## 1. Define workflow

To define a process or work-flow the ED must be able to specify a dataset or collection of datasets selected from a set of preloaded data. From the data the ED needs to define a geographical region or single point to study. The ED requires a high level of control of the temporal dimensions of the data. They will define the start and end dates and may need to specific that analysis takes place on recurring time-periods within the data (such as an afternoon in a specific season). The ED needs to specify which variables from the data to analyse and potentially assign a weight to each to define its relevant importance. Once the dataset and spatial and temporal subset has been defined and the dimensionality specified, the ED will select an analytic or data mining technique from a set of predefined (and described) methods. At any point in this process the ED may need to save, save as and or editing the defined work-flow.

## 2. Run workflow + monitor resources

Once the work-flow is defined the ED may need an estimate of how long the selected analytics will take. The workflow will then be submitted for processing and its progress will be monitored in terms of its status (ready, executing, staging, completed). The ED may also need to monitor the state of processing resources. During processing the should be able to stop, pause or cancel the workflow.

## 3. Output + visualise results

Following analysis results from the workflow shall be stored and accessible to the ED. Numerical and statistical summaries of the results will be generated by the ED using one of a range of predefined methods. These results would be stored and available for download in formats that can be further analysed or shared in spreadsheets. The ED may need to use simple visualisation of results (histograms, pie charts, line charts). Visualisation tools should include the ability to represent results by geospatial mapping. The graphics generated will be downloaded as high-quality images or vector graphics and used in reports. Within the interface the ED may need to compare two or more visualisations side-by-side to undertake a visual analysis. The ED also needs to share visualisation results with collaborators by providing a secure link to a webpage where the graphics can be viewed and download.

## 4. Manage design strategies

The design strategies specific to individual ED’s, it is necessary, therefore, to manage the predefined general design strategies used by the system. Some EDs will wish to add new strategies, removing unwanted strategies and editing existing ones. A standard unambiguous method for describing a strategy is required.

# Cycle 1 Requirements

The scope of the development cycle is to establish a basic workflow management system that allows set up and testing of some predefined simple workflows. This scope serves two goals; The first is to acquire working technical knowledge of how analysis using Hadoop Map Reduce and Spark can be run using Amazon’s Web Services (AWS) elastic map reduce (EMR) and simple storage services (S3). Secondly to start investigating the primary hypothesis of the thesis – that these technologies when applied to Colombia’s climate data can enable knowledge discovery that can ultimately support low-energy building construction.

To enable this a rudimentary desktop user interface is required (ultimately this should be web based). The UI must allow a user to select from one of a few predefined workflows. The UI must provide the ability to upload data sets and the Spark and Map Reduce JAR files required for the workflows. For this cycle the datasets and JAR files will be pre-processed on a local machine. Via the UI users should be able to start workflows, check their progress and have the option of cancelling a workflow. When a workflow completes users should be able to view results as simple graphs and (when possible) plotted on georeferenced maps.

The application will be developed with the AWS Java SDK and use JavaFx to develop the user interface. Test Driven Development will be used

Unit testing will be deployed for f

Mock

Integration testing for checking the UI and integration with the AWS API

## Infrastructure

1. The system must be based on AWS cloud infrastructure.
2. The infrastructure must support data processing and analytics with Hadoop, Apache Spark and Apache Mahout.
3. For the purposes of the current development cycle the system should be accessed via a desktop based user interface

## Workflow definition

For the purposes of the development cycle workflows will be predefined.

Using a predefined data set hosted on AWS S3:

1. The system should be able to process a k-means clustering with Apache Spark and Hadoop
2. The system should be able to process a Hadoop Map Reduce workflow
3. The system should be able to process a linear classification with Apache Spark and Hadoop

## Workflows

Control + monitoring:

The user must be able to start, stop and cancel running workflows.

The system should provide an interface for monitoring workflow status.

Data management:

The system should provide a method for uploading datasets

The system should provide a method for uploading JAR files with data processing (MapReduce and Spark) functions written in Java

## Output + visualise results

The system must provide the ability to visualise results with simple graphs.

The system must provide a method to visualise georeferenced results on a map.

## Data required:

For this cycle monthly data from two sources will be used

3. Monthly gridded data from CRU-TS 3.0 Climate Database for 1901-2008. Variables available: cloud cover, diurnal temperature range, frost day frequency, precipitation, daily mean temperature, monthly average daily minimum temperature, monthly average daily maximum temperature, vapour pressure, wet day frequency (Cgiar-csi.org., 2012).

4. Daily gridded data from CCMP Version-2.0 for 1988-2014. Variables available: wind speed and wind direction (Wentz *et al.*, 2015).

