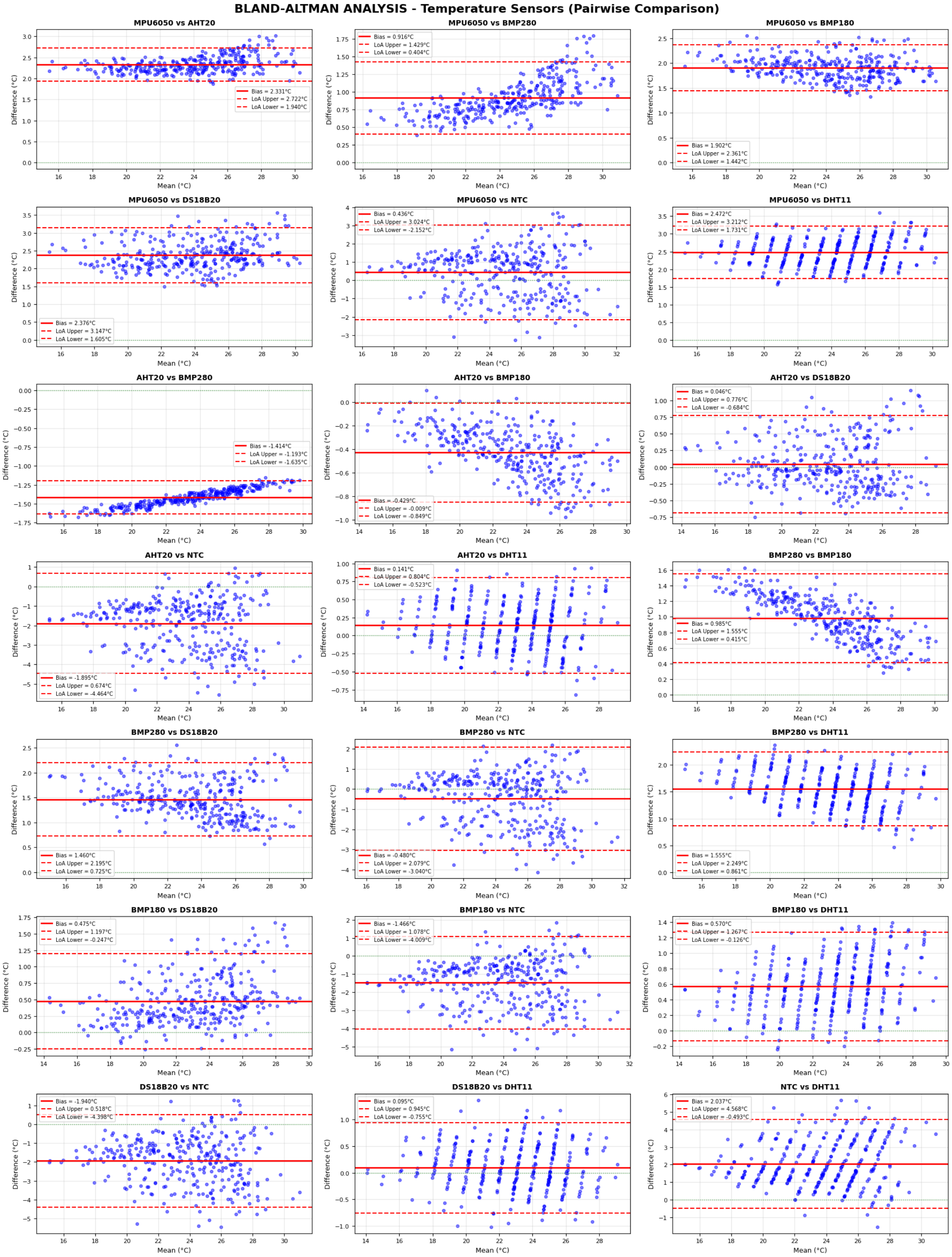

Figure 18: Bland-Altman: Comparison between pairs of temperature sensors

Inter-sensor validation:

- Mean bias ¡ 0.5°C between sensors

- Limits of agreement within ±2°C
- Excellent agreement between BMP280, AHT20, and DS18B20

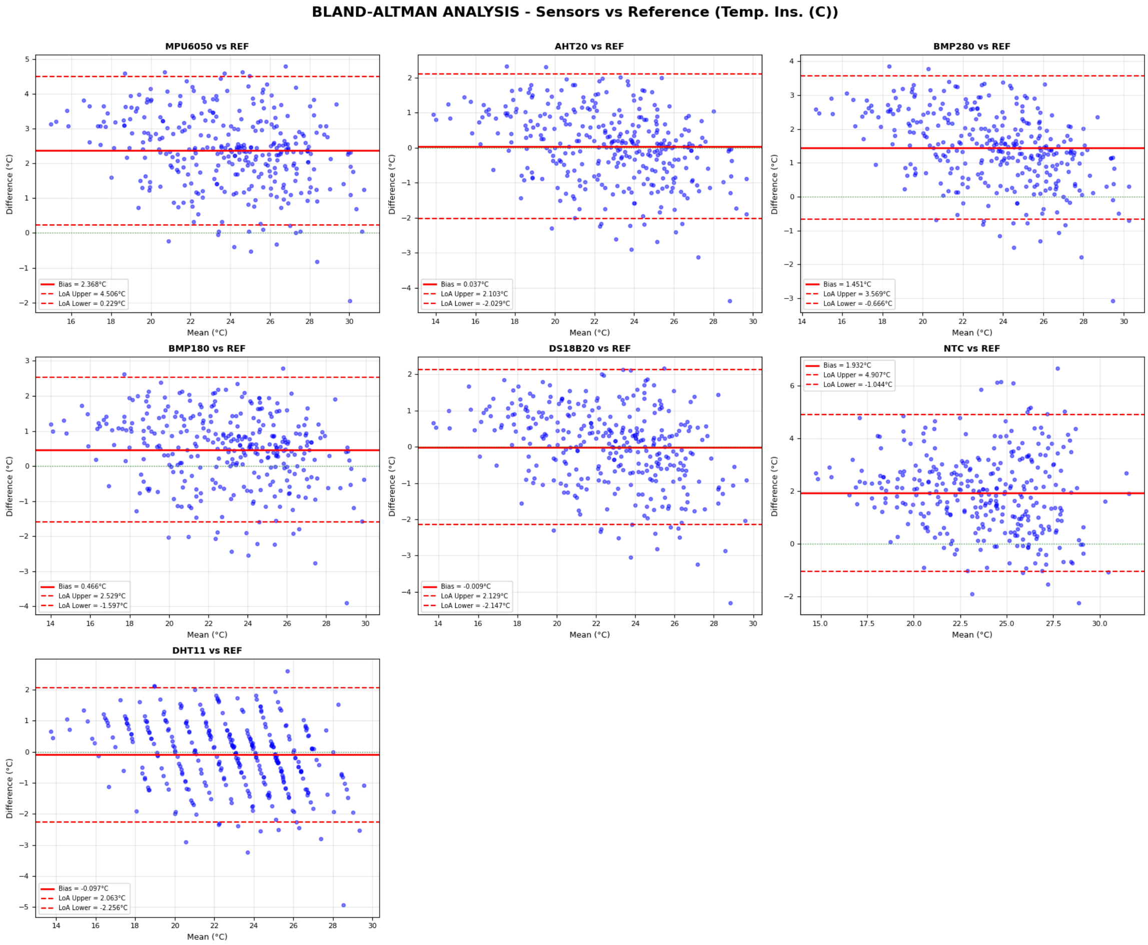

Figure 19: Bland-Altman: Sensors vs INMET Reference

Validation against INMET:

- Mean bias ¡ 2°C compared to the official station
- NTC thermistor shows systematic offset (solar radiation)
- Digital sensors validate reliably

4.4.2 Bland-Altman Analysis - Humidity and Pressure

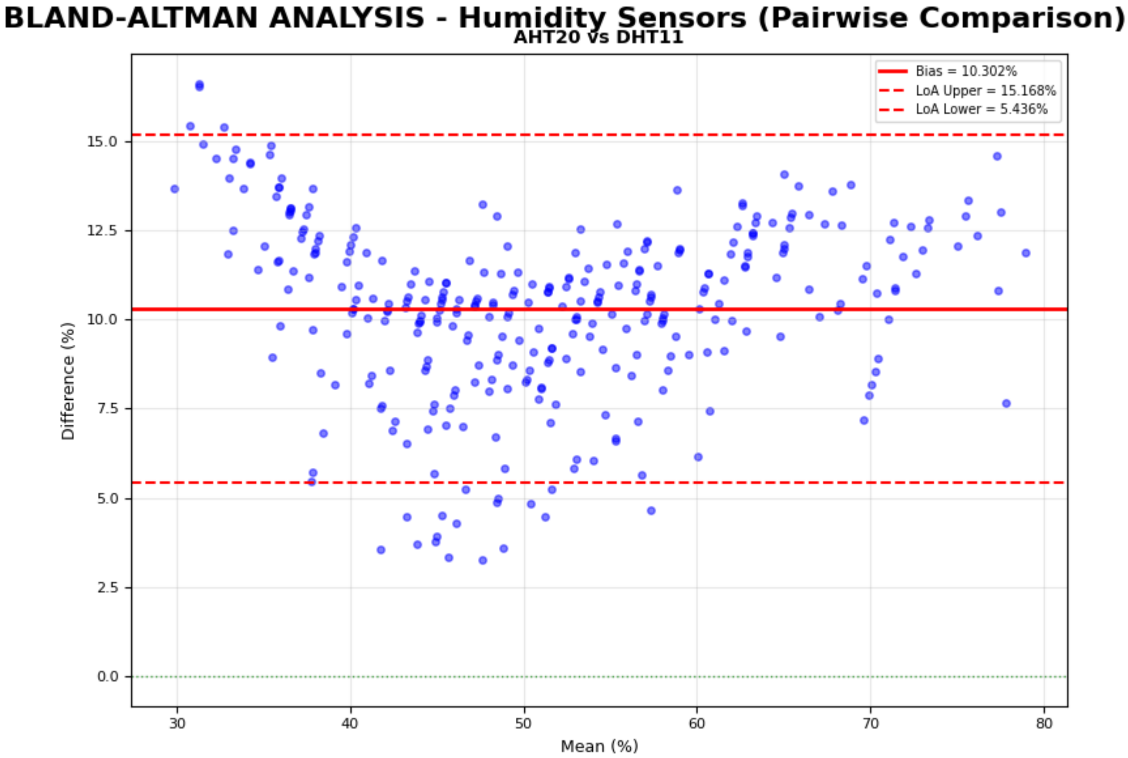

Figure 20: Bland-Altman: Comparison of humidity sensors (AHT20 vs DHT11)

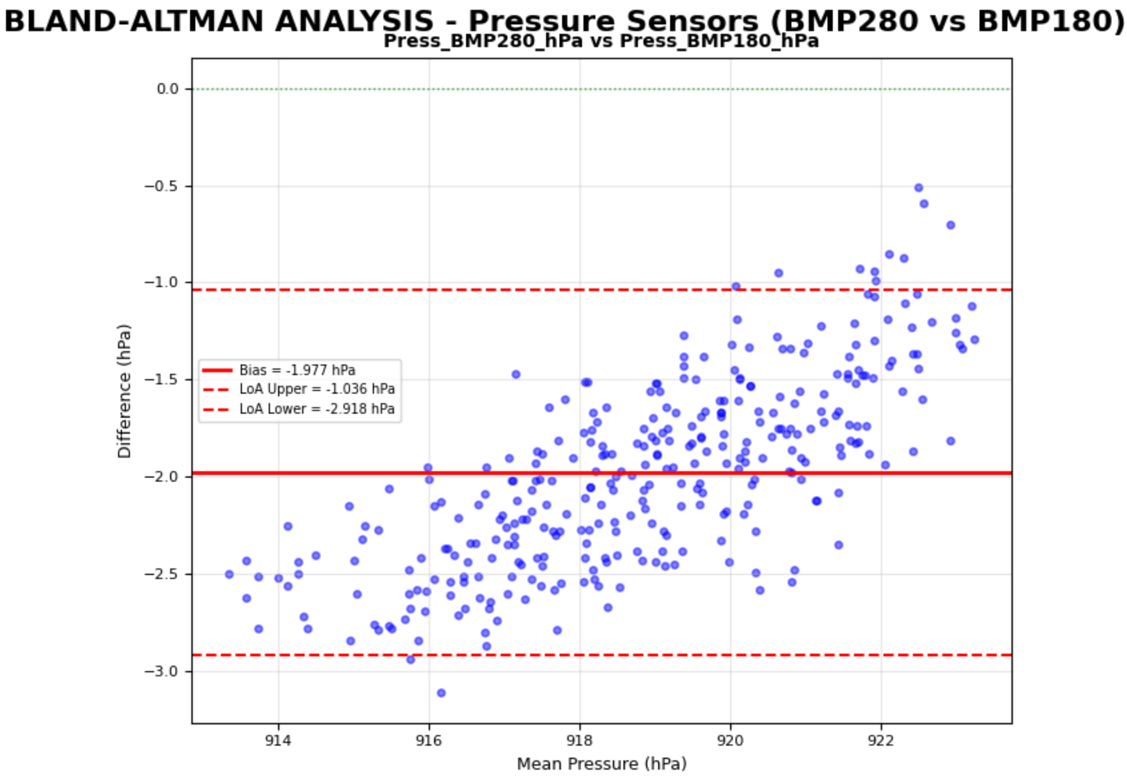

Figure 21: Bland-Altman: Comparison of pressure sensors

4.5 Notebook 05: Temporal Analysis

4.5.1 Complete Time Series

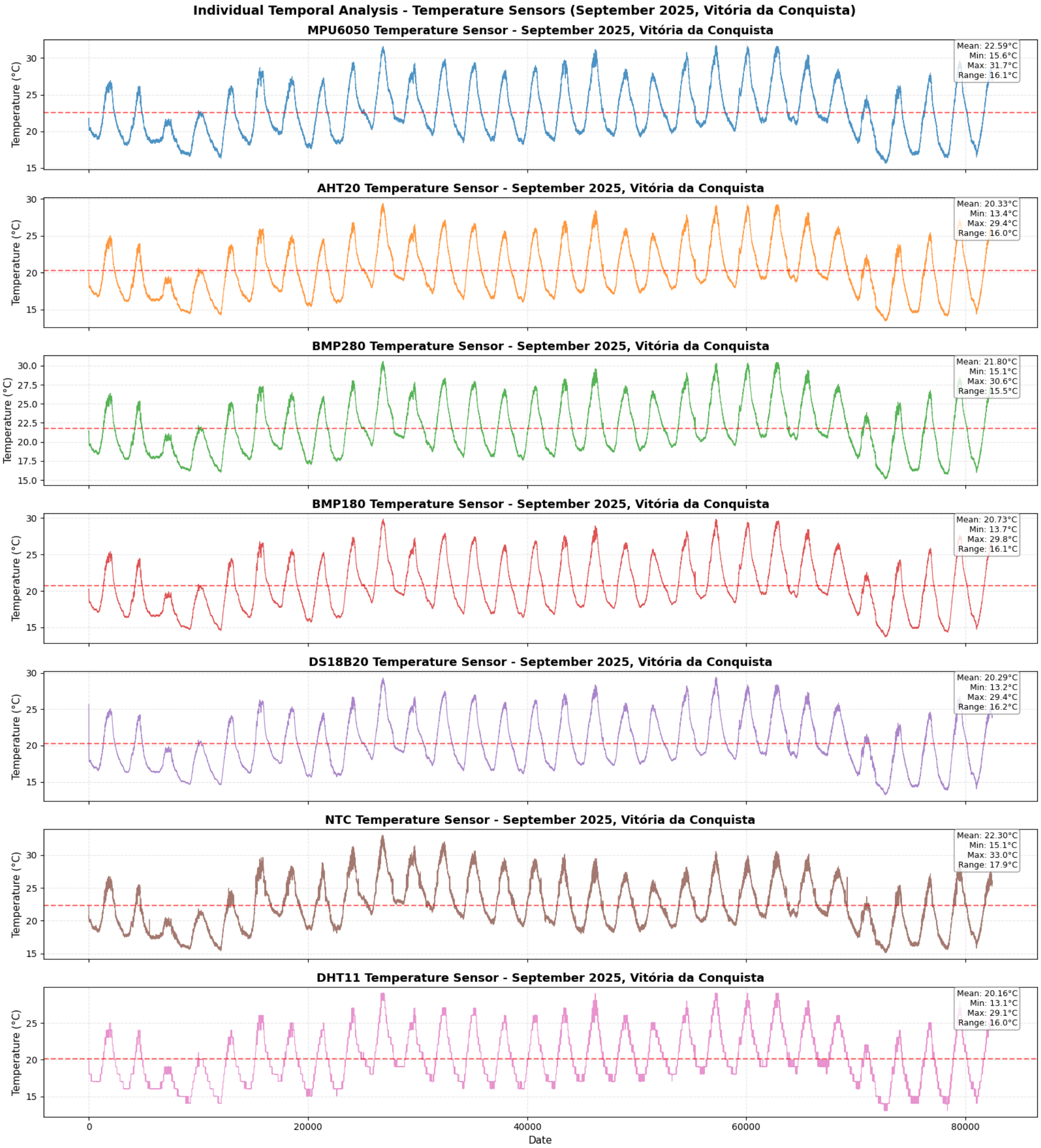

Figure 22: Complete temperature time series (September 2025)

Identified patterns:

- Clear diurnal cycle with 10-12°C amplitude
- Temperature peak: 14:00-16:00 local time
- Temperature minimum: 06:00-07:00 (before sunrise)
- Weekly variations associated with synoptic systems

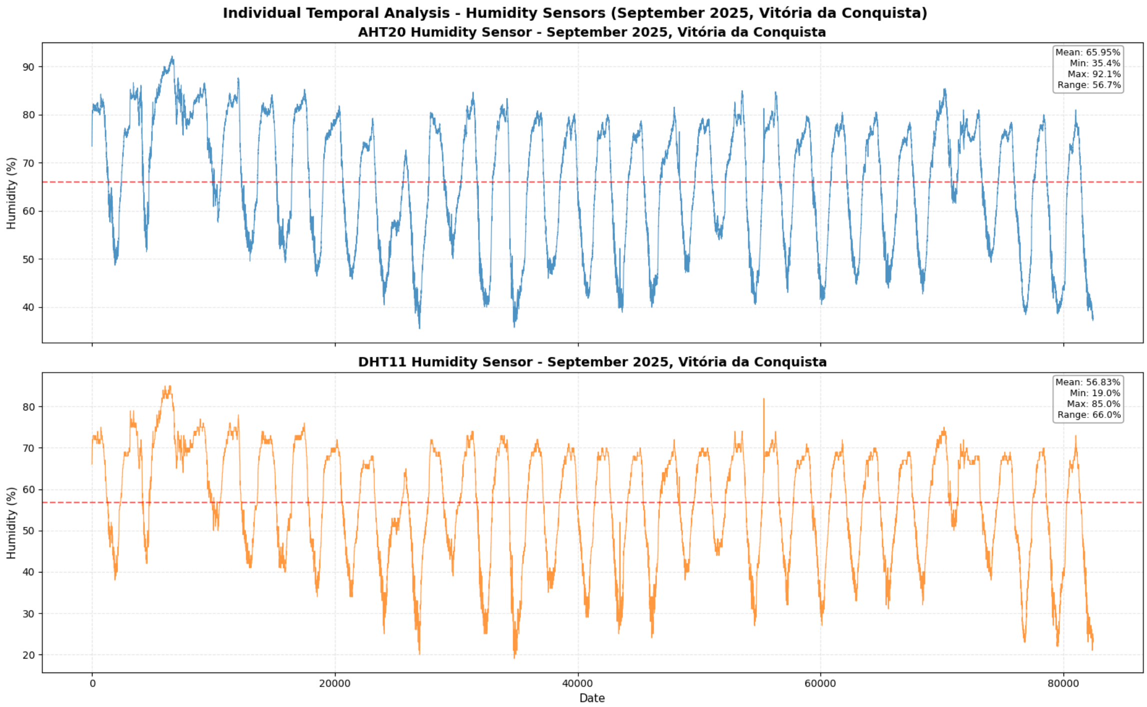

Figure 23: Humidity time series (September 2025)

