Detecting Trends in Water Temperature in Estuaries of the United States

Sofia Catalan, Michael Mickelson, Owen O'Connor, and Eryn Wheeler

Introduction

- **Goal:** Analyze long term water monitoring data to detect and characterize trends in water temperature for U.S. estuaries
- **Variables of interest** (for trend detection) are water temperature, air temperature, and water temperature with influence of air temperature removed
- Full data set used for initial input, but final analysis broken down by individual estuaries

Dataset

- 16GB uncompressed, 1.9GB compressed
- historical data with temporal and spatial coverage across the U.S. to study and monitor estuaries
- Features: monitors dedicated to track specific data, such as:
  - water quality, nutrient, meteorological monitors

Image source: https://cdmo.baruch.sc.ed

Water Quality Monitors			
Water Temperature	Salinity	Dissolved Oxygen	pH
Turbidity	Specific Conductivity	Chlorophyll-a (optional)	

Nutrient Monitors			
Nitrate (NO_3-)	Nitrate (NO_2-)	Ammonium (NH_4^+)	Orthophosphate (PO_4^{3-})
Total nitrogen (TN)	Total phosphorus (TP)	Silicate (SiO_2)	

Meteorological Monitors			
Air temperature	Relative humidity	Barometric pressure	Wind speed
Wind direction	Precipitation	Solar radiation	

Data Preprocessing

- Hadoop MapReduce with Java
- Down-sampled air temperature and water temperature data to get monthly average values
- **Mapper:** Took in full dataset, filtered relevant rows and prepared a key-value pair based on month
- **Reducer:** Aggregated our readings for each key and calculate the average

Mapper Input

StationCode	isSWMP	dateTimeStamp	waterTemperature
acebbwq	P	3/3/1995 18:00	14.2

Mapper Output (Key-Value Pair)

Key: acebbwq-199503, Value: 14

Reducer Input

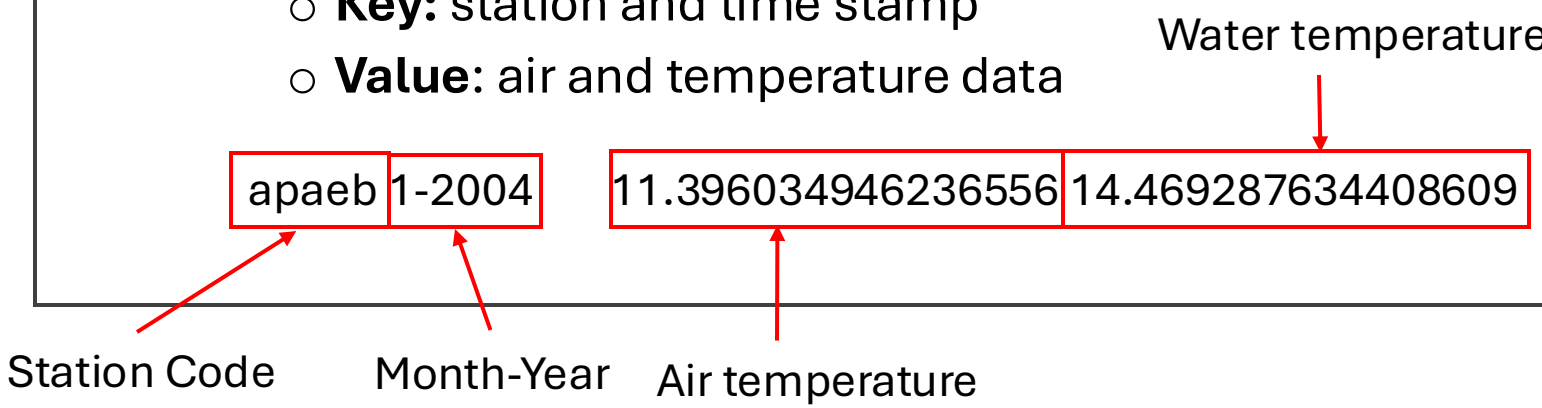
Key: acebbwq-199503
Values: [14, 14, 15]

Reducer Output

Key: acebbwq-199503, Value: 14

Composite Join

- Hadoop MapReduce with Java
- Joined two separate datasets for **air** and **water** temperature
- Join on **two foreign keys**: station and time stamp
- Map-only job
  - **Key:** station and time stamp
  - **Value**: air and temperature data



Mapper

```
public static class CompositeMapper extends MapReduceBase implements Mapper<Text, TupleWritable, Text, Text> {
        private Text outputKey =  new Text();
        private Text outputValue = new Text();

        public void map(Text key, TupleWritable value, OutputCollector<Text, Text> output, Reporter reporter)
                        throws IOException {

            // stationCode and dateTimeStamp
            outputKey.set(key.toString().trim() + "\t");

            // Obtain air and water data, removing extra spaces
            String airInfo = ((Text) value.get(0)).toString().trim();
            String waterInfo = ((Text) value.get(1)).toString().trim();

            outputValue.set(airInfo + " " + waterInfo);

            output.collect(outputKey, outputValue);
        }
}
```

