EMPLOYING SOFTWARE ENGINEERING PRINCIPLES TO ENHANCE MANAGEMENT OF CLIMATOLOGICAL DATASETS FOR CORAL REEF ANALYSIS

Mark Jenne[1], Alex Zimmerman[2], Hasan Kurban[1], Claudia Johnson[2], M.M. Dalkilic[1]

***Abstract*—The challenges presented by data to scientific inquiry and hypothesis testing in an oceanographic setting are not new problems. Indeed, the challenges are at least a century old. The problems are not with the data itself, but rather with the attention to the management of the "data ecology" in the information systems. Data needs to be accessible as an input to scientific inquiry—a requirement that goes far beyond simply centralizing the available data. Our research focuses on the development of a proof-of-concept system that properly handles an information ecology. The power of such a strong foundation is then demonstrated in two ways: (1) through our data driven hypothesis generation system, built and employed for analysis of the relationship between coral disease and temperature in the Caribbean; (2) through programmatic search for patterns and anti-patterns that verify, falsify, or demonstrate no discernible relationship for a set of variants on a particular temperature–disease hypothesis.**

I. MOTIVATION

Many oceanographic data repositories have come online in the last few decades. Some repositories are large oceanographic datasets (World Ocean Database (WOD) [1] and World Ocean Atlas (WOA) [2]), while others have more specific content (ReefBase Coral Bleaching GIS [3] and Global Coral Disease Database [4]). Data stored in these repositories, particularly the WOD and WOA, are vast and invaluable. Making the data accessible as a proprietary product, however, is not sufficient for driving large-scale analyses needed to understand the effects of climate change on coral reefs.

The focus of our research is to make climate and coral data available for scientific inquiry in a data-driven approach. We recognize that significant effort and resources are required to ensure longevity of data in an information system. Our proof-of-concept software for data governance and robust management of the data ecology is developed with this in mind. We demonstrate the power of such an approach by building an algorithm for a high-level analysis of coral disease in relation to ocean temperature in the Caribbean on top of our information ecology framework that uses not thousands, nor hundreds of thousands, but millions of data points. Together, the components of our system allow for programmatic search of the coral disease and temperature space to form testable hypotheses—a methodology referred to as data-driven hypothesis generation. It is these data-driven approaches that allows us to perform large-scale analyses needed to address the questions looming large for coral reefs under climate change.

The robustness of our information ecology management system is further demonstrated through a reciprocal approach where the complete data space is searched for patterns and anti-patterns that verify, falsify, or demonstrate no discernible relationship for a temperature–disease hypothesis formulated as an conditional rule-set and fed back into the system. Thus our system acts as a framework for both hypothesis generation and testing.

II. METHOD

The software systems behind the two primary components in this research are: (1) the information ecology framework; (2) the algorithms for the analysis of the relationship between coral disease and temperature. We include only an overview of these components here, but provide formal notation for the algorithms behind the data-driven hypothesis generation and hypothesis pattern search. The configuration for the system, or experimental parameterization, behind the data-driven hypothesis extraction and testing is then described.

Corresponding author: M Jenne, mjenne@indiana.edu [1]School of Informatics and Computing, Indiana University [2]Department of Geological Sciences, Indiana University

Rather than following the traditional computational science approach of ushering all of the data to the algorithm and building out the algorithm to incorporate the elements of transformation, management, and processing of the data, we isolate the non-research related procedures and push them to the data. Together, this set of procedures and the controlling software around them, form the backbone of what we are calling our information ecology framework. This system accomplishes two major goals: (1) provisioning a robust data manager with all necessary extraction, transformation, and loading procedures; (2) isolating the processes involved in the scientific inquiry of the algorithm development.

With the data ecology framework in place, the algorithmic components focus on the search of the data space and extraction of isolated trends in the data for hypothesis testing (Alg. 1) and assessment of those trends against a particular hypothesis or hypotheses formulated as a logical rule-set (Alg. 2). As a first pass, our approach leverages a naive quasi-clustering technique for establishing spatial bounds for the geographic extent of coral disease outbreaks and is followed by a search of the large temperature data space for local representative ocean temperature data. The new spatially and temporally associated data are then used to produce visual analytic tools for expert analysis from which individual, testable hypotheses are extracted for further consideration. Hypotheses formed from these trends can then be plugged back in as rule-sets guiding pattern search through the data. This methodology employs data-driven hypothesis generation by trading in the necessity for explicit, testable claims to drive experimental setup in favor of general pattern search within the data pertaining to the relationship between coral disease and temperature.