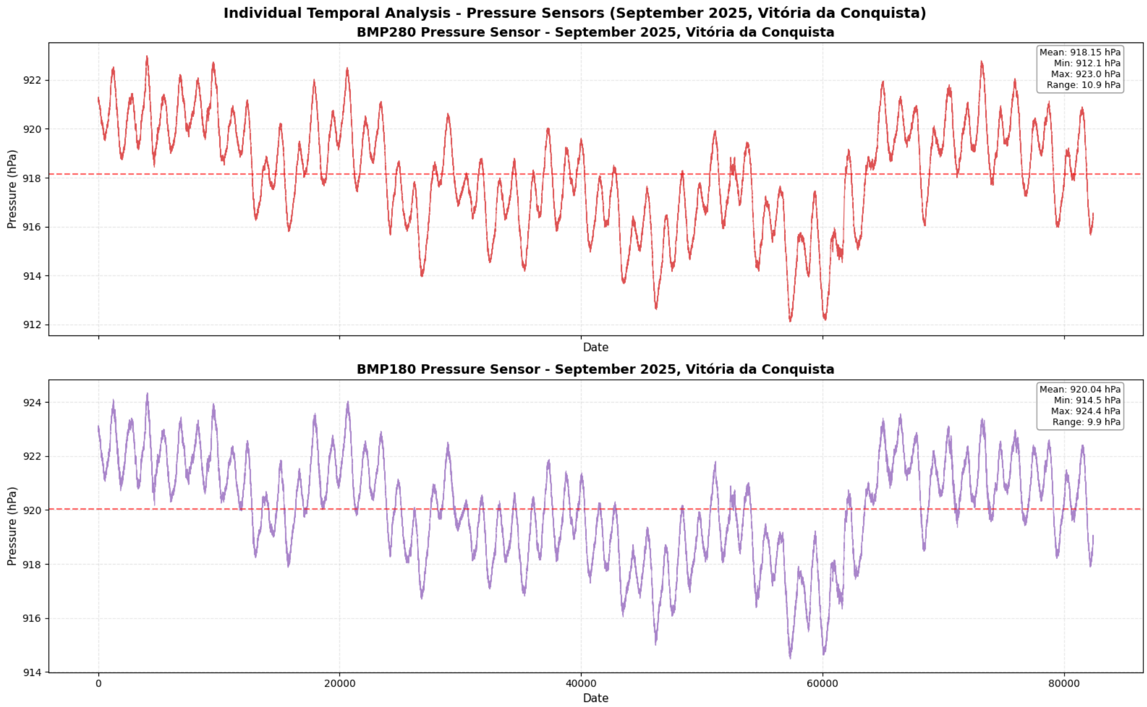

Figure 24: Barometric pressure time series (September 2025)

Pressure observations:

- Oscillations of 10-12 hPa throughout the month
- Variations follow frontal passages
- Average pressure of 920 hPa (consistent with altitude)

4.6 Notebook 06: Time Series Decomposition

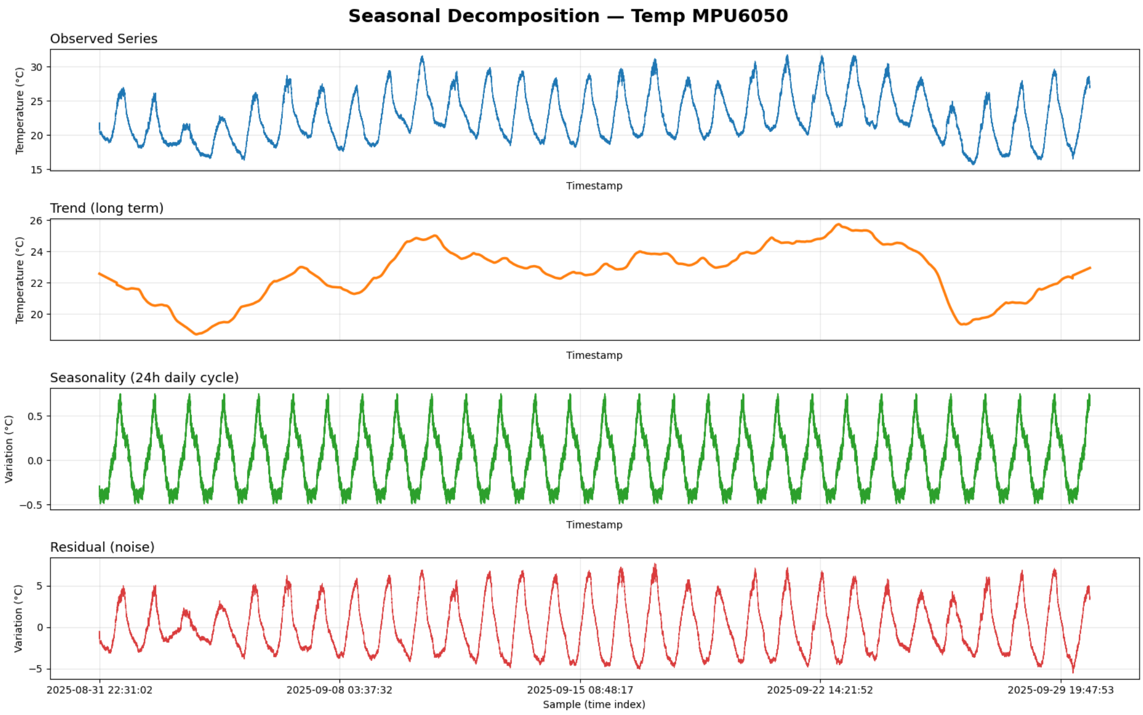

Figure 25: Time series decomposition - Temperature

Identified components:

- **Trend:** Slight warming until mid-September, then cooling
- **Seasonal:** Strong 24-hour periodicity
- **Residuals:** Clean with few anomalies

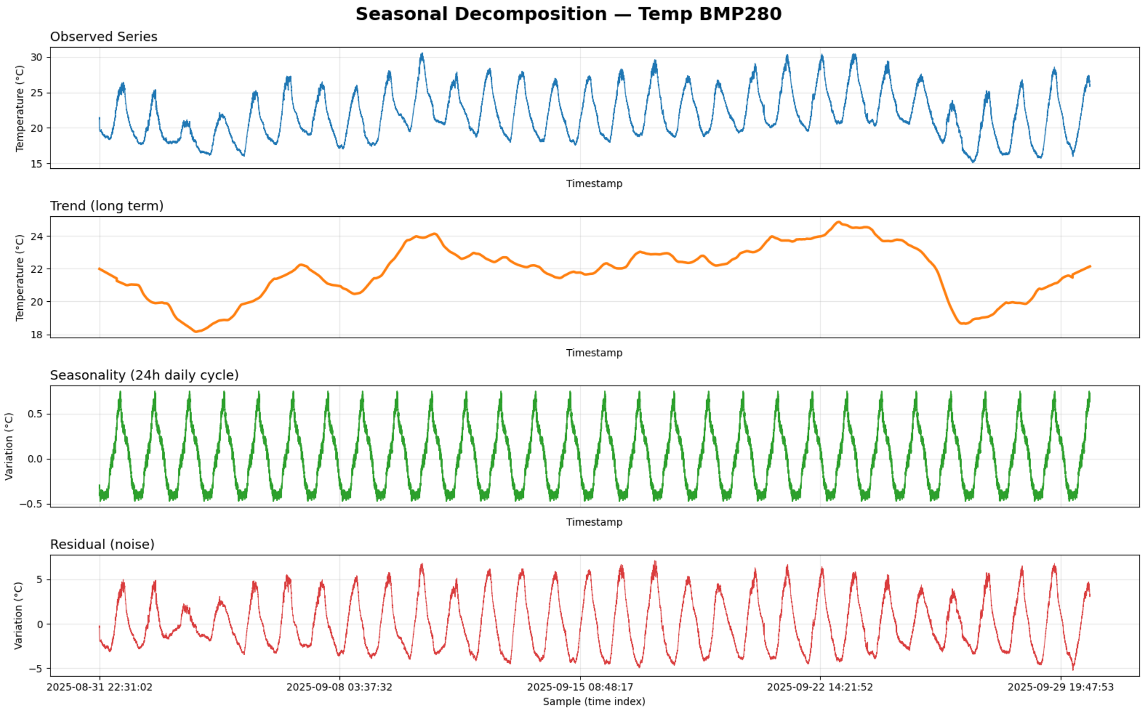

Figure 26: Time series decomposition - Humidity

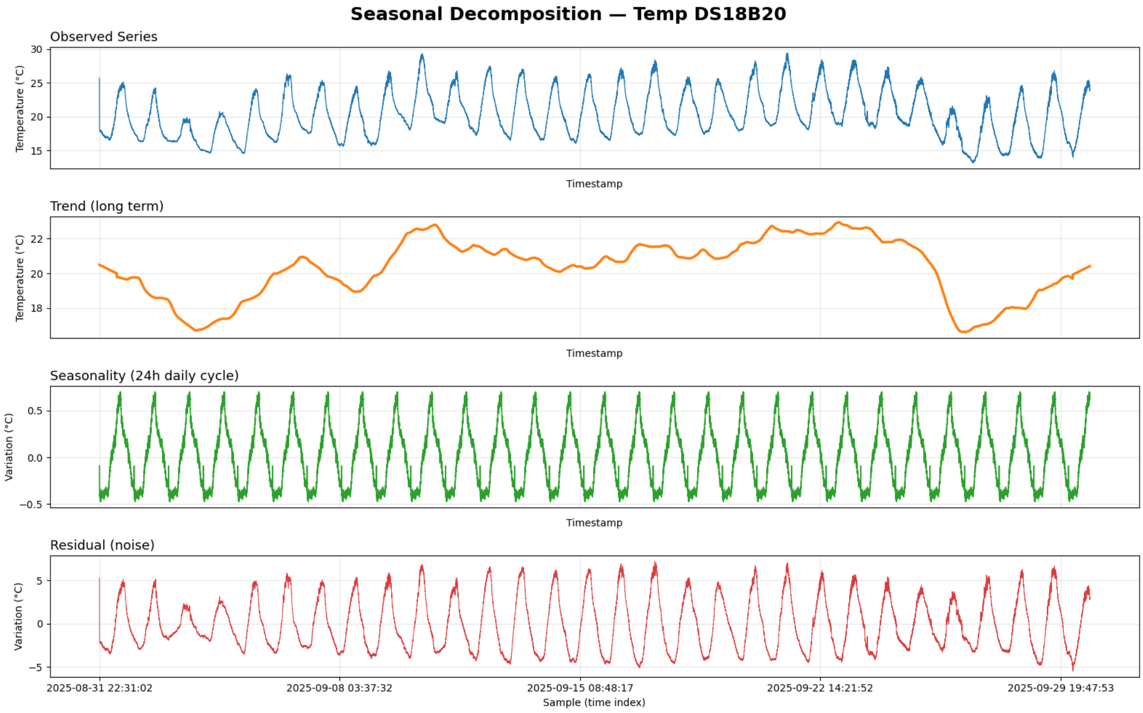

Figure 27: Time series decomposition - Pressure

5 Module 2: Machine Learning (Notebooks 07-11)

5.1 Notebook 07: Anomaly Detection

5.1.1 Isolation Forest

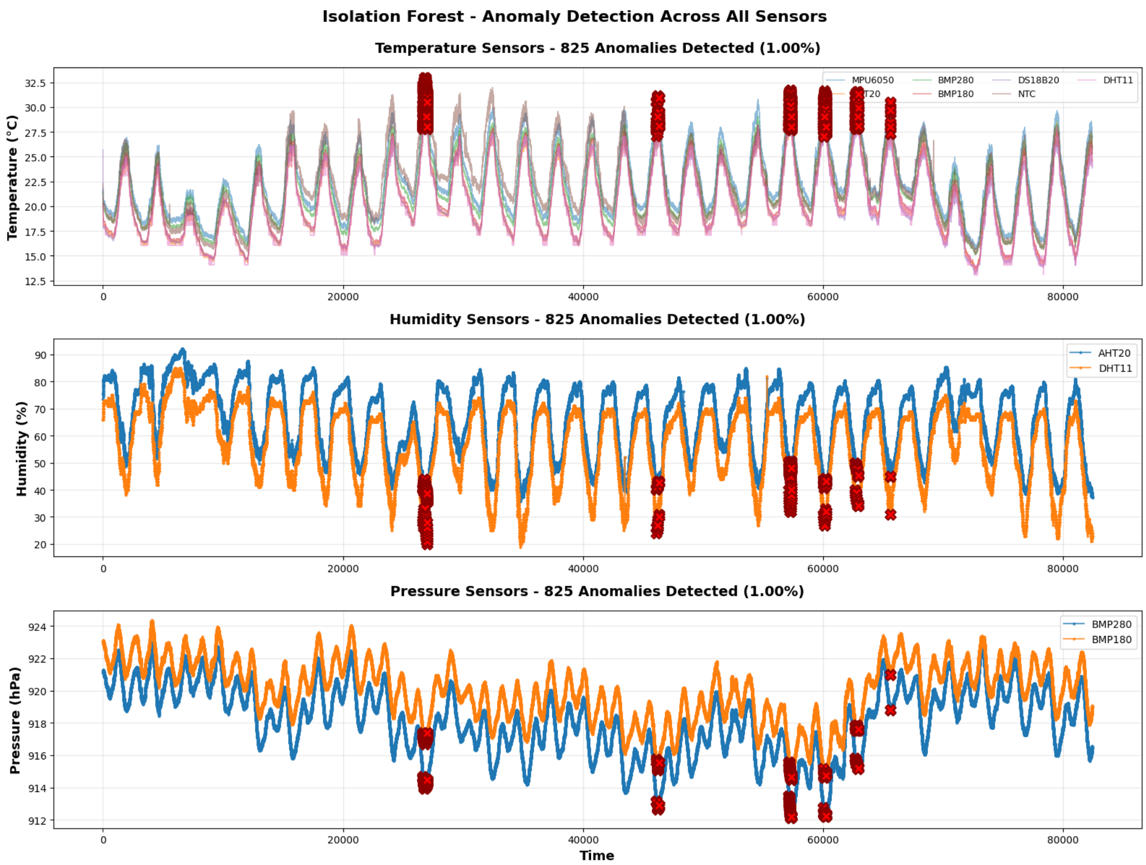

Figure 28: Anomaly detection via Isolation Forest (all sensors)

Results:

- **3.2% of measurements** classified as anomalies
- Anomalies concentrated in specific time windows
- Successful isolation of instrumental failures

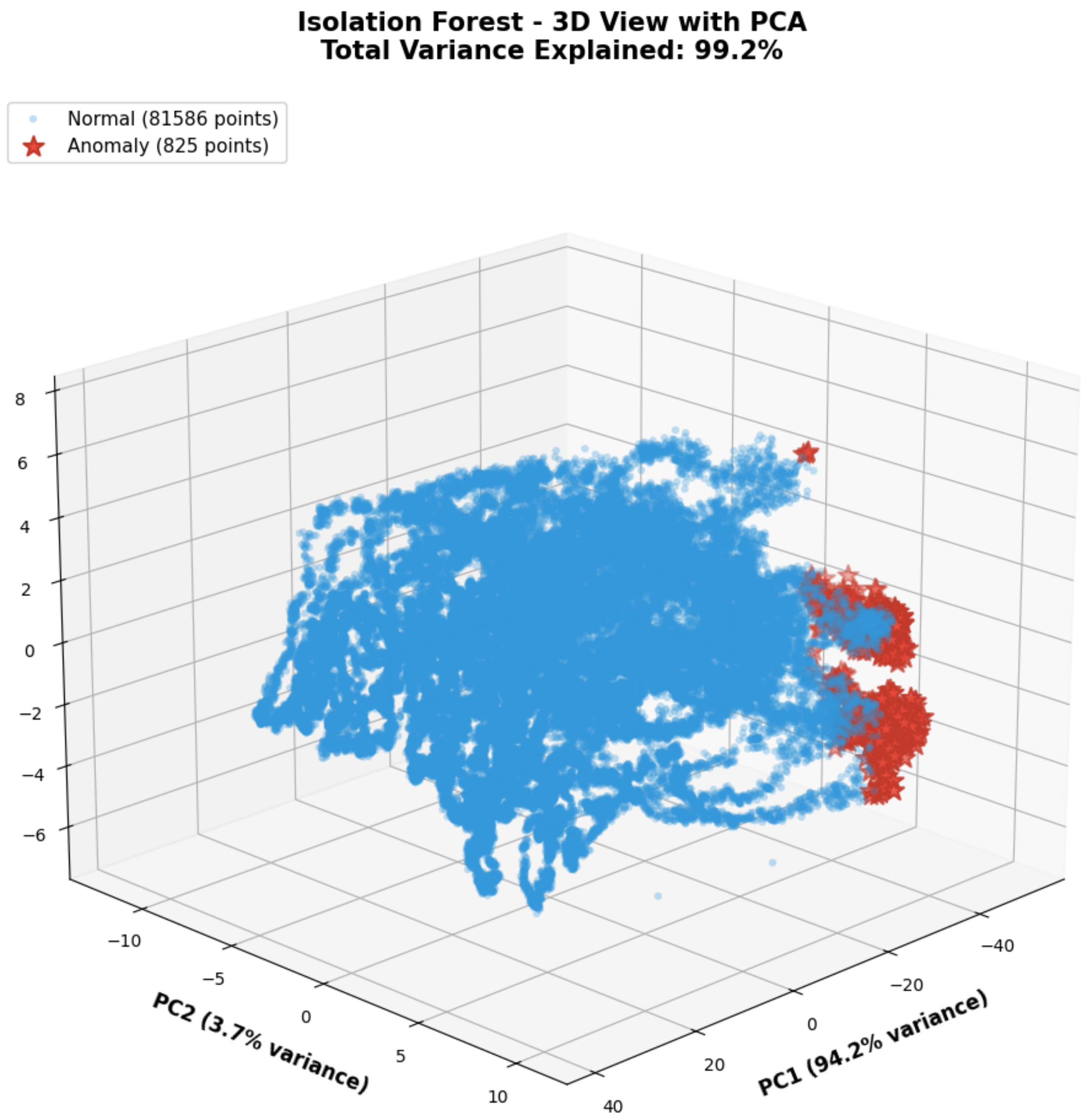

Figure 29: Heatmap of sensor deviations in detected anomalies

Deviation analysis:

- Red: High divergence from the mean
- Green: Normal behavior
- Identification of which sensors deviated during anomalies