## The Design:

### Architecture overview focus elements for the cycle show in black - future elements show in grey

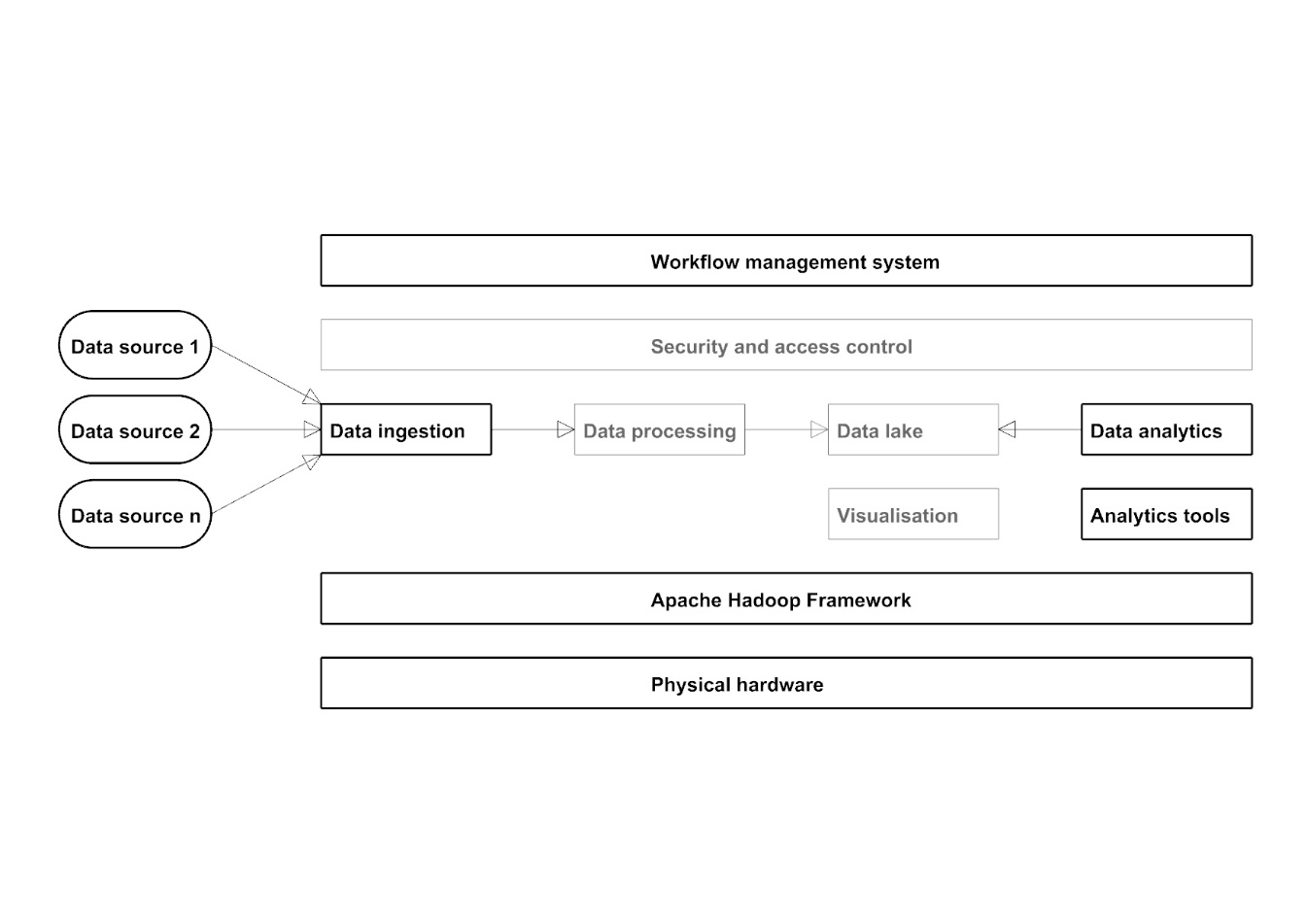
**

Figure 4. Proposed architecture overview based on Hadoop framework (extended from: Lopes, Palmer and O’Sullivan, 2017).