Reviewing individual geographic locations reveals a potential causal relationship between thermal stress anomalies followed by disease outbreaks. Or stated succinctly: for a particular geographic location, where the annual average sea surface temperature exceeds the regional average during the time period of 1970 to 2009 by some threshold (antecedent), we expect to see an increase in coral disease at that location in the following year (consequent). Forming a similar metric to those presented by Selig, et al [5] for testing temperature–coral disease trends, we translate this hypothesis into a logical rule-set and programmatically search all geographic locations for all instances that support and refute the hypothesis. The consequent of this rule is whether there is an increase [verifies], a reduction [falsifies], or no change [inconclusive] in

Algorithm 1 Coral Disease Temperature Analysis

```
INPUT data $\{\Delta_1, \Delta_2\}$, config $\Phi$
OUTPUT Temperature–Disease Timelines $A_1, \ldots, A_n \in \mathsf{A}$
%% assume that each $A_i$ is a tuple $(lat, lon, \mathsf{D} \in \Delta_1, \mathsf{T} \in \Delta_2, \mathsf{Y})$
%% where each $Y_i \in \mathsf{Y}$ is $(y, \mathsf{D}_i \subset \mathsf{D}, \mathsf{T}_i \subset \mathsf{T}, \mathsf{C})$
%% $y$ is the year, $\mathsf{D}_i$ is the subset of coral diseases at this location for year $y$, $\mathsf{T}_i$ is the subset of temperatures at this location for year $y$
%% $\mathsf{C}$ is the list of corals affected by disease at this location
%% cluster disease instances
for $\mathbf{x} \in \Delta_1$ do
    flag $\leftarrow$ false
    for $A_i \in \mathsf{A}$ do
        if $\mathbf{x}$.distance$(A_i.lat, A_i.lon) \leq \Phi.radius$ then
            $A_i.\mathsf{D} \leftarrow A_i.\mathsf{D} \cup \mathbf{x}$
            $flag \leftarrow true$
        end if
    end for
    if !flag then
        $\mathsf{A} \leftarrow \mathsf{A} \cup A(\mathbf{x})$
    end if
end for
%% associate temperature data with disease clusters
for $A_i \in \mathsf{A}$ do
    for $Y_j \in A_i.\mathsf{Y}$ do
        $itr \leftarrow 0$
        while $\|resultSet\| = 0 \wedge itr < \Phi.maxItr$ do
            $results \leftarrow Query(\Delta_2, A_i, Y_j.y, \Phi.rad + (\Phi.rad * (itr/2)))$
        end while
        if $\|resultSet\| = 0$ then
            $A_i.\mathsf{T} \leftarrow Query(\Delta_2, Y_j.y)$
        else
            $A_i.\mathsf{T} \leftarrow resultSet$
        end if
    end for
end for
%% process temperature-disease timelines
for $A_i \in \mathsf{A}$ do
    for $Y_j \in A_i.\mathsf{Y}$ do
        $Y_j.\mathsf{D}_j \leftarrow A_i.\mathsf{D}.Where(x => x.year = Y_j.y)$
        $Y_j.\mathsf{T}_j \leftarrow A_i.\mathsf{T}.Where(x => x.year = Y_j.y)$
        $Y_j.\mathsf{C} \leftarrow Y_j.\mathsf{D}_j.Select(x.genusSpecies => x)$
    end for
end for
```

coral disease instances following a true evaluation of the rule antecedent. Four variants of the rule with the threshold ranging from 1°C to 4°C are tested. Results are presented in Table 1.

Algorithm 2 Temperature–Disease Hypothesis Search

```
**INPUT** Temperature–disease timelines $A_1, \ldots, A_n \in \mathsf{A}$, Hypothesis Rule $R$
**OUTPUT** Pattern sets that verify $\mathsf{V}$, falsify $\mathsf{F}$, and demonstrate no discernible relationship $\mathsf{I}$
%% assume that each $A_i$ and all of the constituent variables have the same meaning as those presented in Algorithm 1.
**for** $A_i \in \mathsf{A}$ **do**
  **for** $Y_j \in A_i.\mathsf{Y}$ **do**
    **if** $AntecedentMatch(R.Ant, Y_j.\mathsf{T}_j)$ **then**
      **if** $ConsequentVerify(R.Con, Y_j.\mathsf{D}_j)$ **then**
        $\mathsf{V} \leftarrow \mathsf{V} \cup Y_j$
      **end if**
      **if** $ConsequentFalsify(R.Con, Y_j.\mathsf{D}_j)$ **then**
        $\mathsf{F} \leftarrow \mathsf{F} \cup Y_j$
      **end if**
      **if** $ConsequentInconclusive(R.Con, Y_j.\mathsf{D}_j)$ **then**
        $\mathsf{I} \leftarrow \mathsf{I} \cup Y_j$
      **end if**
    **end if**
  **end for**
**end for**
```