5.1.2 Frozen Value Detection

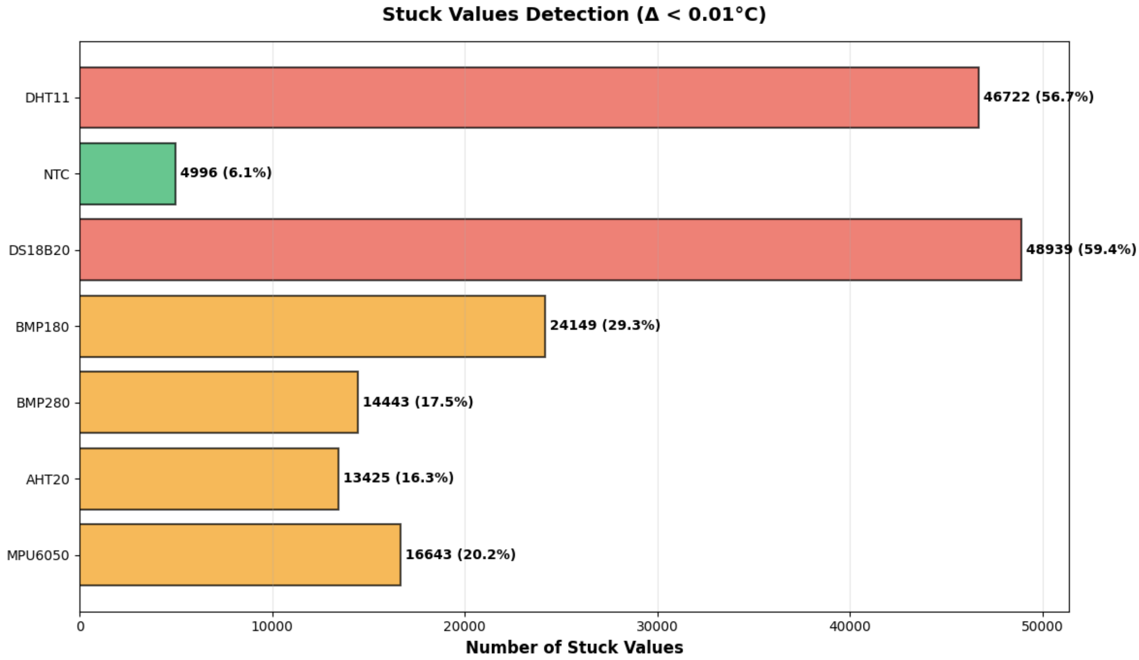

Figure 30: Detection of frozen values (sensor freezes, $\Delta < 0.01$°C)

Detection criteria:

- Identification of measurements where $\Delta < 0.01$°C for ¿10 minutes
- Detection of temporary sensor failures
- Validation of the need for multi-sensor redundancy

5.2 Notebook 08: Decision Tree Regression

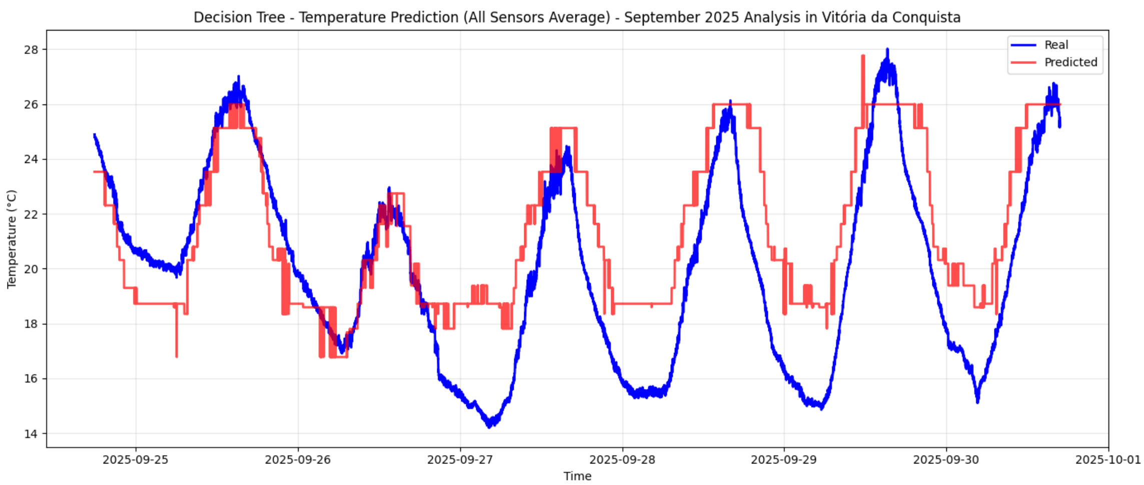

Figure 31: Decision tree structure for temperature prediction

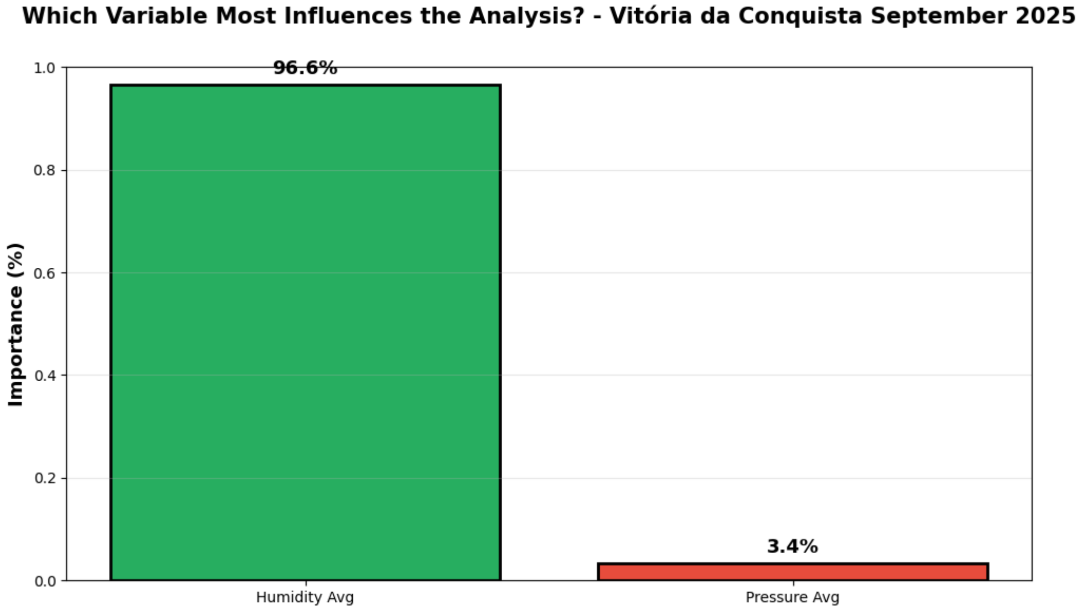

Figure 32: Feature importance for temperature prediction

Most important features:

- AHT20 and BMP280 among the main predictors
- Humidity has strong predictive power for temperature
- Model presents interpretable rules

5.3 Notebook 09: Clustering with Gaussian Mixture Model (GMM)

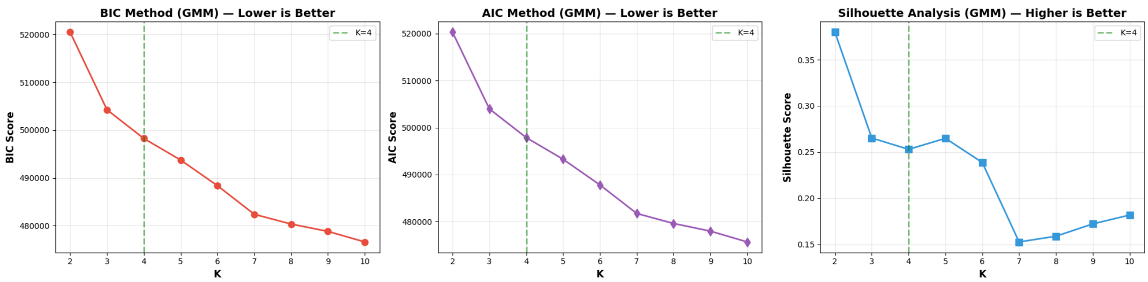

Figure 33: GMM Clustering - 2D visualization (K=4 climate regimes)

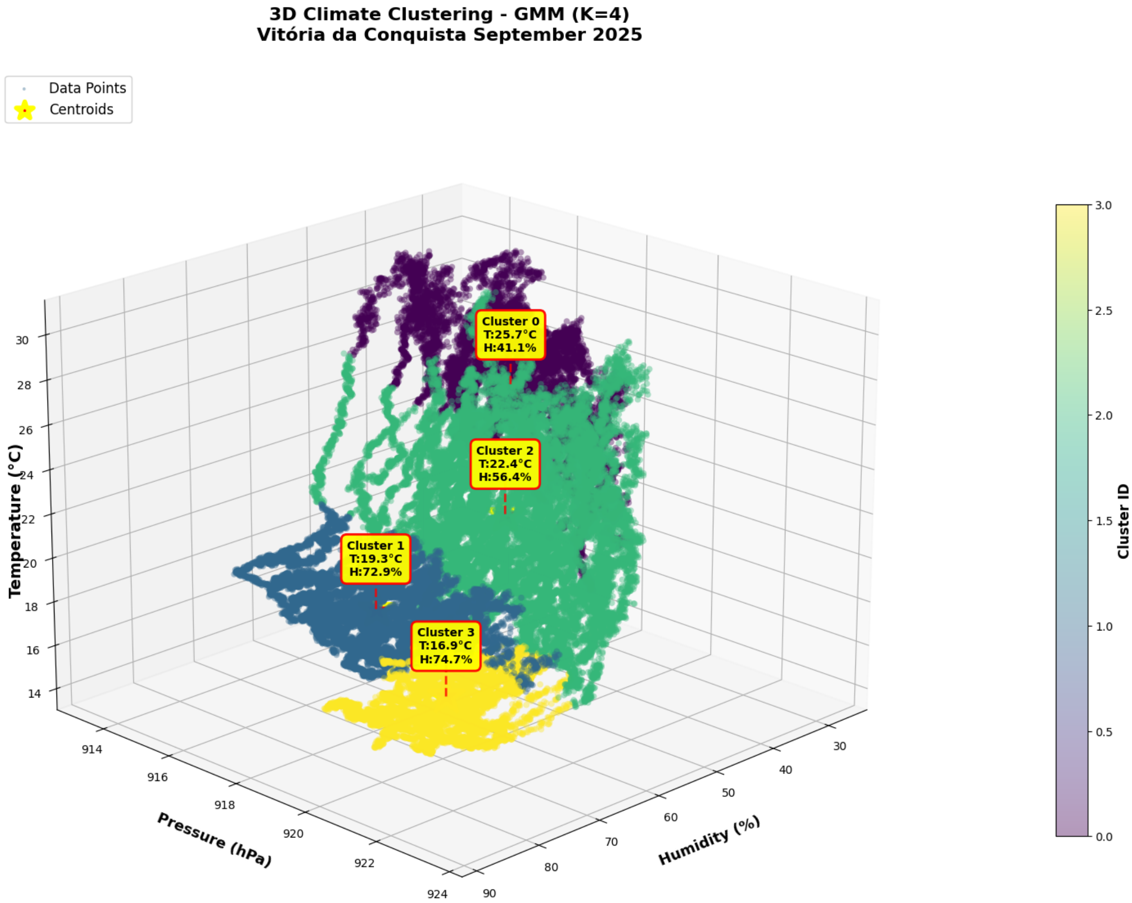

Figure 34: GMM Clustering - 3D visualization (Temperature, Humidity, Pressure)

4 identified climate regimes:

1. **Cluster 1 (Blue):** Cold-Humid (nighttime with fog)
2. **Cluster 2 (Orange):** Hot-Dry (afternoon with maximums)
3. **Cluster 3 (Green):** Morning transition (progressive warming)
4. **Cluster 4 (Red):** Evening transition (cooling)

5.4 Notebook 10: Clustering with KMeans

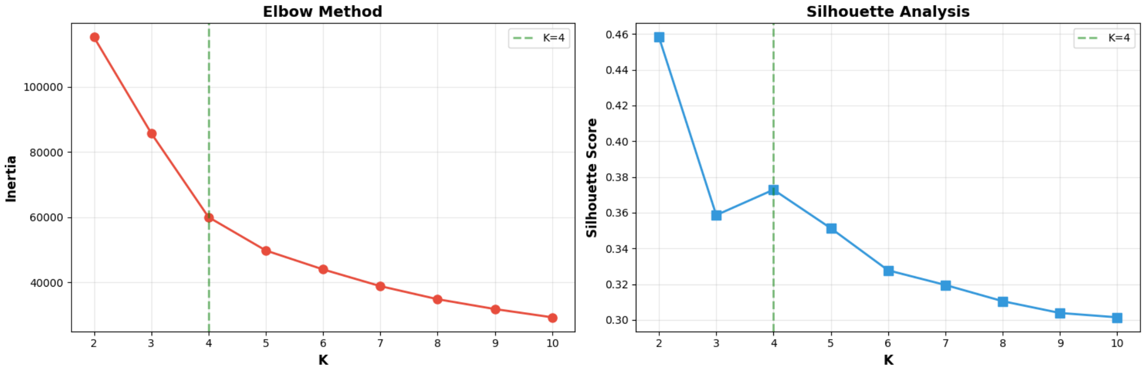

Figure 35: KMeans clustering - 2D visualization (K=4)

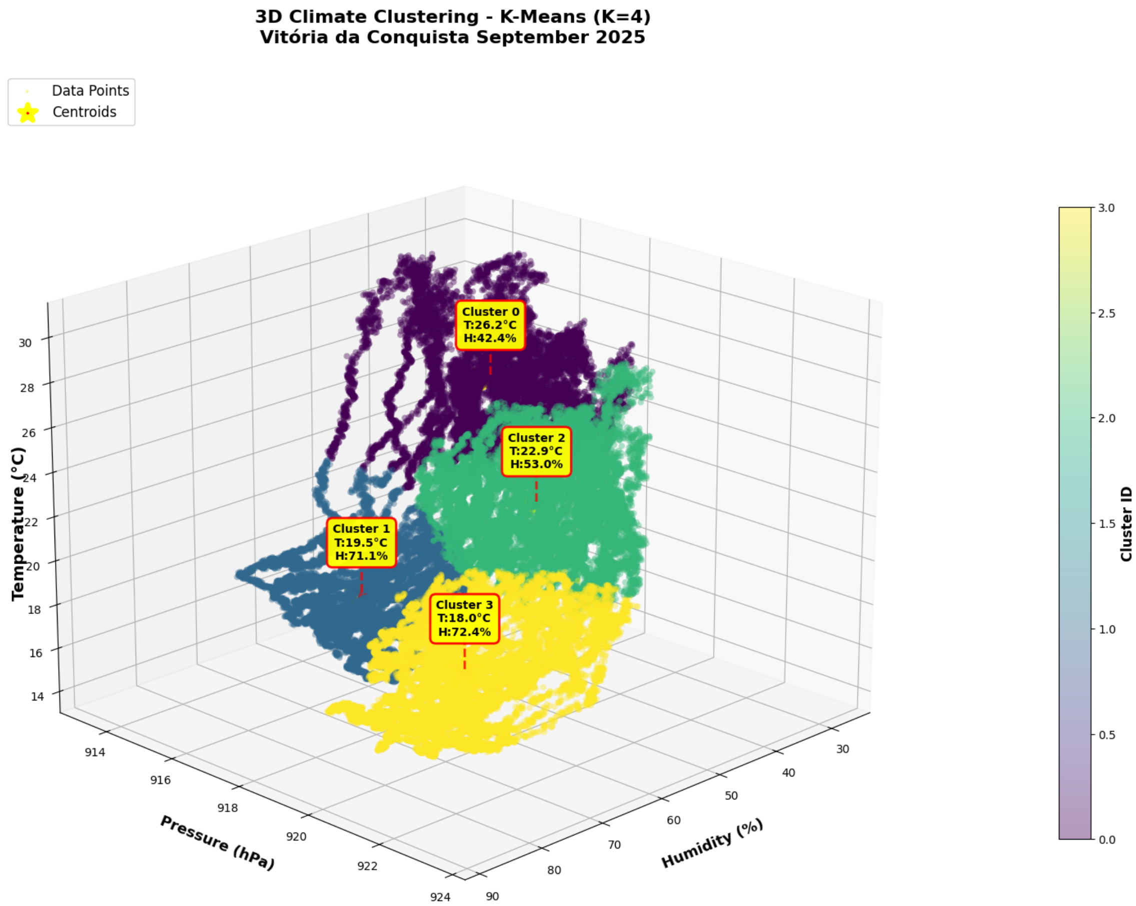

Figure 36: KMeans clustering - 3D visualization

Consistency between methods:

- KMeans validates the 4 climate states identified by GMM
- Well-separated clusters in three-dimensional space

- Confirmation of climate patterns throughout September

5.5 Notebook 11: Prediction with LSTM Neural Networks

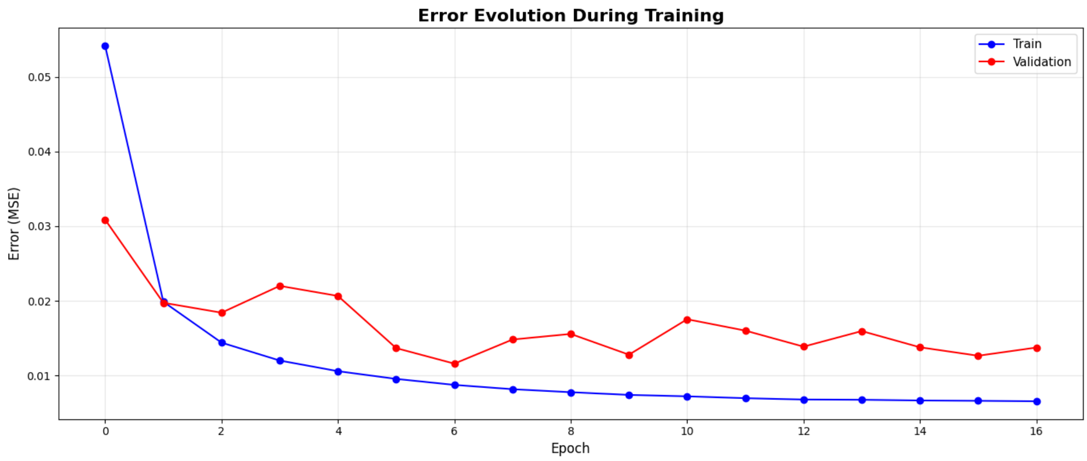

Figure 37: Error evolution during LSTM model training

Model characteristics:

- LSTM architecture with 30-minute time window
- Convergence observed after 50 epochs
- Consistent reduction of training and validation error

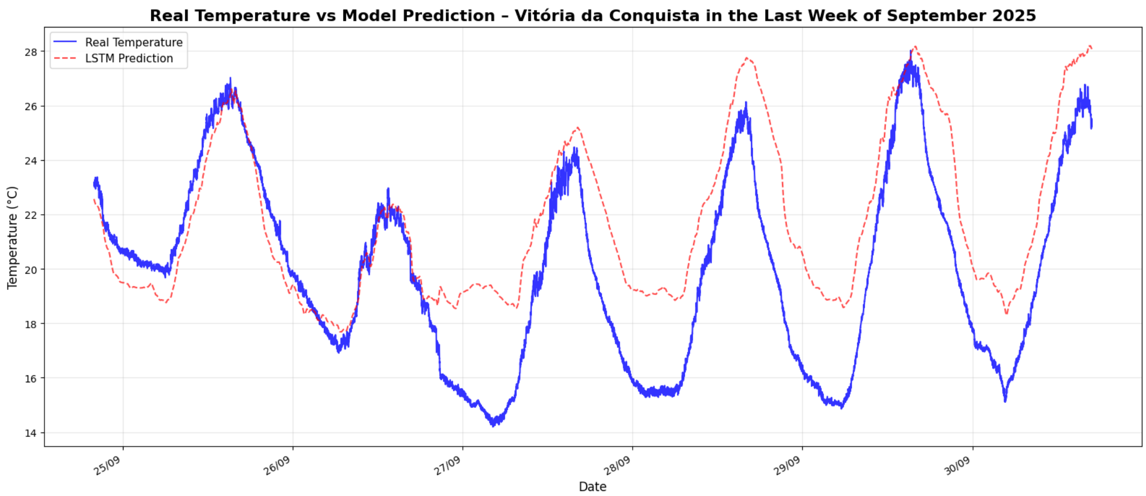

Figure 38: Real Temperature vs LSTM Predictions (last week of September 2025)

Model performance:

- **MAE ¡ 1°C** for 1-hour ahead prediction
- Model accurately captures diurnal patterns
- Predictions fit real temperatures in the validation period
- Temporal dependencies effectively captured