### System components overview

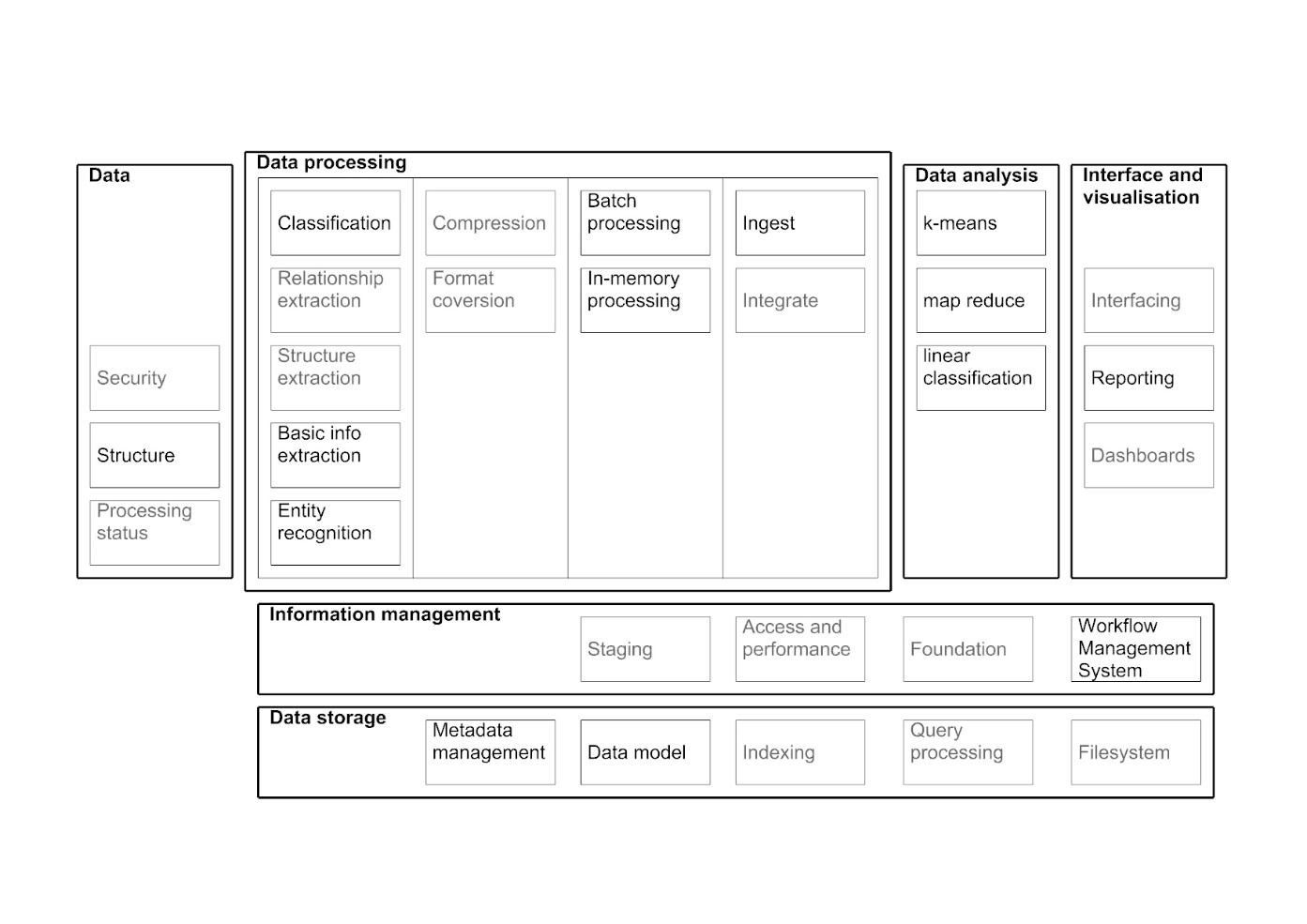
**

Figure 5. System components overview based on big data reference architecture (extended from: Avci Salma, Tekinerdogan and Athanasiadis, 2017).