Driver snippet

```
// input
format      conf.setInputFormat(CompositeInputFormat.
class);

// specify join type (e.g., inner join)
// the input format for the data sources
(KeyValueTextInputFormat)
// corresponding paths (airPath, waterPath).
conf.set("mapreduce.join.expr",
         CompositeInputFormat.compose(joinType,
         KeyValueTextInputFormat.class, airPath,
waterPath));
```

Linear Regression

- Computed linear regression residuals
  - Represent the water temperature minus the impacts of the air temperature
- Apache Spark using the ML library

```
LinearRegression regression = new LinearRegression();

LinearRegressionModel model = regression.fit(vectorized);

Dataset<Row> residuals = model.summary().residuals();
```

Seasonal Mann-Kendall Test

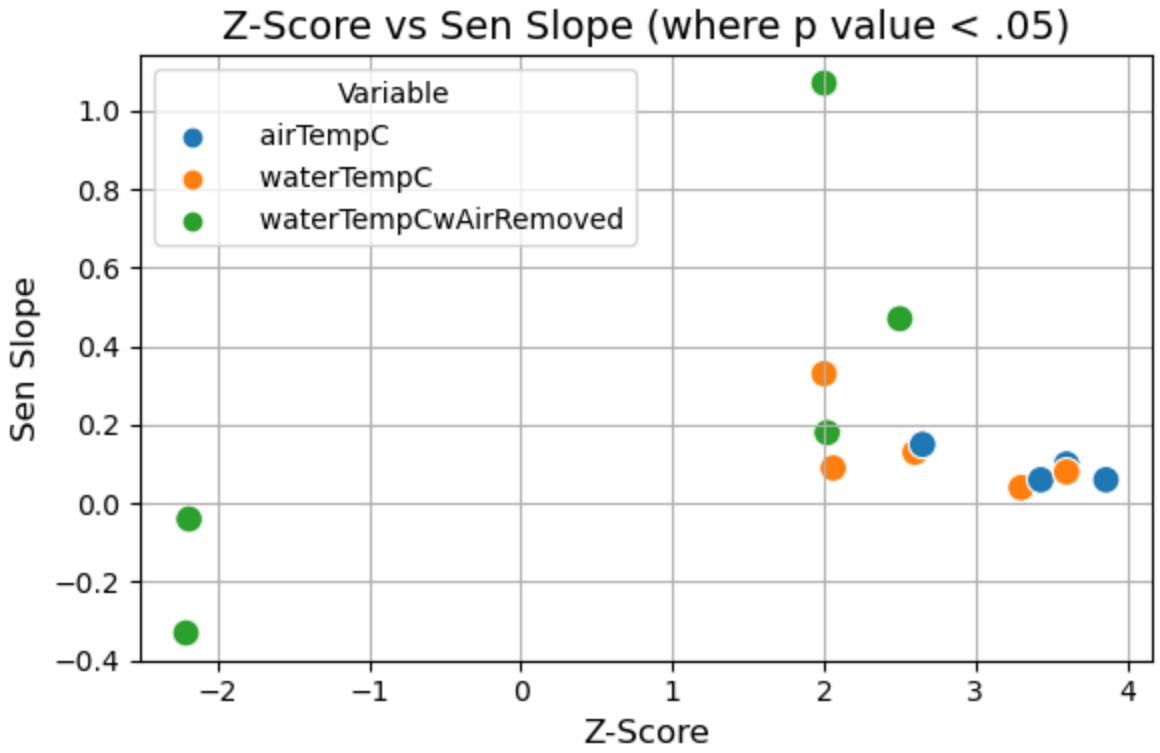
- Calculated for all three variables of interest for each estuary
  - Identifies and characterizes long-term monotonic trends if any
  - Water, air, and isolated water temperatures
- Compares historical data from similar seasons
  - e.g., winter to winter, summer to summer. In our case, we compared values from the same months but from different years.

Seasonal Mann-Kendall Test

- Mann-Kendall Tests
  - Compare data points to all previous datapoints and assigns a –1,+1 or 0. Sums all these to get an S score.
  - Tau normalizes this value between –1 and 1.
  - Z score calculated using the S score and the variance of the data.
  - If a significant trend is detected, a Sen slope estimate can give an estimate of the slope of the trend. It takes the median of all the slopes between all the points.
- Seasonal Mann-Kendall Test
  - Compare only values from like seasons.
  - S score is calculated for each season and then summed.
  - Variance calculated with specific formula and for each season, and then summed together.

Analysis

- SMK test was used to determine stations with statistically significant results
- 85% of those stations had an increase in water temperatures
- Some stations required more data to be collected before the test could be considered significant



Conclusion

- Data collected from across the United States
- Distilled and aggregated the water and air temperatures
- Performed the SMK Test, removing statistically insignificant results
- For estuaries with stronger trends detected, the water and air temperature trends were dominantly positive.

Next Steps & Future Directions

- Determine better methods for determining the trends
- Utilize stations near each other to supplement missing data
  - Air temperature missing from large portions of the dataset
- Tweak hyperparameters for linear regression
  - Improve the accuracy of the residuals
- Remove less data during aggregation step, focusing on more granular timeframes

Questions?