6 Module 3: Digital Signal Processing (Notebooks 12-13)

6.1 Notebook 12: Digital Filters

6.1.1 Filter Comparison - Temperature

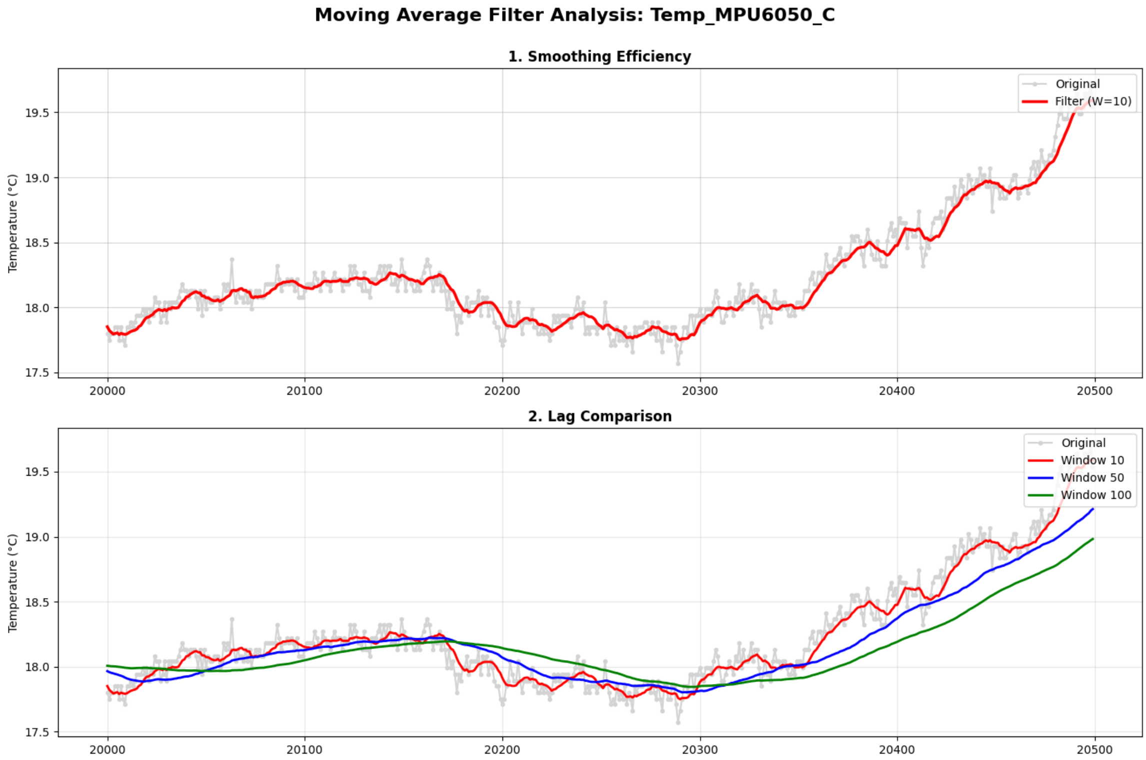

Figure 39: Comparison of digital filters - Temperature Sensor 1 (Moving Average)

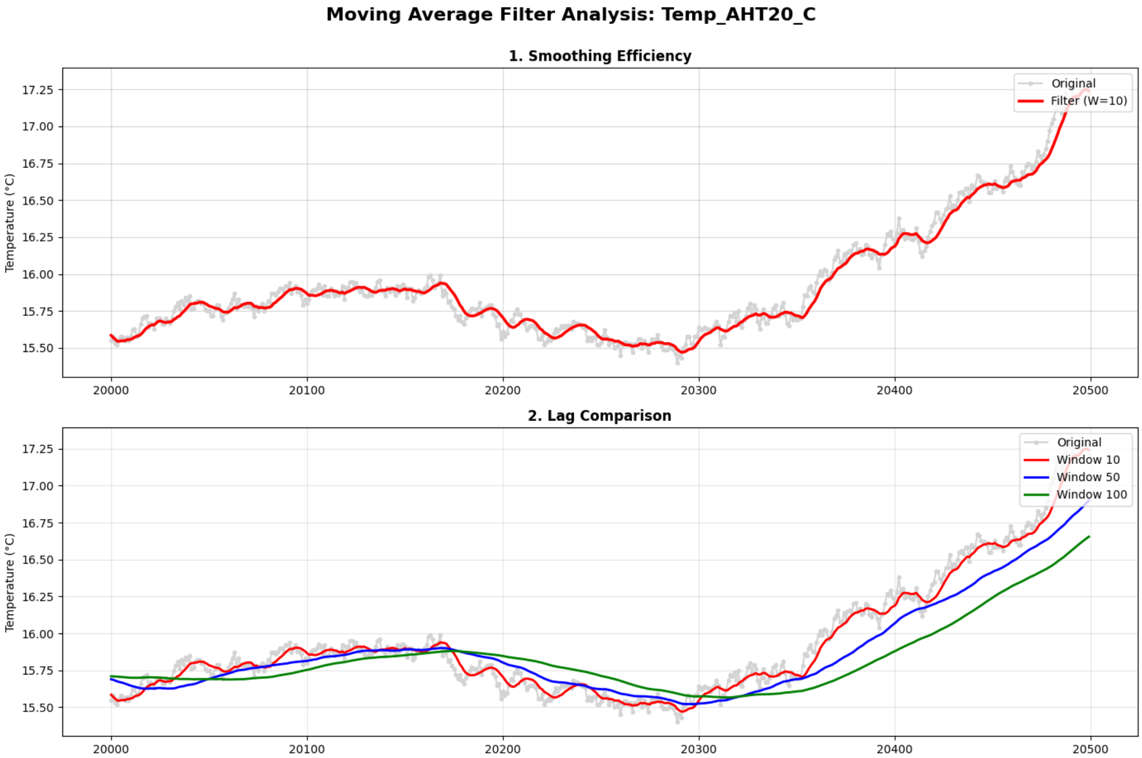

Figure 40: Comparison of digital filters - Temperature Sensor 2 (Moving Average)

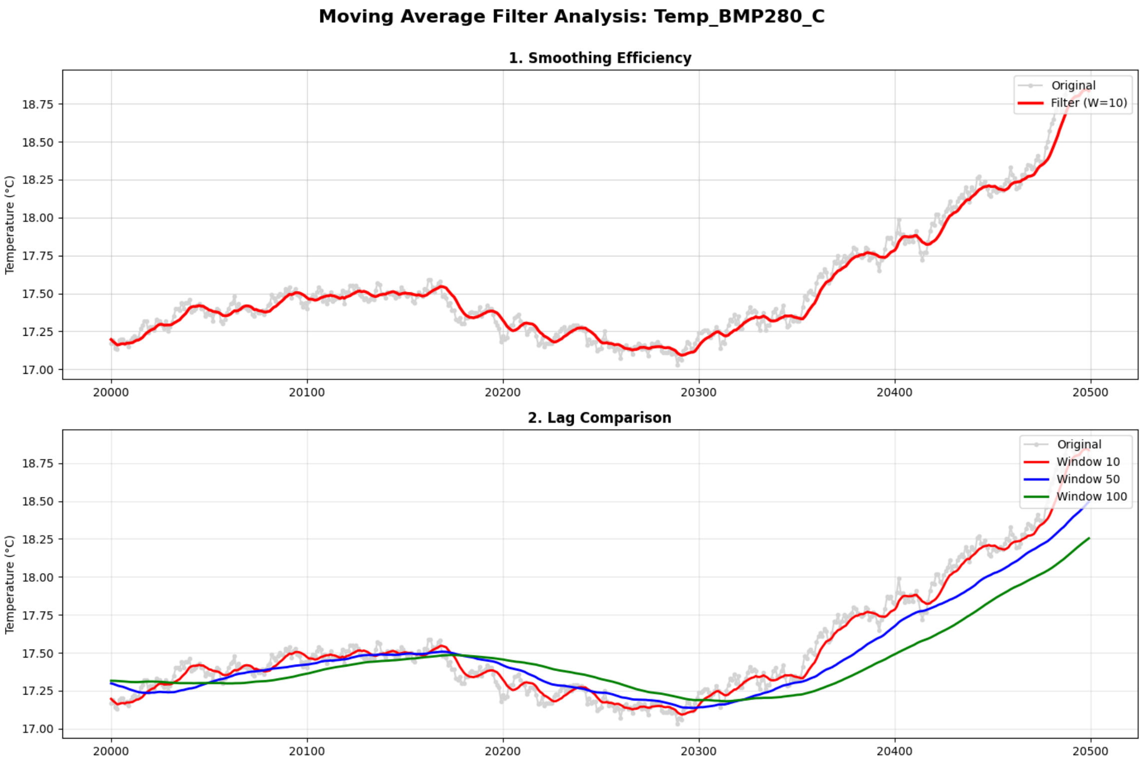

Figure 41: Comparison of digital filters - Temperature Sensor 3 (Moving Average)

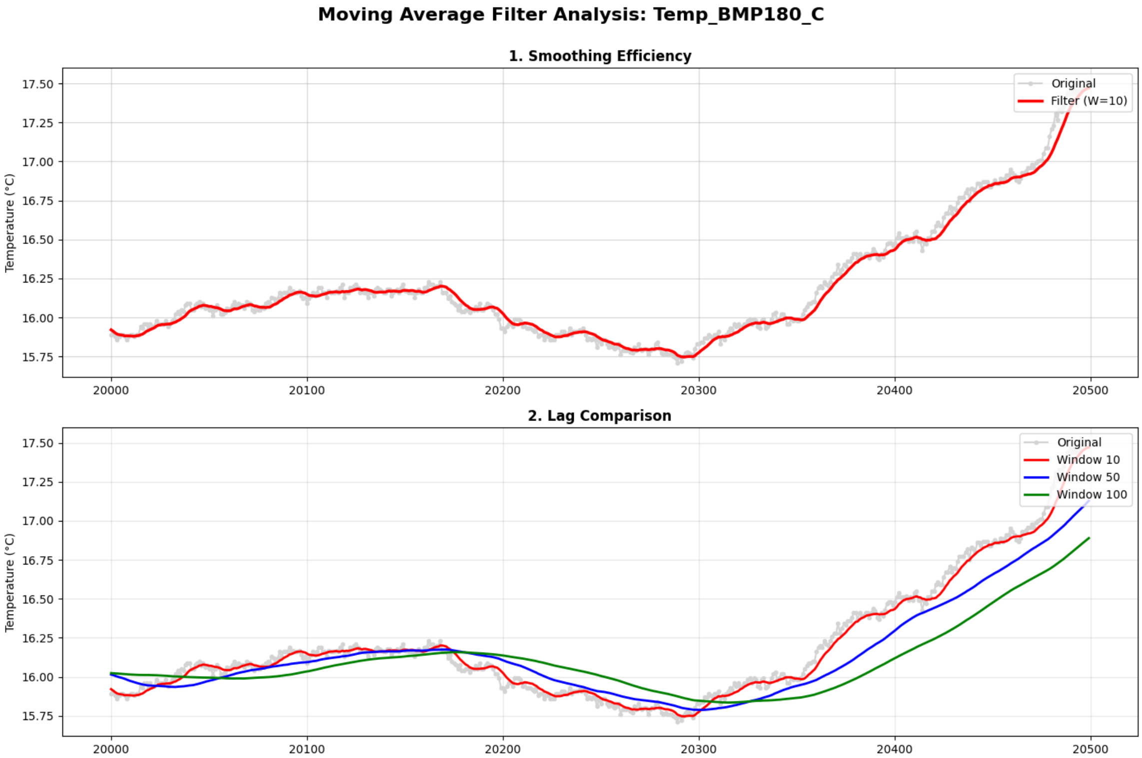

Figure 42: Comparison of digital filters - Temperature Sensor 4 (Moving Average)

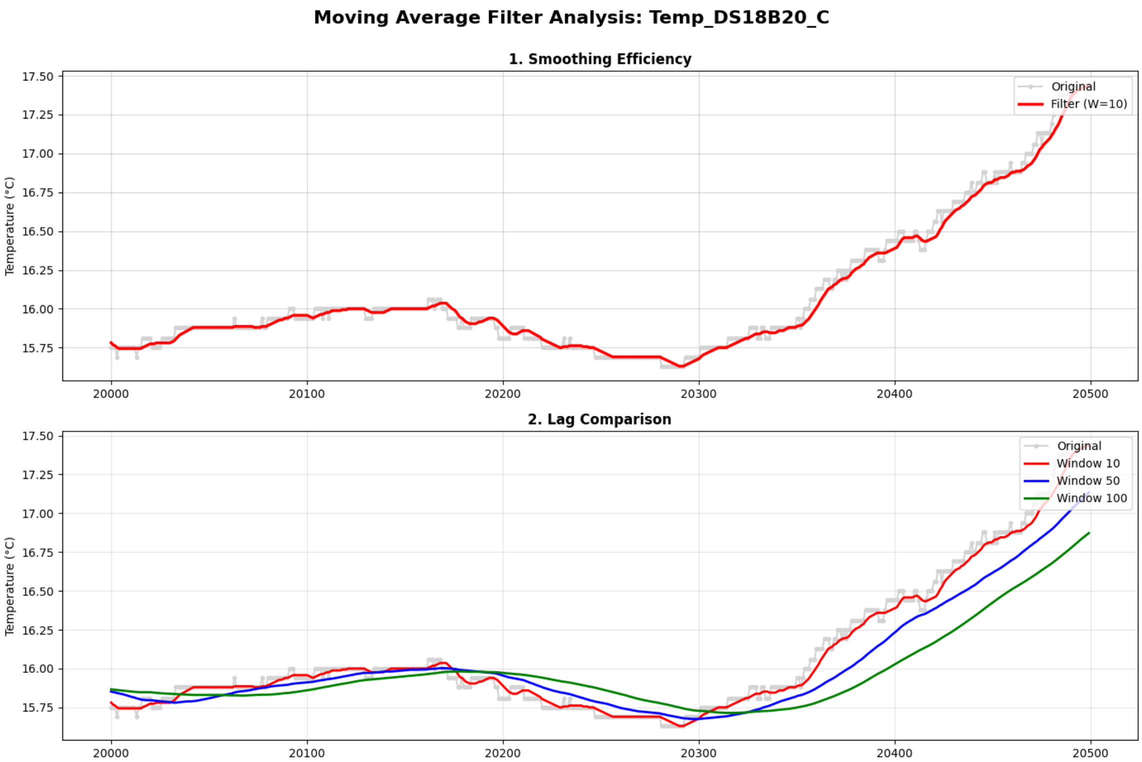

Figure 43: Comparison of digital filters - Temperature Sensor 5 (Moving Average)

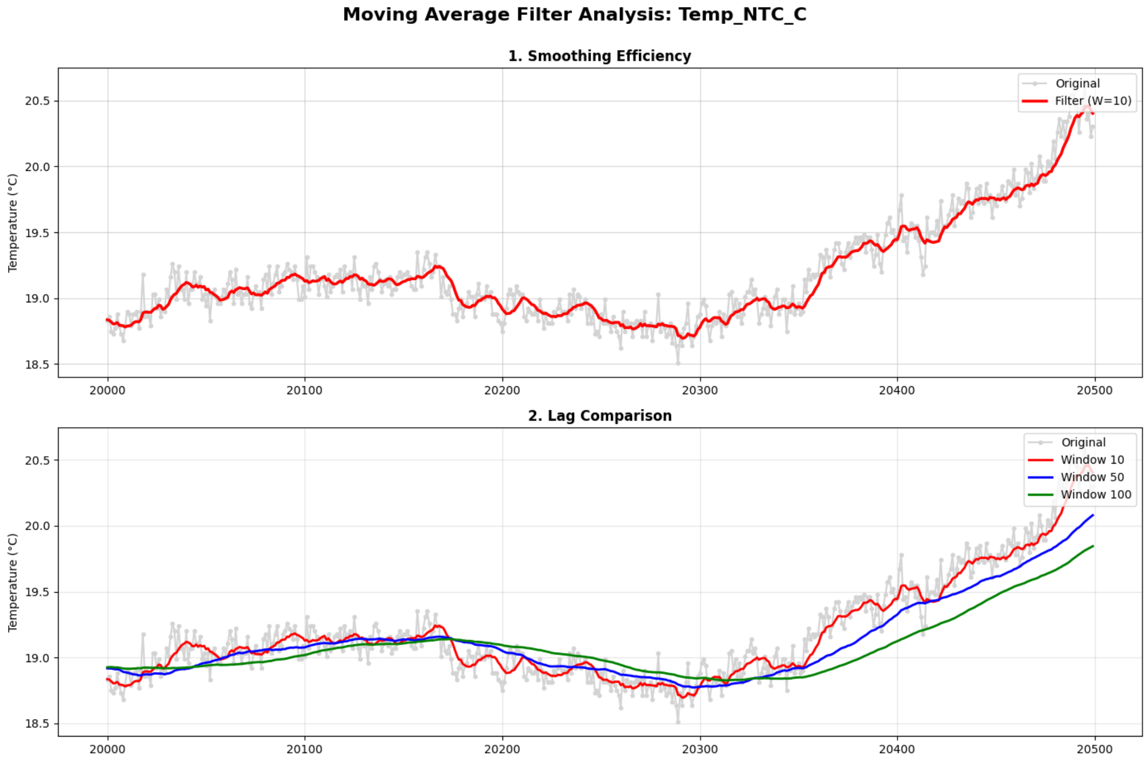

Figure 44: Comparison of digital filters - Temperature Sensor 6 (Moving Average)

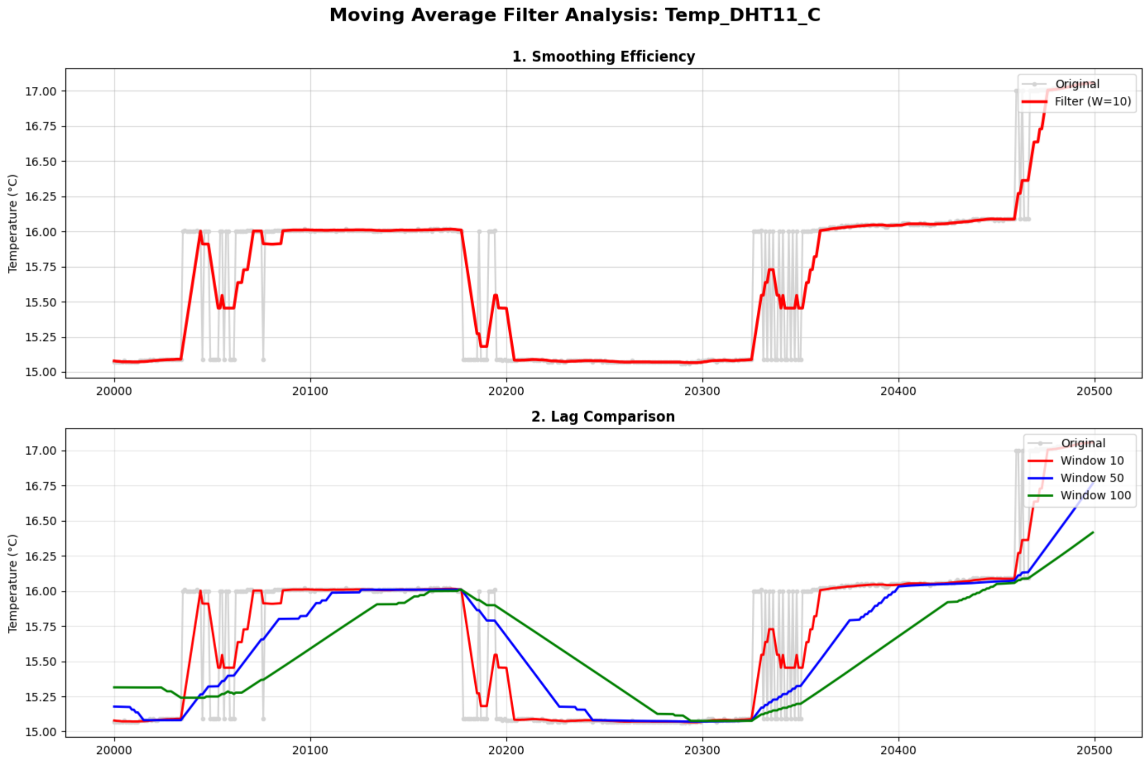

Figure 45: Comparison of digital filters - Temperature Sensor 7 (Moving Average)