III. EVALUATION

The results presented here are in the context of both a big data problem and large-scale analyses. Making use of our data ecology framework, our algorithm for data-driven hypothesis generation regarding the temperature-coral disease relationship in the Caribbean was able to integrate and process the complete coral disease catalog presented by ReefBase and the complete ocean temperature data set hosted in the WOD. Grouping the 5,038 coral disease records into spatial clusters yielded 293 distinct geographic locations for analysis. At each location, respective temperature data subsets were selected from the more than 62 million data points available. The resulting coral disease and temperature sets were grouped together and visualized for extraction of testable hypotheses. We now address components of a compound hypothesis stating that a 2°C temperature rise and pH reduction of about 0.1 are more than sufficient to cause extensive stress and mortality to corals [6]. Our hypothesis regarding regional thermal stress anomalies preceding coral disease outbreaks was formed as a logical rule-set and fed into the system to see if the trends in the data verify or falsify the hypothesis. The results are found in TABLE I. It is informative that in the Caribbean, the data show verification/falsification counts are similar for ≤ 2 °C, but > 2 °C the hypothesis appears valid. As importantly, we observe that the data show relatively few instances of coral diseases at these temperatures, likely because these temperatures are rarely observed in the Caribbean, especially when all depths of the ocean are pooled together. This data-driven hypothesis generation technique suggests, for the high temperatures, an analysis of the shallow water temperatures separate from the deeper, cooler waters would be warranted, and would further test the hypothesis relating high temperatures to coral diseases.

TABLE I: Compound Hypothesis Analysis

TSA	Verify	Falsify	Inconclusive
$\geq$ 1°C	68	67	1625
$\geq$ 2°C	29	28	539
$\geq$ 3°C	11	3	126
$\geq$ 4°C	4	0	23

Cases that verify, falsify, or demonstrate no discernible relationship for the temperature–coral disease hypothesis for *thermal stress anomalies (TSA)* ranging from 1°C to 4°C.

This data-driven hypothesis generation and testing approach explores a proof-of-concept through a hypothesis rule-set and, as such, suffers from an incomplete picture of the biotic data. Incorporation of additional data *e.g.*, reef coverage, coral counts, coral mortality, would help the construction of a more complete model.

Here we have demonstrated the benefit that a robust information ecology management system lends to hypothesis generation and testing. Use of our system made the individual data sets involved easily accessible as input to our scientific inquiry, which allowed us to perform Caribbean-wide analyses in exploring the relationship between ocean temperature and coral disease. Proper data management in concert with these data-driven approaches will further allow us to perform large-scale analyses needed to address the questions looming large for coral reefs under climate change.

REFERENCES

[1] T. Boyer, J. A. O. Baranova, C. Coleman, H. Garcia, A. Grodsky, D. Johnson, R. Locarnini, A. Mishonov, T. O'Brien, C. Paver, J. Reagan, D. Seidov, I. Smolyar, and M. Zweng, "World ocean database," *NOAA Atlas NESDIS*, vol. 72, p. 209, 2013.

[2] R. A. Locarnini, A. V. Mishonov, J. I. Antono, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, C. R. Paver, J. R. Reagan, D. R. Johnson, M. Hamilton, and D. Seidov, "World ocean atlas 2013, volume 1: Temperature," *NOAA Atlas NESDIS*, vol. 73, p. 40, 2013.

[3] "Reefbase coral diseases." Accessed: 2016-06-15.

[4] "Global coral disease database." Accessed: 2016-06-15.

[5] E. R. Selig, C. Drew Harvell, J. F. Bruno, B. L. Willis, C. A. Page, K. S. Casey, and H. Sweatman, "Analyzing the relationship between ocean temperature anomalies and coral disease outbreaks at broad spatial scales," *Coral reefs and climate change: science and management*, pp. 111–128, 2006.

[6] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, *et al.*, "Coral reefs under rapid climate change and ocean acidification," *science*, vol. 318, no. 5857, pp. 1737–1742, 2007.

[7] S. Levitus, "Climatological atlas of the world ocean," *Eos, Transactions American Geophysical Union*, vol. 64, no. 49, pp. 962–963, 1983.

[8] T. H. Davenport and L. Prusak, *Information Ecology: Mastering the Information and Knowledge Environment*. Oxford University Press, 1st ed., 1997.

[9] A. de Geus, *The Living Company: Habits for Survival in a Turbulent Business Environment*. Harvard Business School Press, 1st ed., 1997.

[10] T. R. McClanahan, M. Ateweberhan, C. A. Muhando, J. Maina, and M. S. Mohammed, "Effects of climate and seawater temperature variation on coral bleaching and mortality," *Ecological Monographs*, vol. 77, no. 4, pp. 503–525, 2007.

[11] R. J. Jones, J. Bowyer, O. Hoegh-Guldberg, and L. L. Blackall, "Dynamics of a temperature-related coral disease outbreak," *Marine Ecology Progress Series*, vol. 281, pp. 63–77, 2004.

[12] J. W. Porter, P. Dustan, W. C. Jaap, K. L. Patterson, V. Kosmynin, O. W. Meier, M. E. Patterson, and M. Parsons, "Patterns of spread of coral disease in the florida keys," in *The Ecology and Etiology of Newly Emerging Marine Diseases*, pp. 1–24, Springer, 2001.

[13] W. L. Hürsch and C. V. Lopes, "Separation of concerns," 1995.

[14] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*. Pearson Education India, 1995.