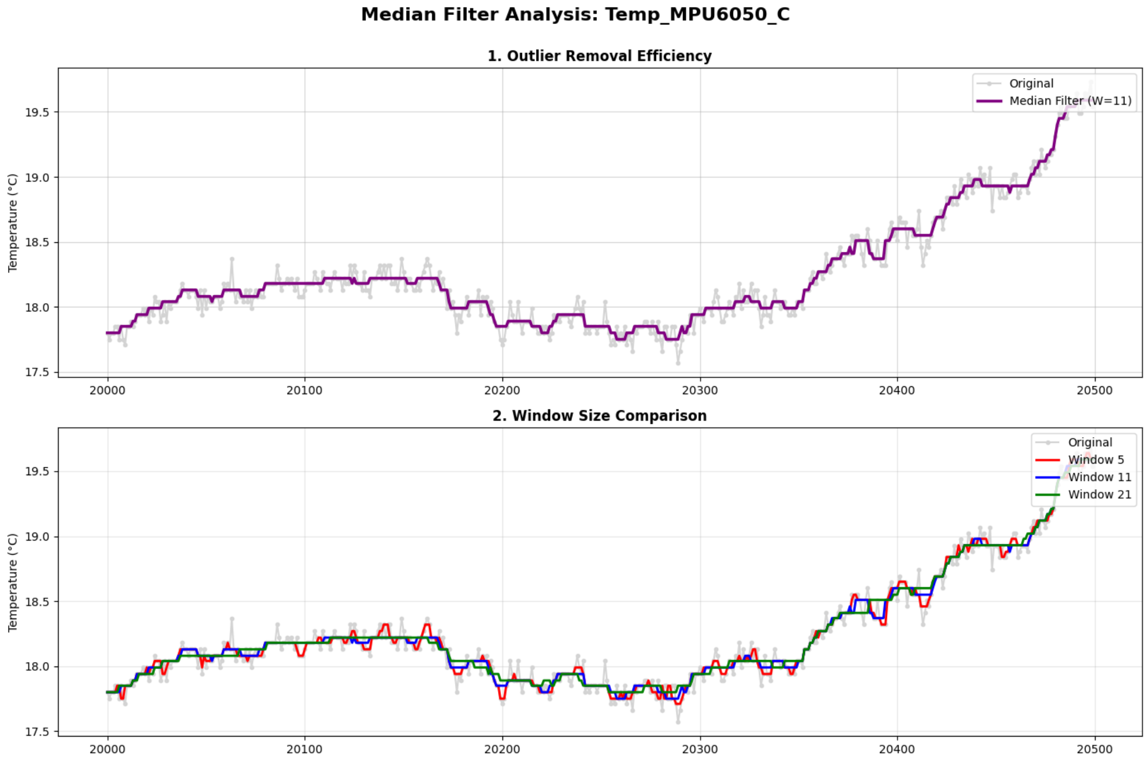

Figure 46: Comparison of digital filters - Temperature Sensor 1 (Median)

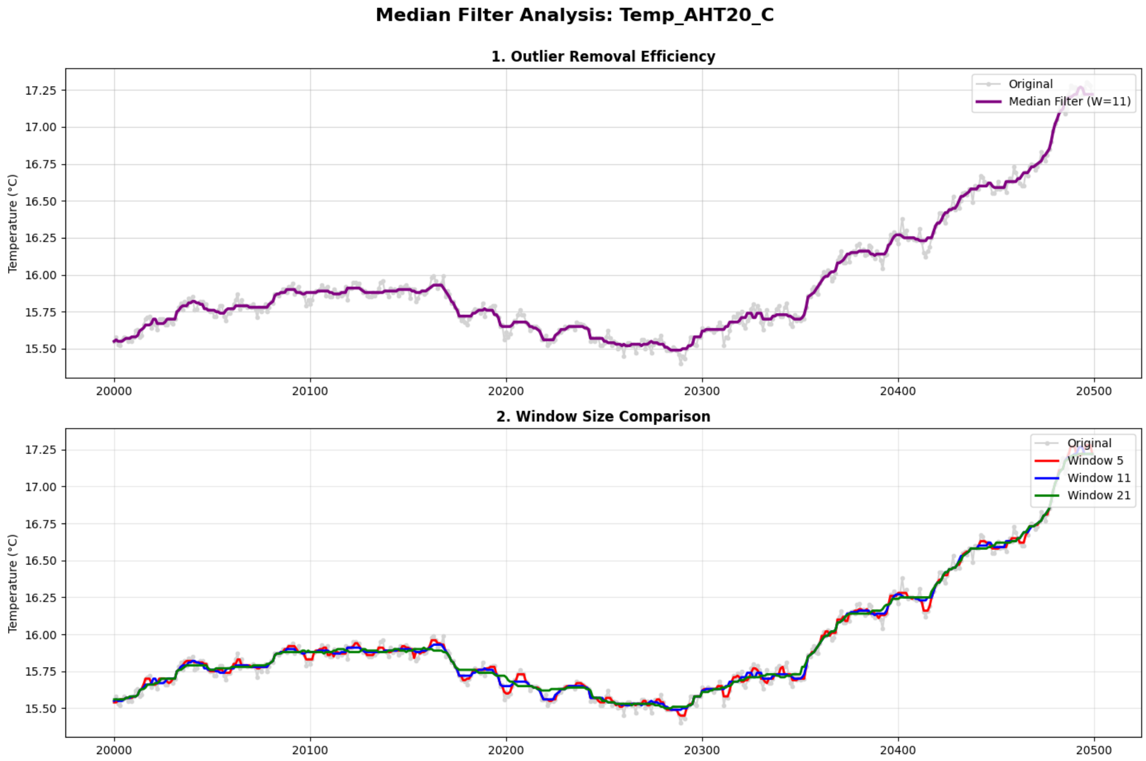

Figure 47: Comparison of digital filters - Temperature Sensor 2 (Median)

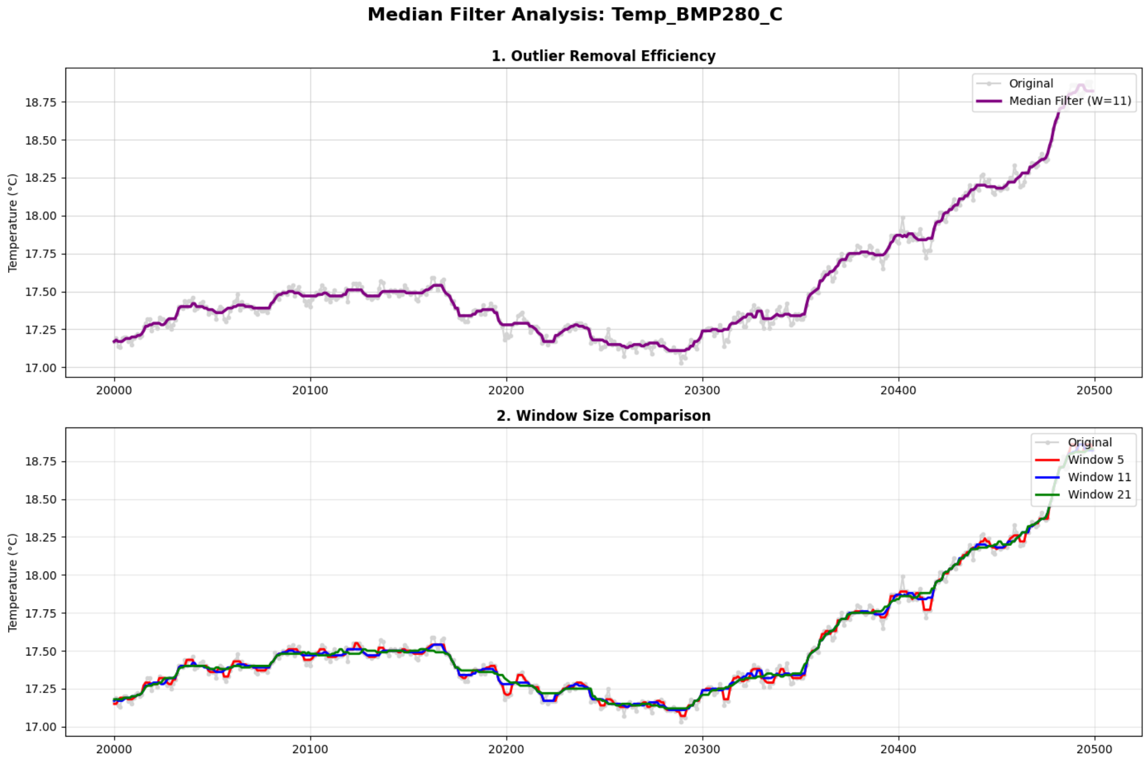

Figure 48: Comparison of digital filters - Temperature Sensor 3 (Median)

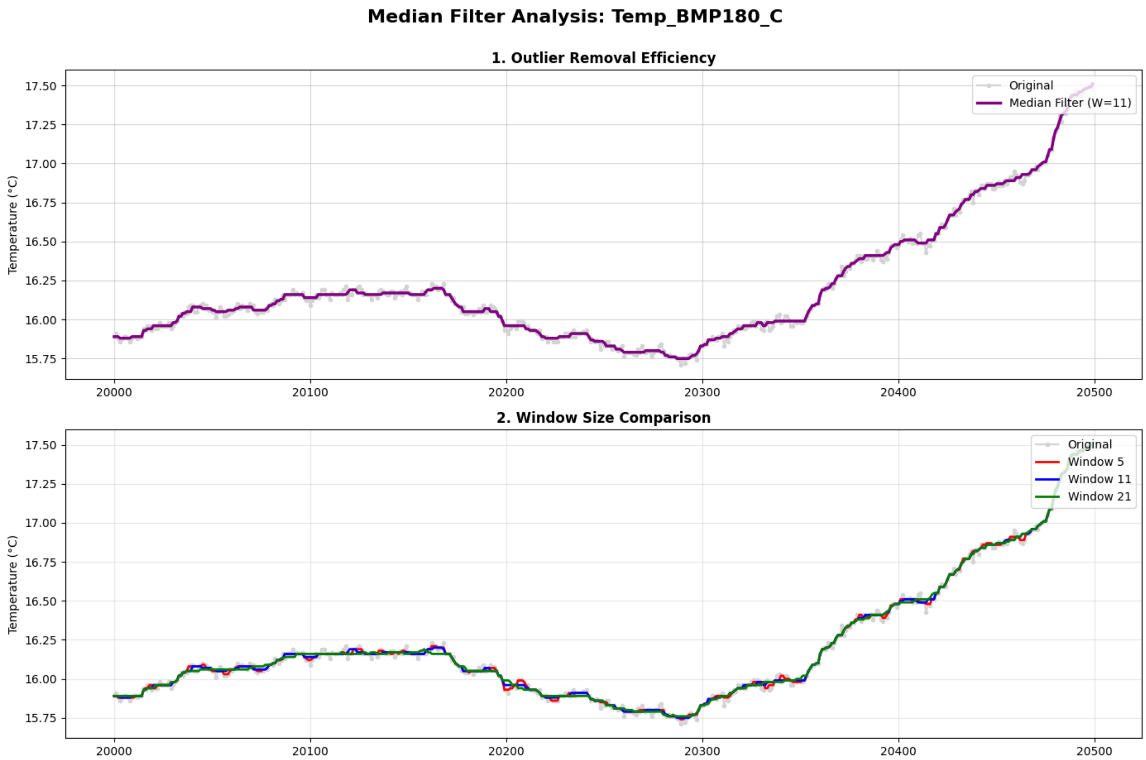

Figure 49: Comparison of digital filters - Temperature Sensor 4 (Median)

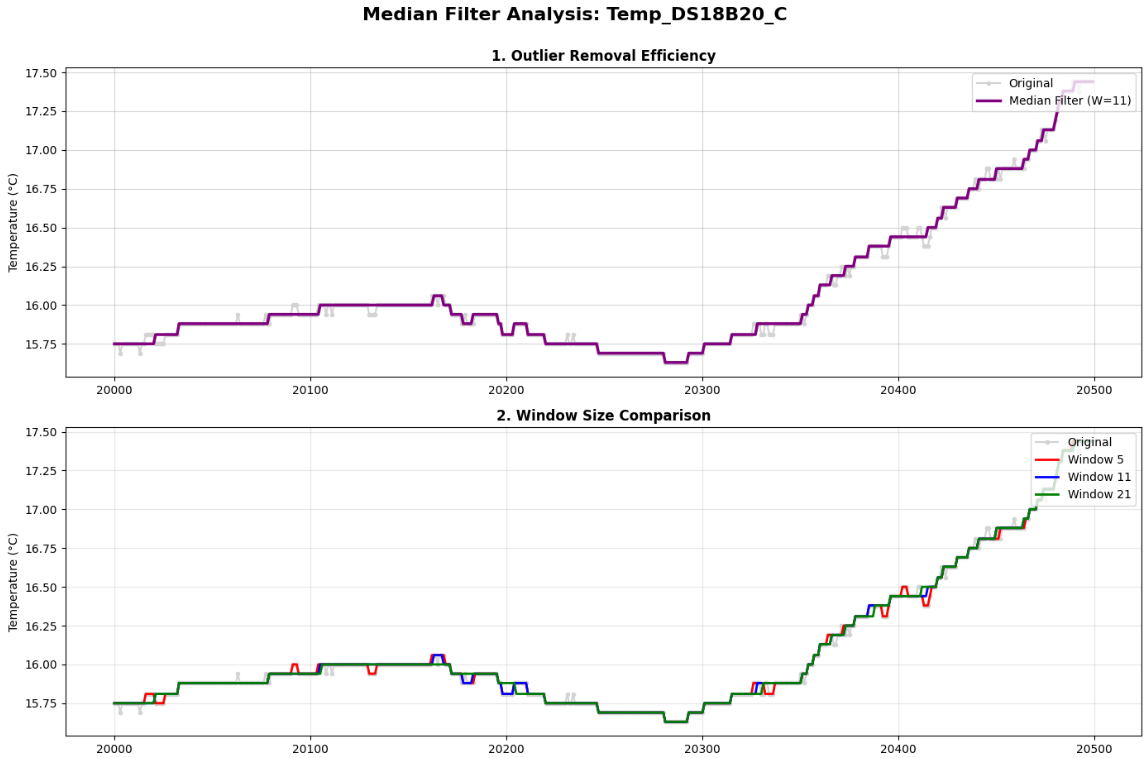

Figure 50: Comparison of digital filters - Temperature Sensor 5 (Median)

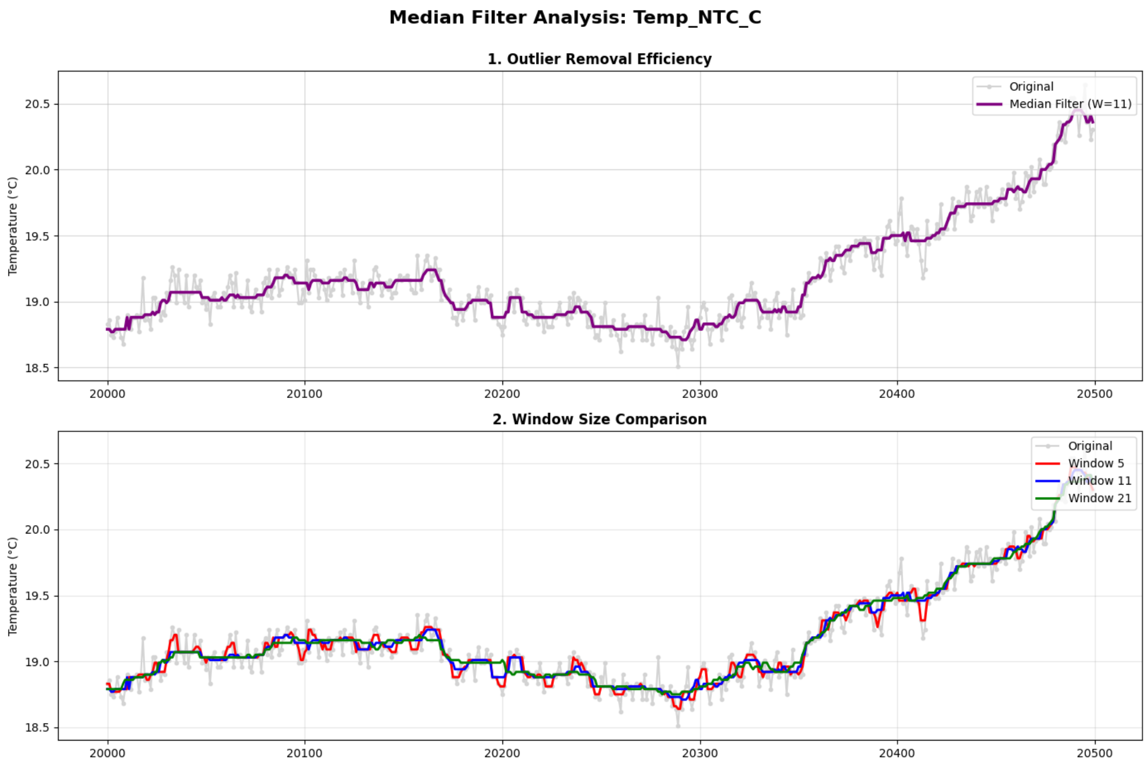

Figure 51: Comparison of digital filters - Temperature Sensor 6 (Median)

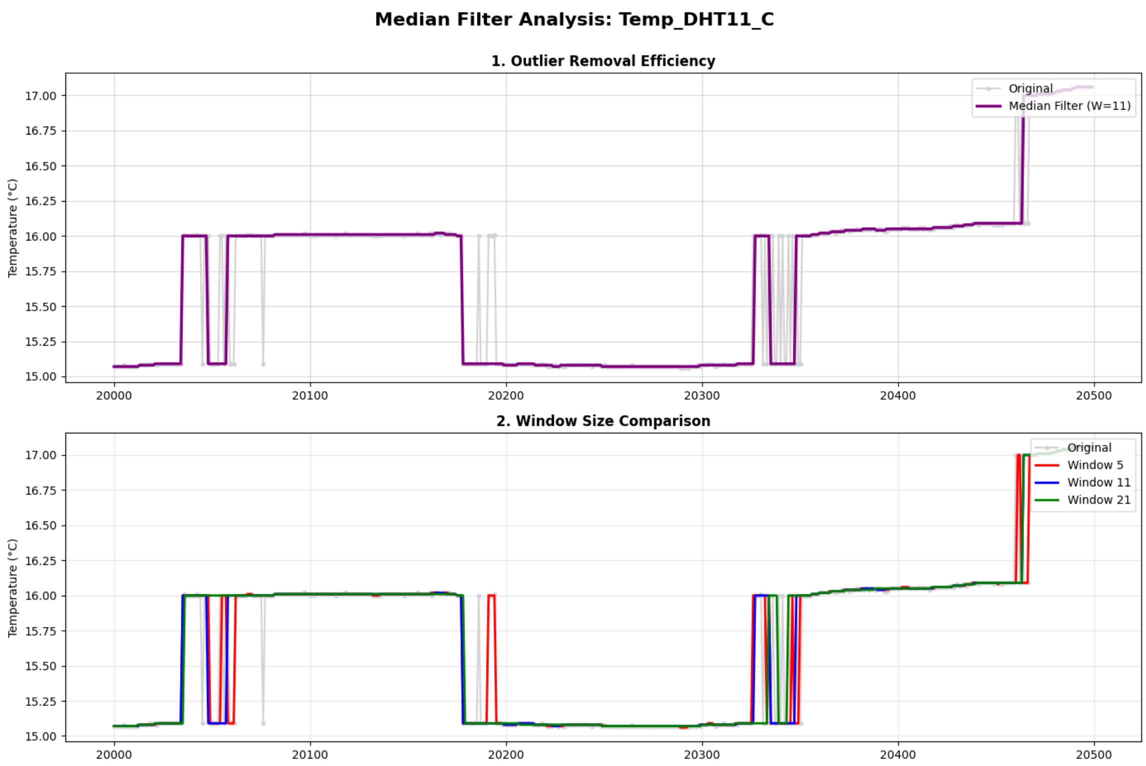

Figure 52: Comparison of digital filters - Temperature Sensor 7 (Median)

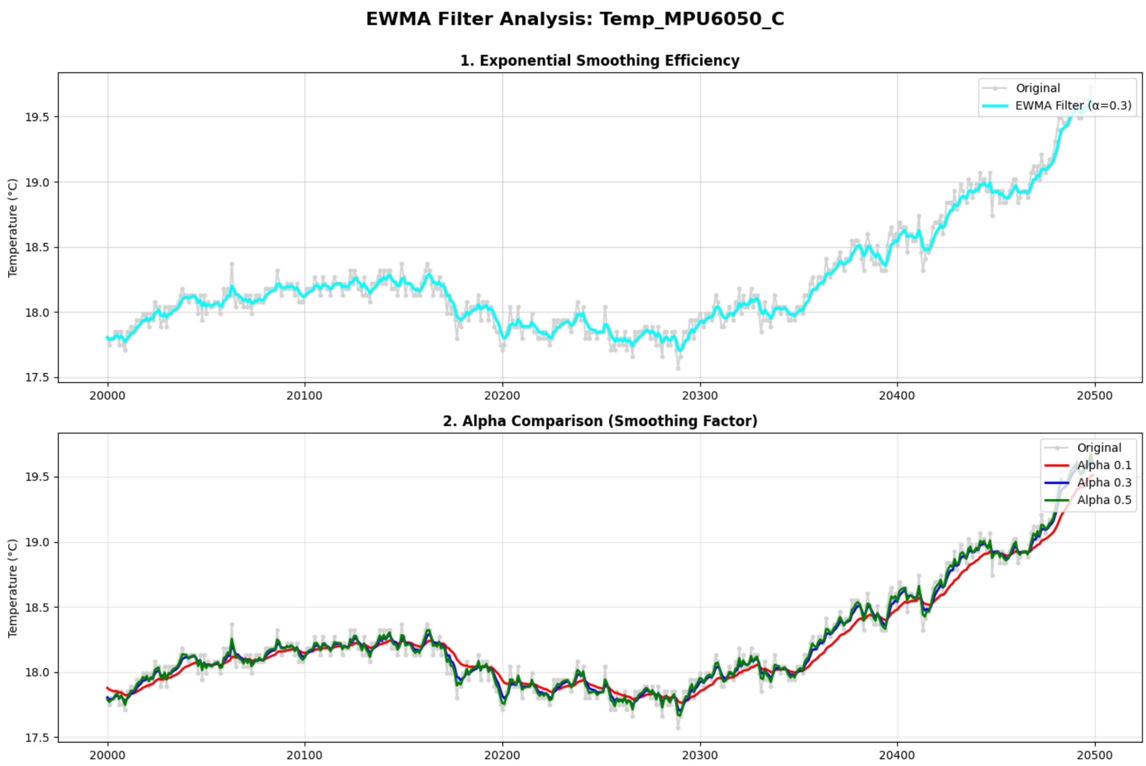

Figure 53: Comparison of digital filters - Temperature Sensor 1 (EWMA)

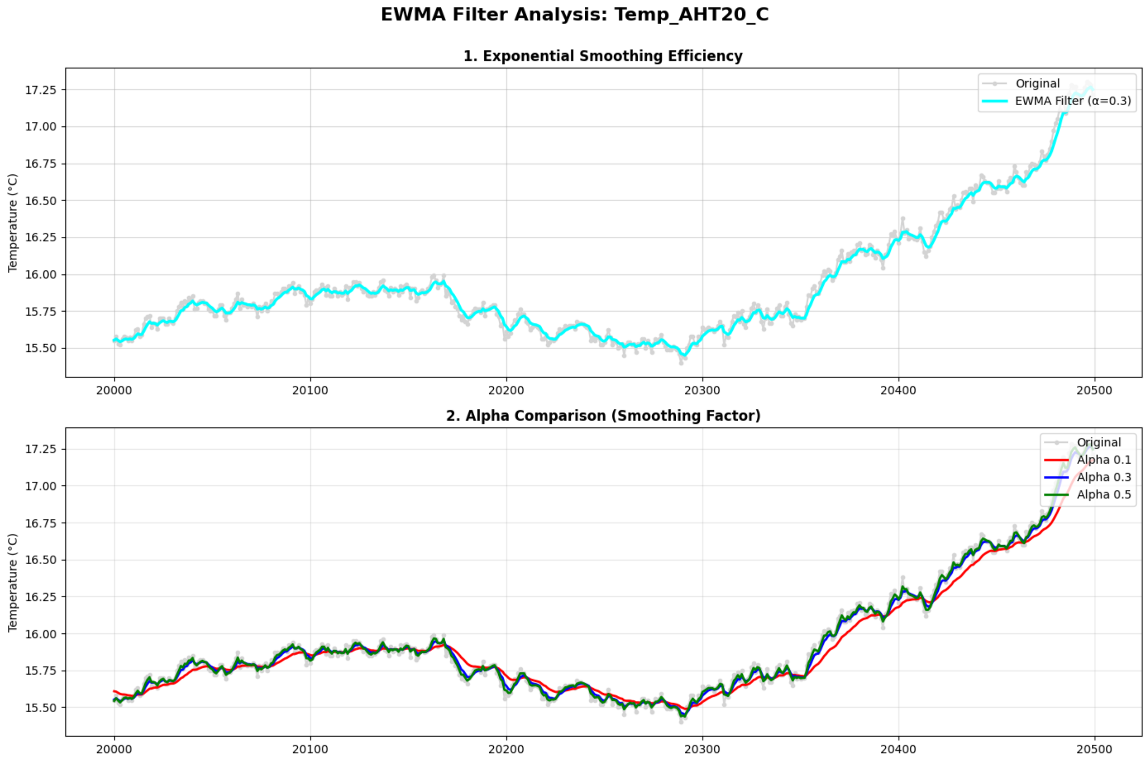

Figure 54: Comparison of digital filters - Temperature Sensor 2 (EWMA)

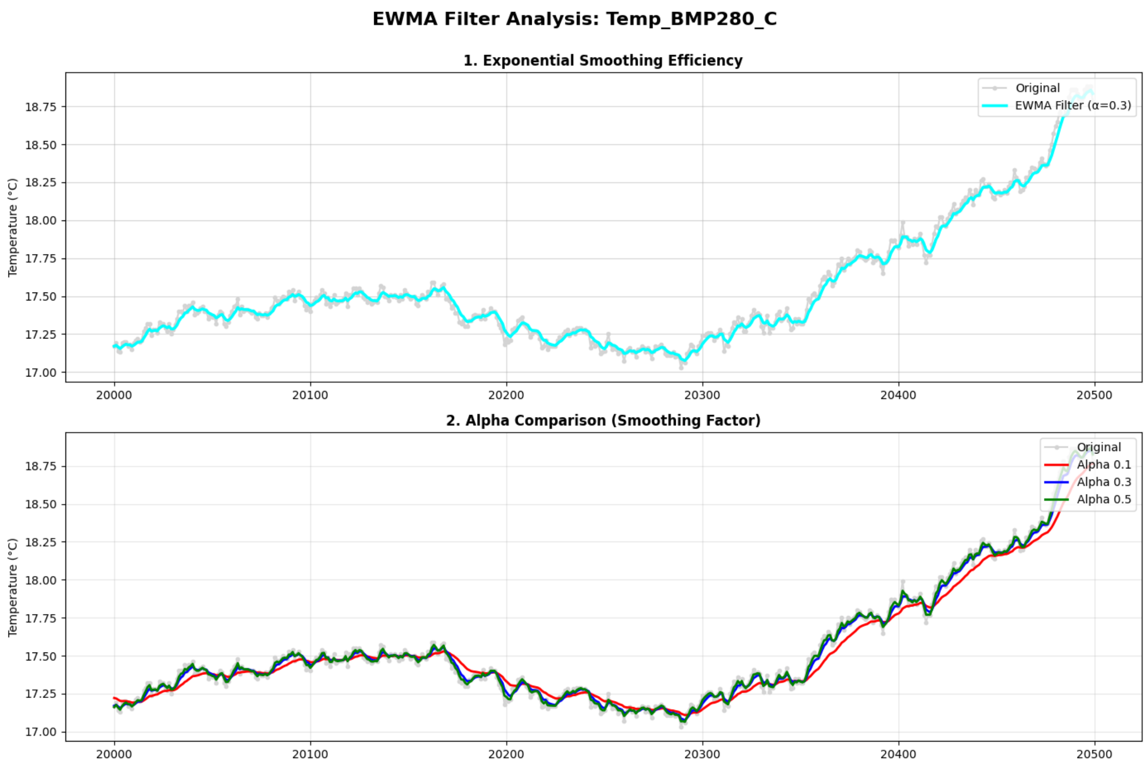

Figure 55: Comparison of digital filters - Temperature Sensor 3 (EWMA)

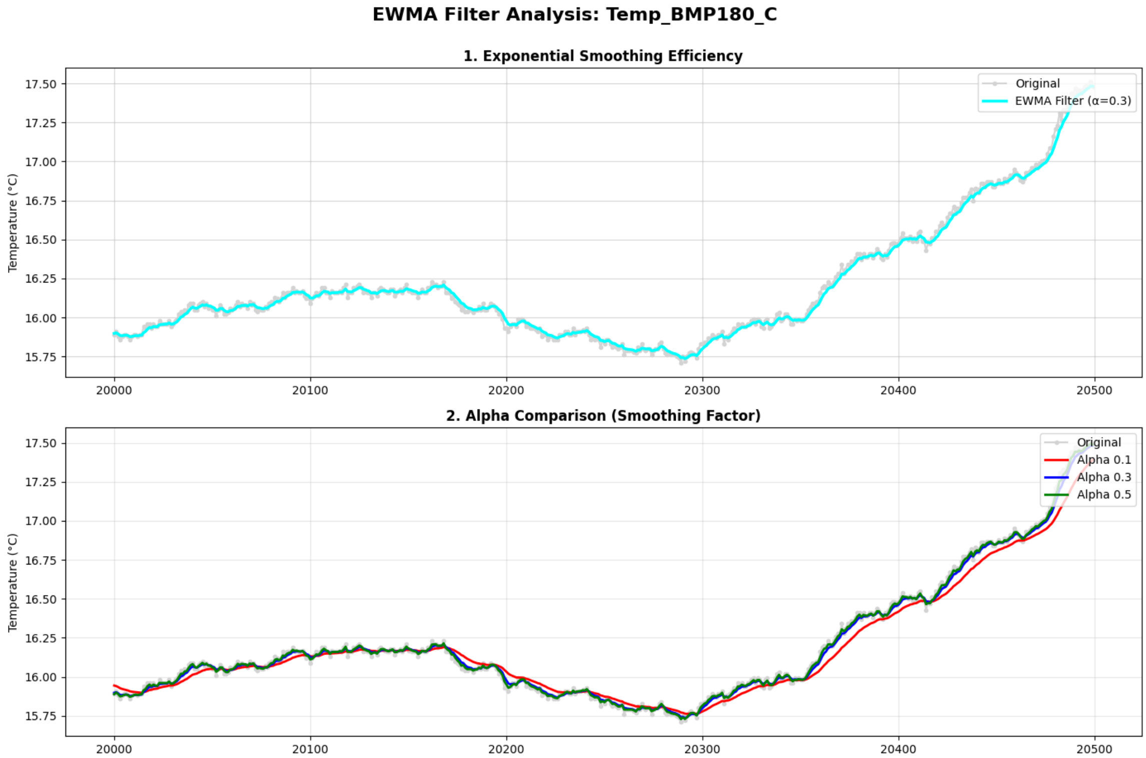

Figure 56: Comparison of digital filters - Temperature Sensor 4 (EWMA)

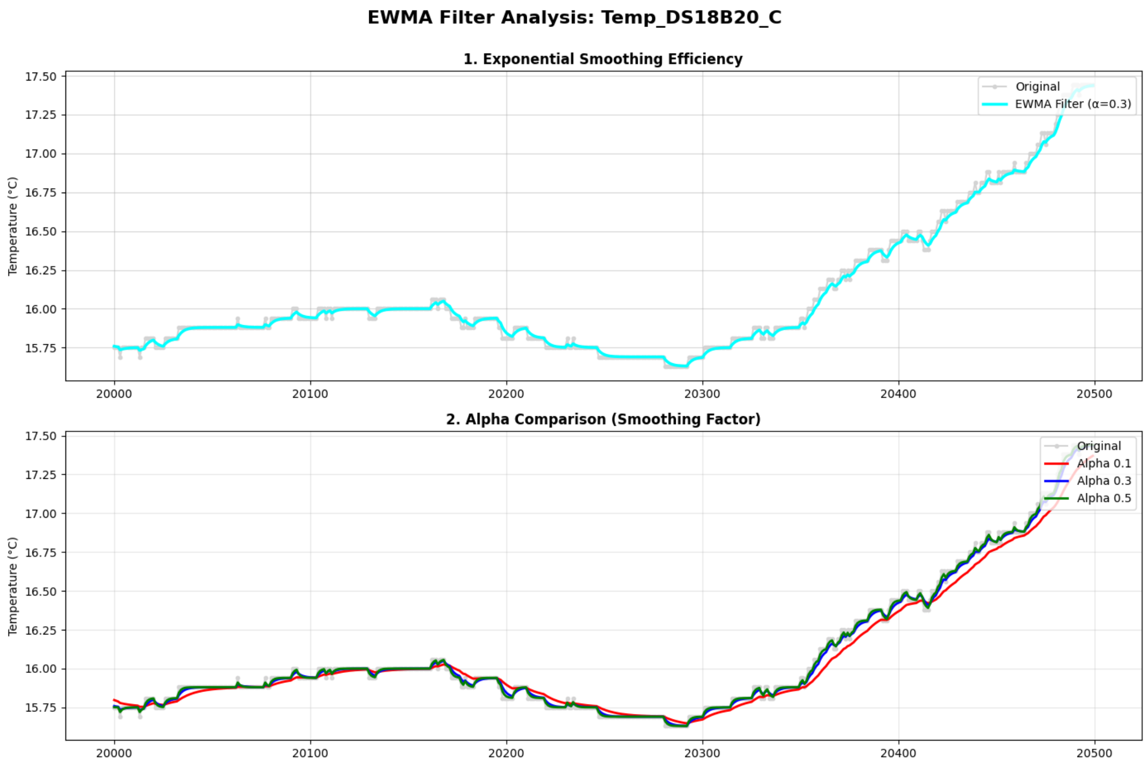

Figure 57: Comparison of digital filters - Temperature Sensor 5 (EWMA)

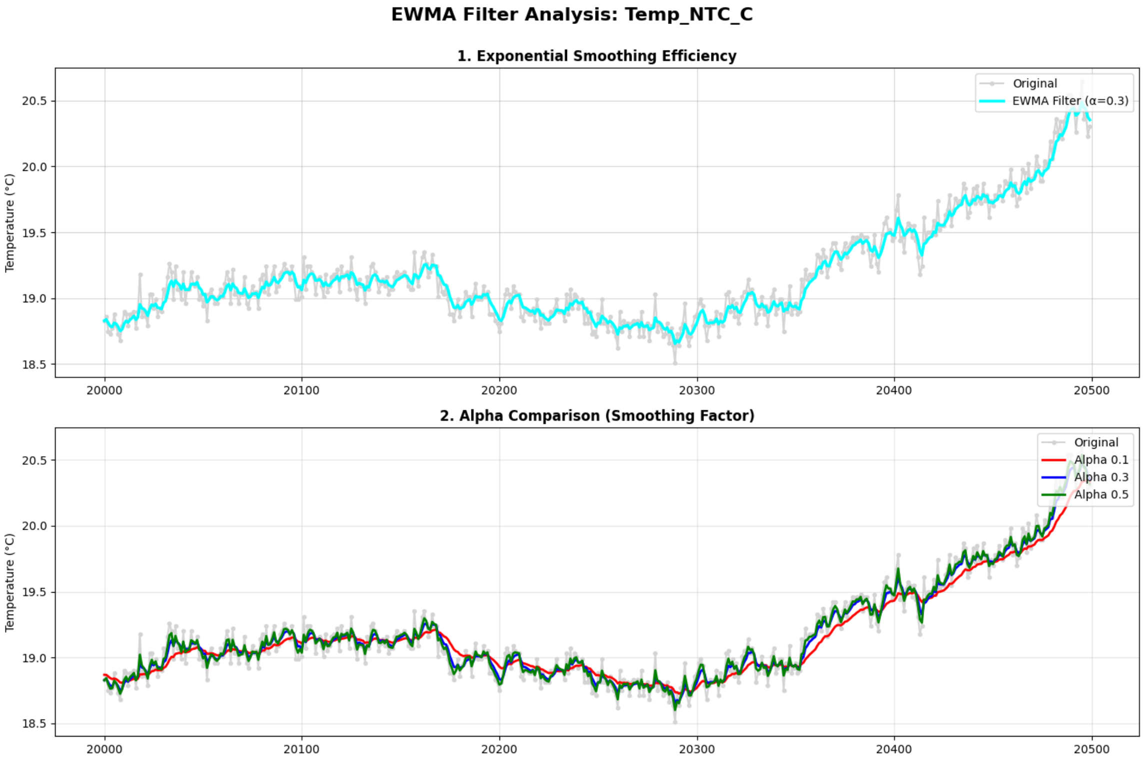

Figure 58: Comparison of digital filters - Temperature Sensor 6 (EWMA)

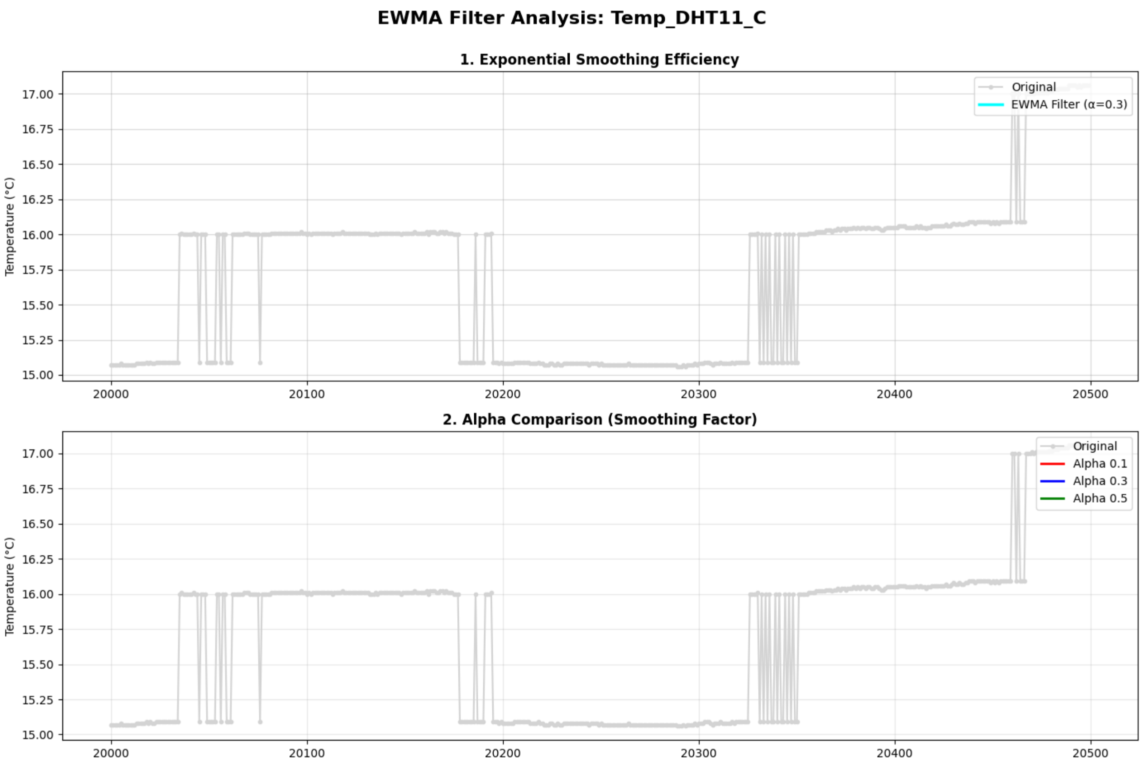

Figure 59: Comparison of digital filters - Temperature Sensor 7 (EWMA)

6.1.2 Filter Comparison - Humidity

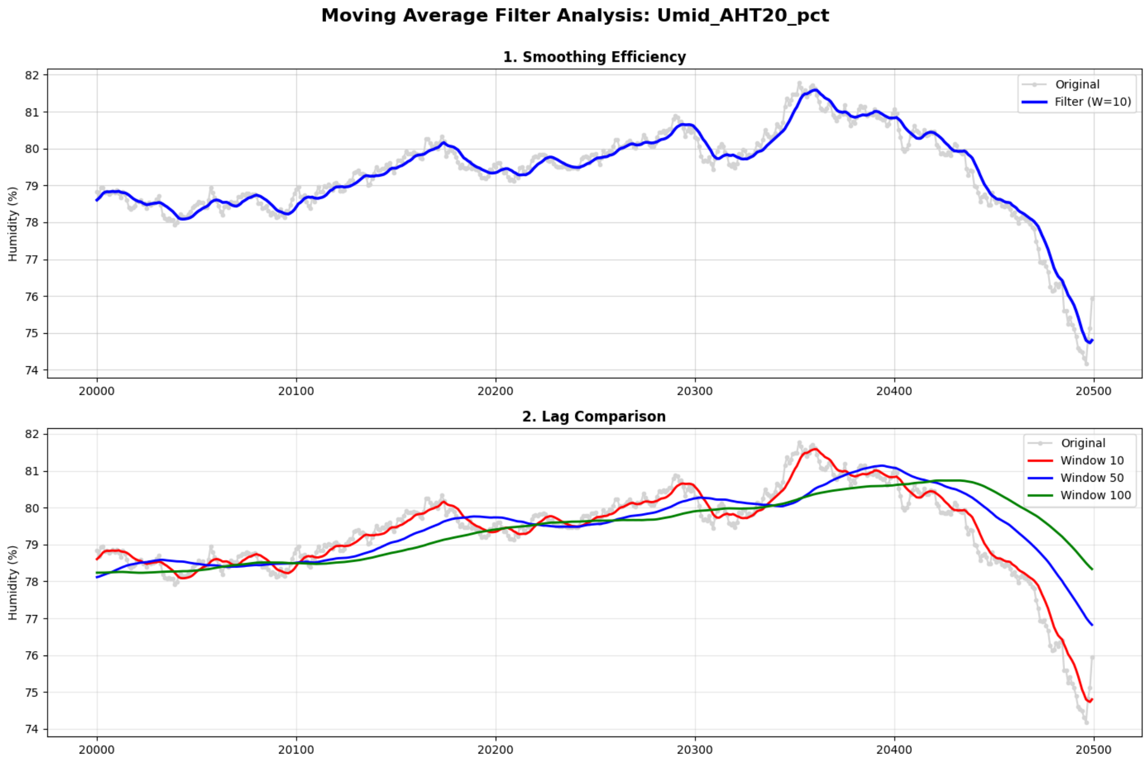

Figure 60: Comparison of digital filters - Humidity Sensor 1 (Moving Average)

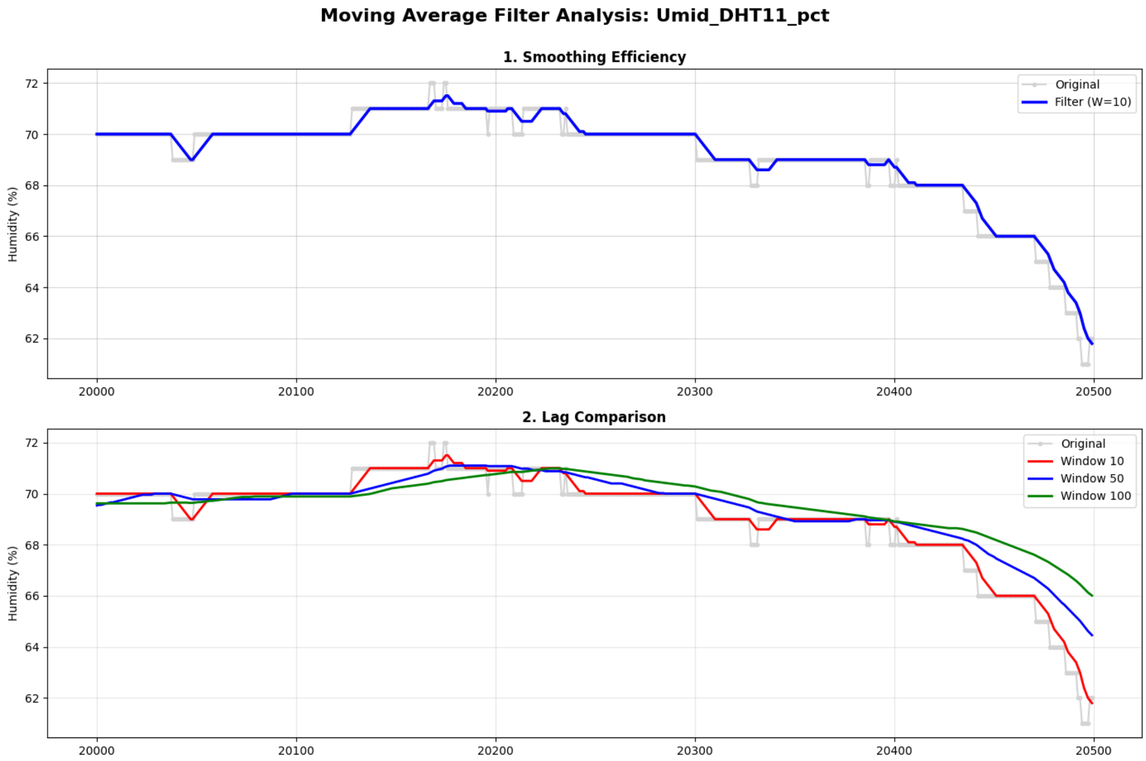

Figure 61: Comparison of digital filters - Humidity Sensor 2 (Moving Average)

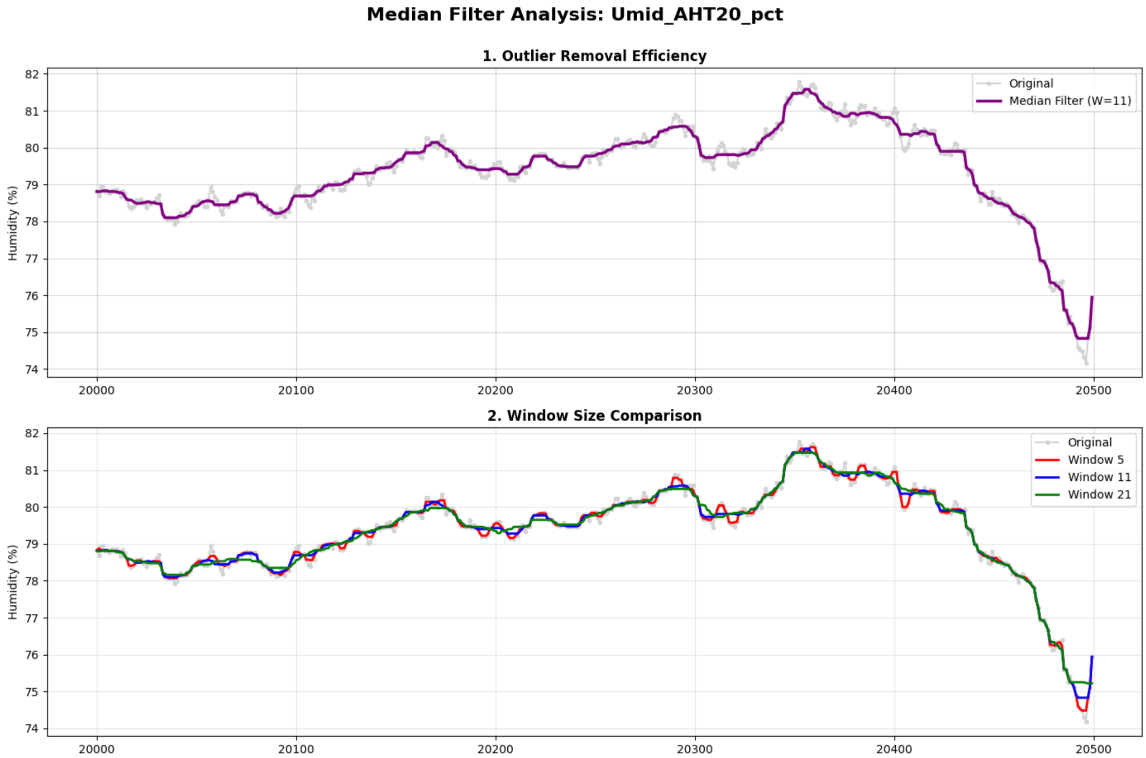

Figure 62: Comparison of digital filters - Humidity Sensor 1 (Median)

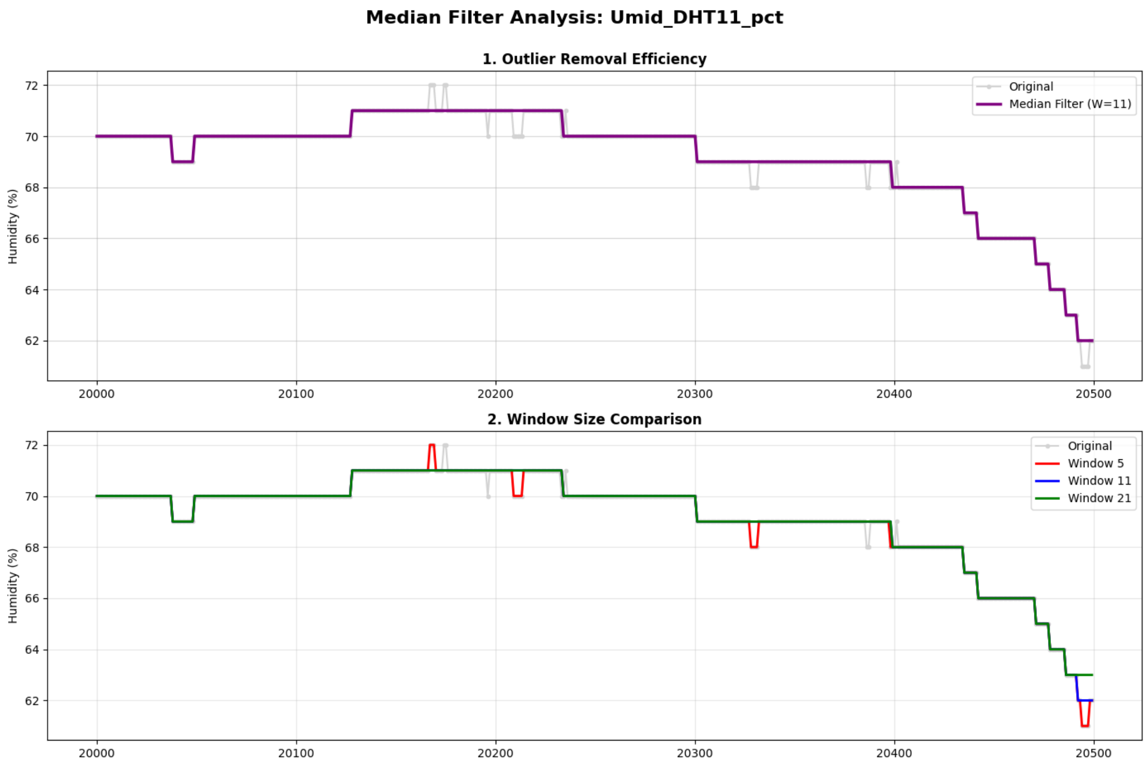

Figure 63: Comparison of digital filters - Humidity Sensor 2 (Median)

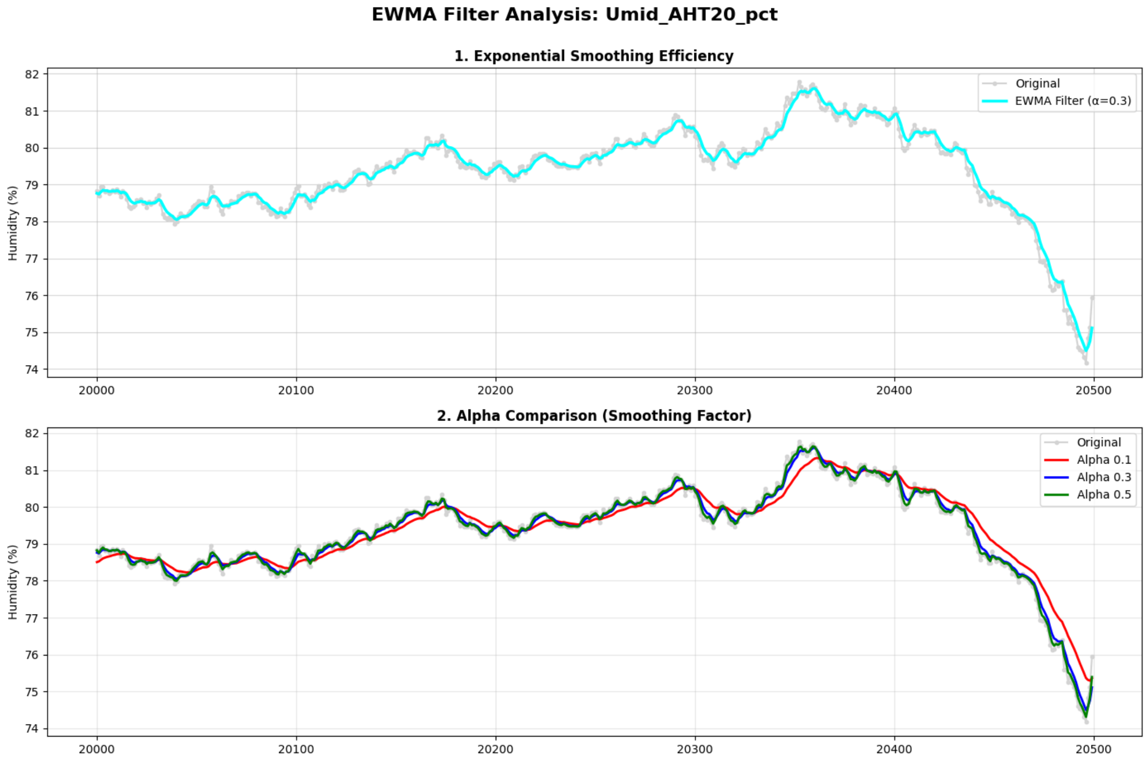

Figure 64: Comparison of digital filters - Humidity Sensor 1 (EWMA)

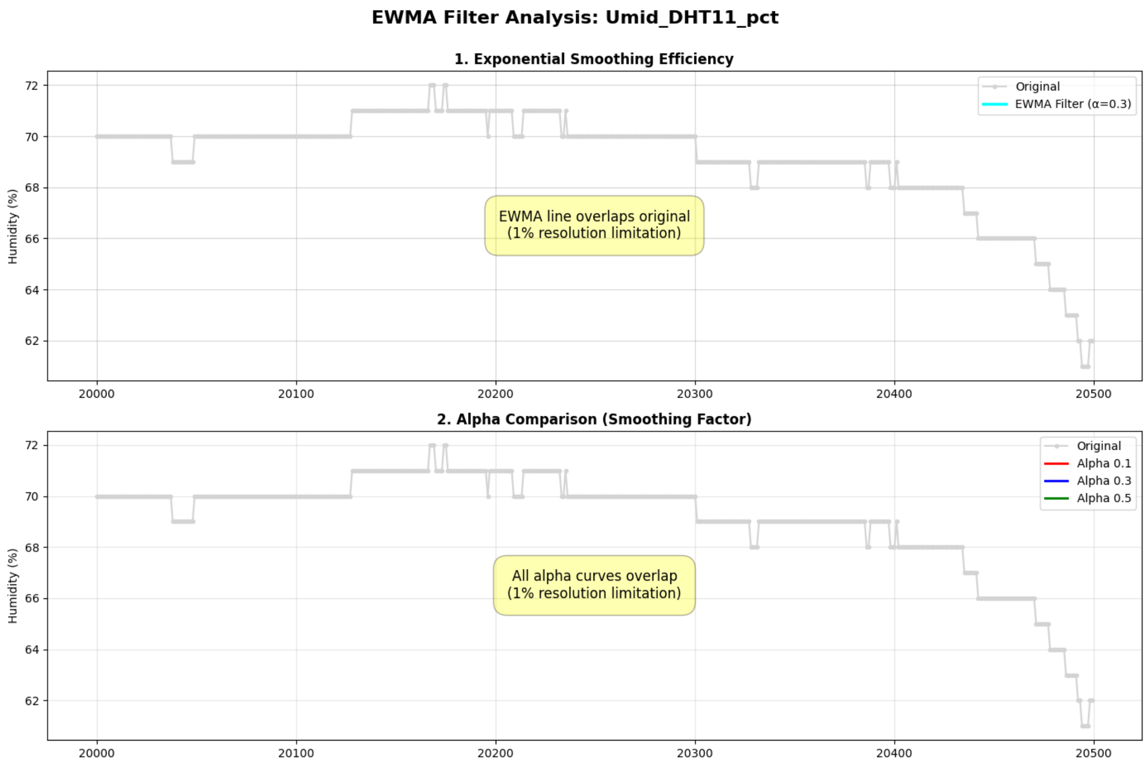

Figure 65: Comparison of digital filters - Humidity Sensor 2 (EWMA)

6.1.3 Filter Comparison - Pressure

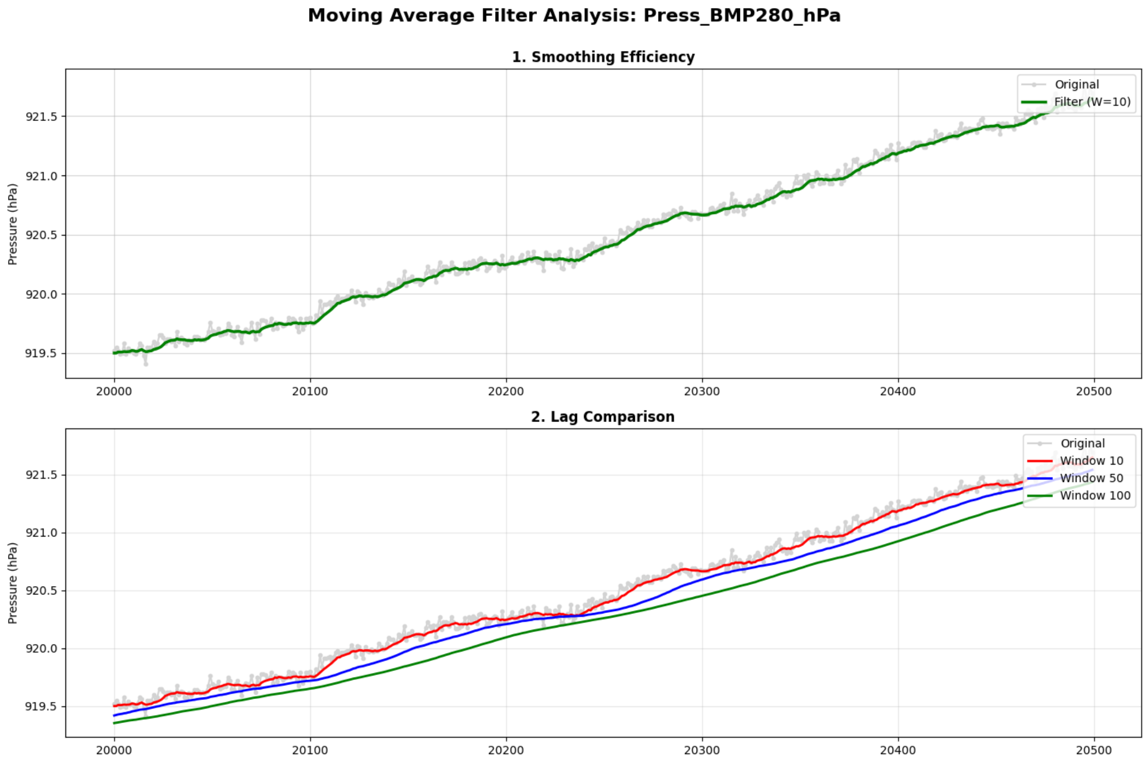

Figure 66: Comparison of digital filters - Pressure Sensor 1 (Moving Average)

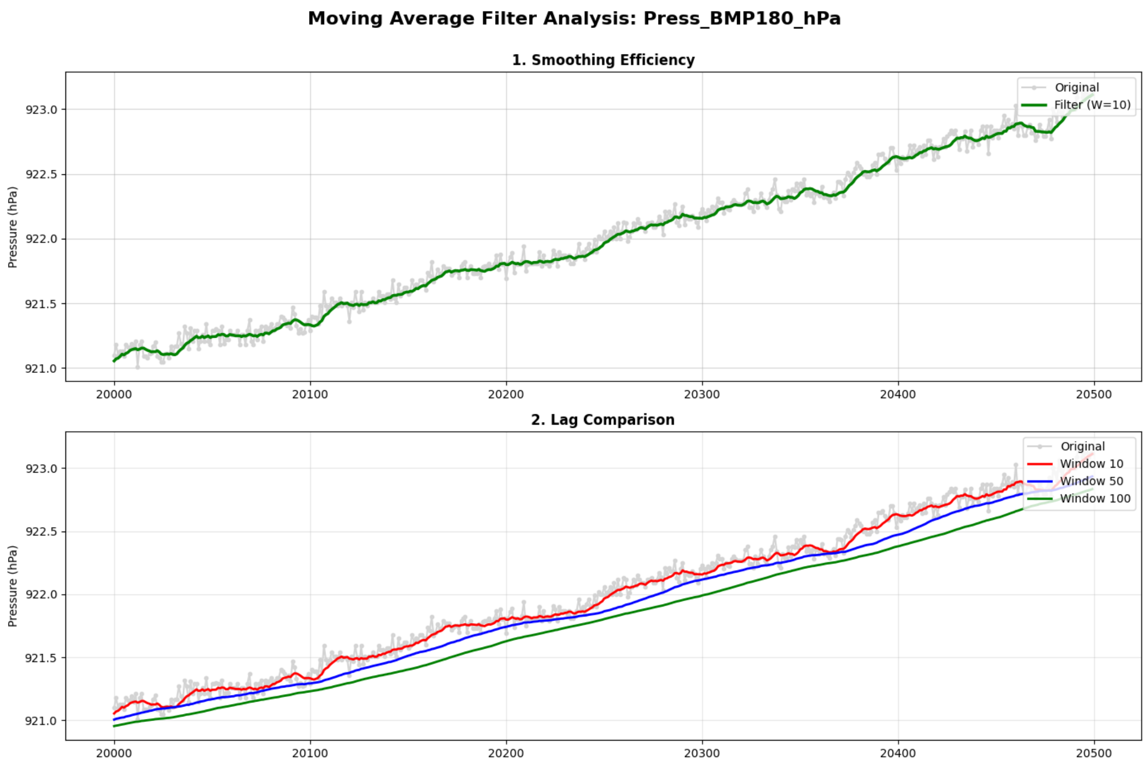

Figure 67: Comparison of digital filters - Pressure Sensor 2 (Moving Average)

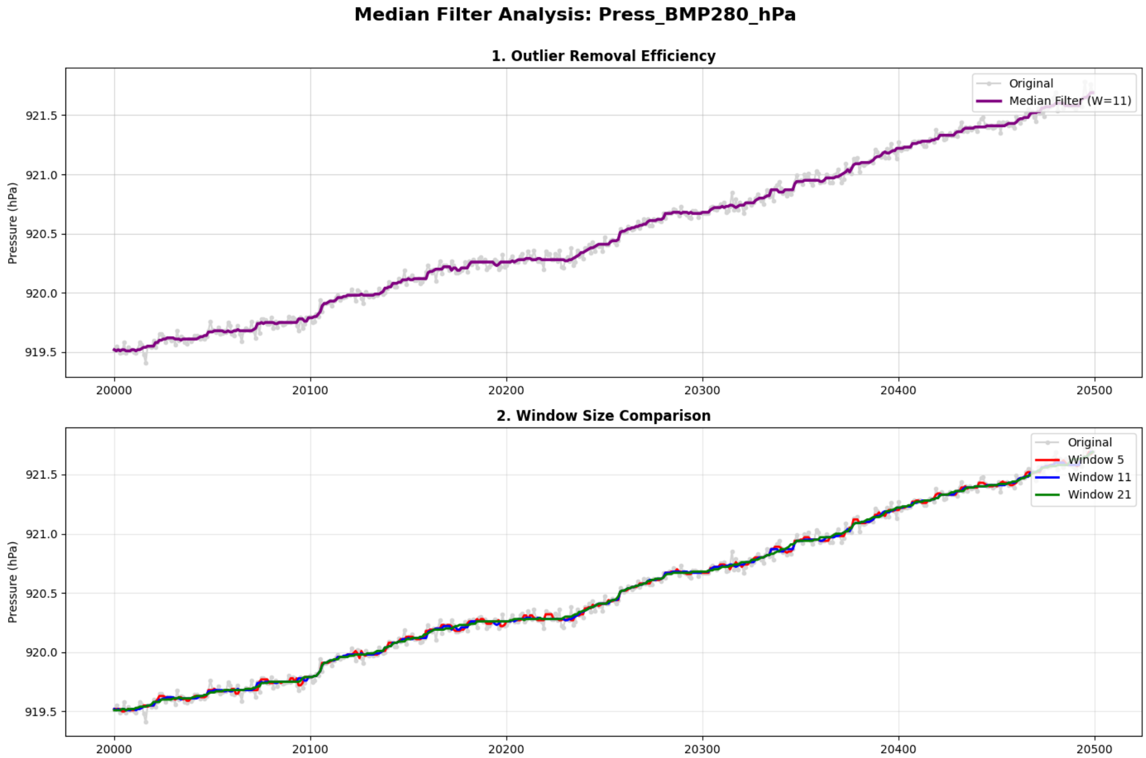

Figure 68: Comparison of digital filters - Pressure Sensor 1 (Median)

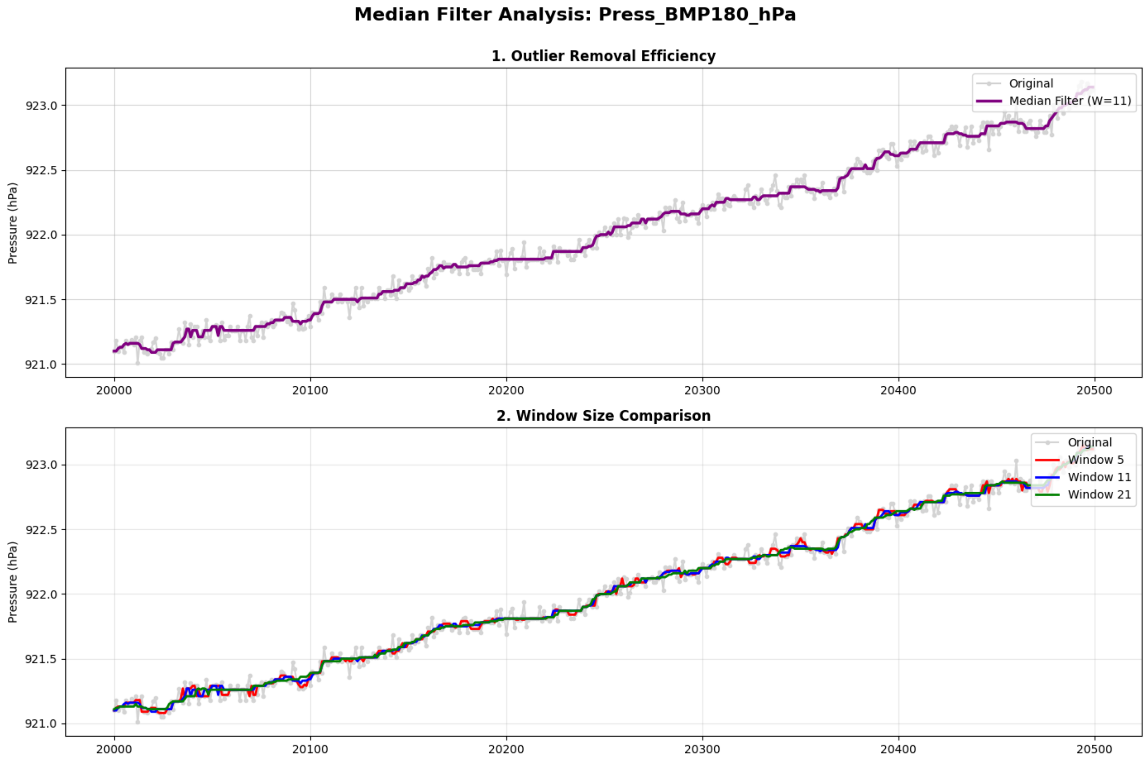

Figure 69: Comparison of digital filters - Pressure Sensor 2 (Median)

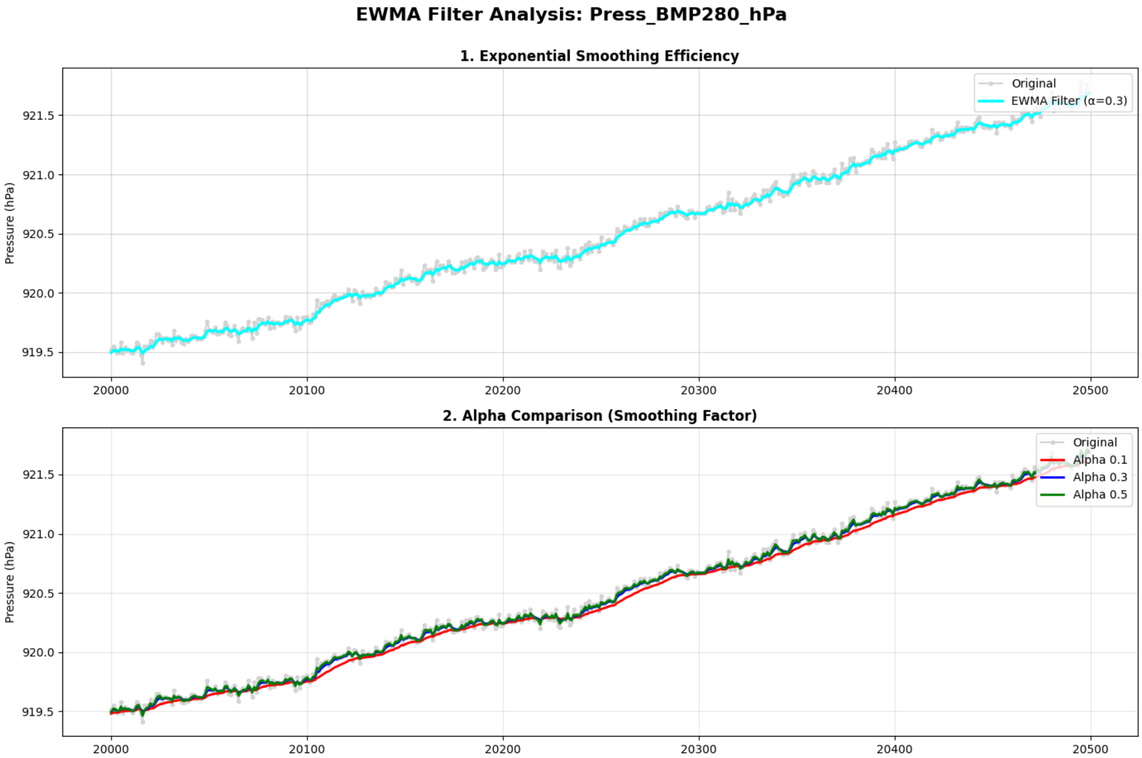

Figure 70: Comparison of digital filters - Pressure Sensor 1 (EWMA)

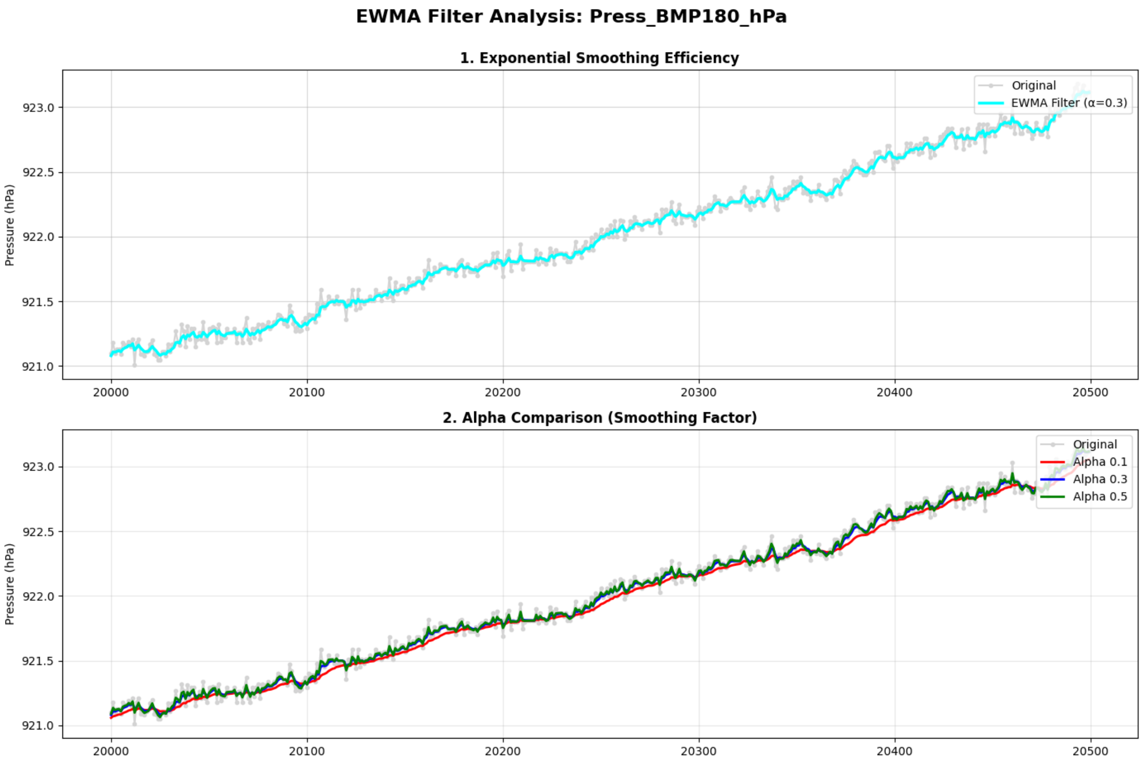

Figure 71: Comparison of digital filters - Pressure Sensor 2 (EWMA)

Digital filtering results:

- **Median Filter:** Best noise suppression without phase lag

- **EWMA:** Smooth transitions, preserves trends
- **Moving Average:** Trade-off between smoothing and responsiveness
- Complete panels with raw, filtered signal, zoom, and frequency spectrum

6.2 Notebook 13: Frequency Analysis (FFT)

Note: FFT analyses were integrated into the digital filter panels (Notebook 12), showing the power spectrum of each signal before and after filtering.

Spectral analysis results:

- **Primary peak:** 23-hour periodicity (diurnal cycle)
- Clear spectral signature of daily heating/cooling cycle
- Validation of the adequacy of the 30-second sampling interval

7 Public Dataset

The complete dataset was made publicly available on Kaggle:

- **Title:** "Vitória da Conquista Weather Data - September 2025"
- **URL:** `https://www.kaggle.com/datasets/jonassouza872/vitoria-da-conquista-weather-`
- **Records:** 82,430 measurements
- **Format:** CSV with 12 environmental variables
- **License:** Free, public access for download

8 Repository Structure

```
PolySense-Station-/
|-- main.py                         # 226 lines MicroPython
|-- requirements.txt                # Python dependencies
|-- lib/                            # MicroPython drivers (9 files)
|-- notebooks/                      # 13 analysis notebooks
|   |-- 01_exploratory_analysis.ipynb
|   |-- 02_correlation_analysis.ipynb
|   |-- 03_missing_data.ipynb
|   |-- 04_sensor_validation.ipynb
|   |-- 05_temporal_analysis.ipynb
|   |-- 06_time_series_decomposition.ipynb
|   |-- 07_anomaly_detection.ipynb
|   |-- 08_decision_tree_regression.ipynb
|   |-- 09_gmm_clustering.ipynb
|   |-- 10_kmeans_clustering.ipynb
|   |-- 11_lstm_prediction.ipynb
```

```
|   |-- 12_digital_filters.ipynb
|   |-- 13_fft_analysis.ipynb
|-- data/raw/                         # 4 CSV files (4.8 MB total)
|   |-- climate_clusters_gmm.csv     # 82,430 records
|   |-- validation_data_cleaned_BRT.csv
|   |-- validation_and_Measured_Data_cleaned_BRT_.csv
|   |-- inmet_weather_station_data_sep_2025_utc.csv
|-- images/                           # 111 generated visualizations
|   |-- data_analysis/                # EDA, correlation, validation
|   |-- machine_learning/             # Clustering, anomalies, regression
|   |-- signal_processing/            # Filter comparisons, FFT
|-- PCB/                              # Custom PCB design
|-- Schematic/                        # Hardware schematics
|-- README.md                         # Complete documentation
```

9 Conclusions

9.1 Main Achievements

1. **Functional collection system:** 100% success over 30 days (82,430 records at 30s interval)
2. **Effective sensor redundancy:** 7 temperature sensors, 2 humidity, 2 pressure with ¿99.9% completeness
3. **Validation against reference:** Mean bias ¡ 2°C compared to official INMET station
4. **Exceptional correlation:** r ¿ 0.98 between redundant sensors
5. **Anomaly detection:** 3.2% outliers successfully isolated with Isolation Forest
6. **Accurate prediction:** MAE ¡ 1°C for 1-hour ahead temperature prediction with LSTM
7. **Climate clustering:** 4 distinct regimes identified and validated (GMM and KMeans)
8. **Spectral analysis:** Clear 24h periodicity confirmed via FFT
9. **Public dataset:** 82,430 measurements available on Kaggle for the scientific community

9.2 Study Limitations

- Data collection limited to 30 days (September 2025)
- Lack of direct solar radiation measurement
- Temporal granularity of 30 seconds (does not capture very fast events)
- External factors not considered (detailed cloud cover, local wind)
- Validation limited to one season

9.3 Future Work

1. Expand collection to 12 months (complete seasonal analysis)
2. Add solar radiation sensor (pyranometer)
3. Implement anemometer for wind speed and direction
4. Develop PCB version 2 with identified improvements
5. Integrate data transmission via LoRaWAN
6. Implement edge computing with real-time LSTM predictions
7. Comparative analysis between multiple stations at different altitudes
8. Develop REST API for real-time data access

10 References

- **Microcontroller:** Raspberry Pi Pico (RP2040), MicroPython Documentation
- **Reference station:** INMET - National Institute of Meteorology
- **Dataset:** `https://www.kaggle.com/datasets/jonassouza872/vitoria-da-conquista-weath`
- **Period:** 08/31/2025 - 09/30/2025 (82,430 records)
- **Python libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, TensorFlow, SciPy
- **ML techniques:** Isolation Forest, Decision Tree, KMeans, GMM, LSTM
- **Signal processing:** FFT, Digital filters (Moving average, Median, EWMA)