### Architecture for WMS

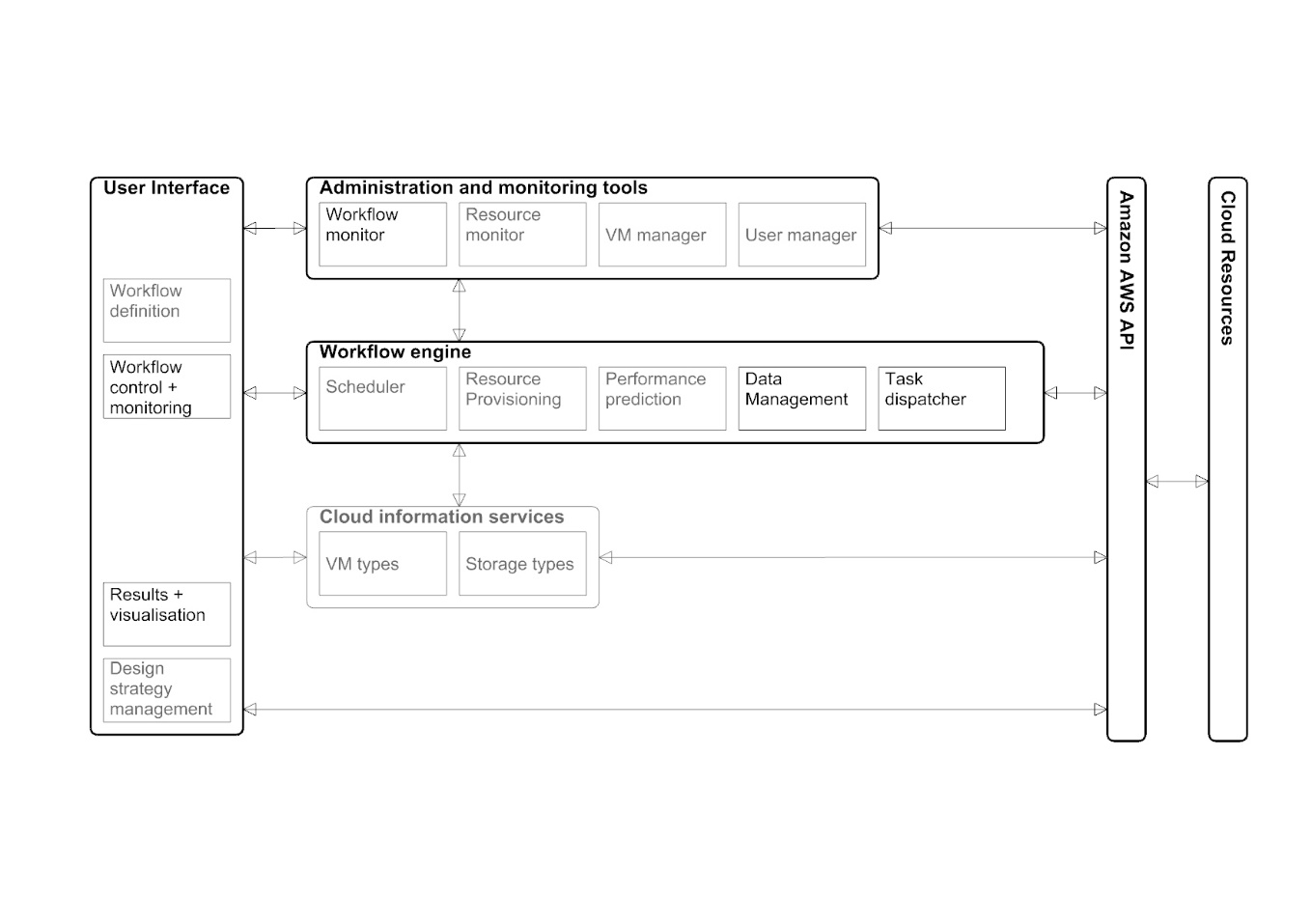


Figure 6. Architecture for the WMS (adapted from: Rodriguez and Buyya, 2017).

Cycle class diagram

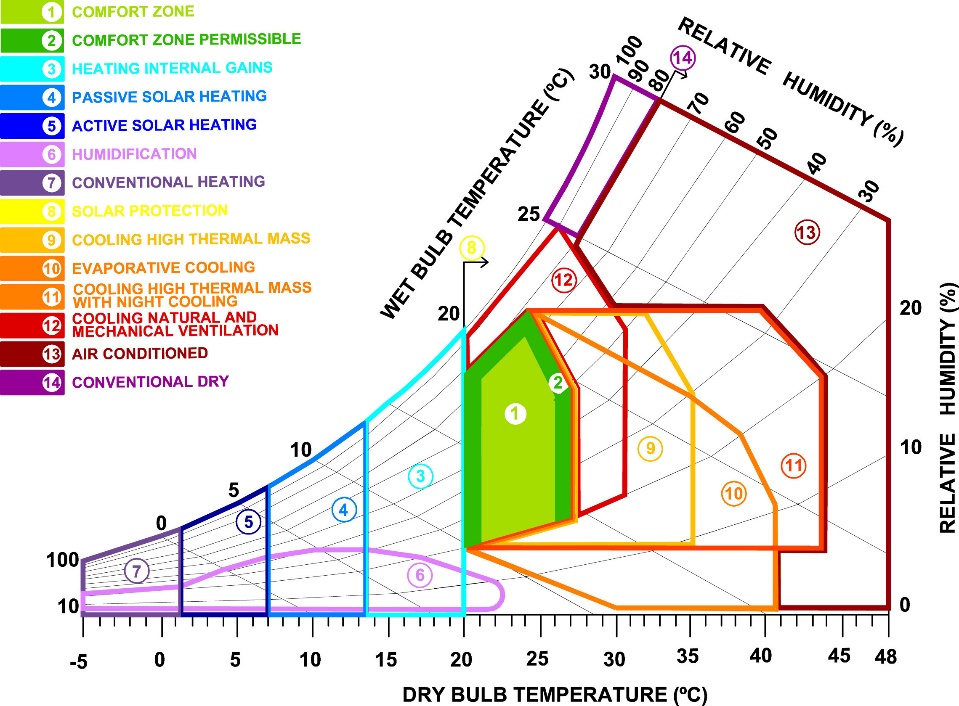


Figure 1. Manzano-Agugliaro *et al.* (2015) adapted version of the psychrometric chart.

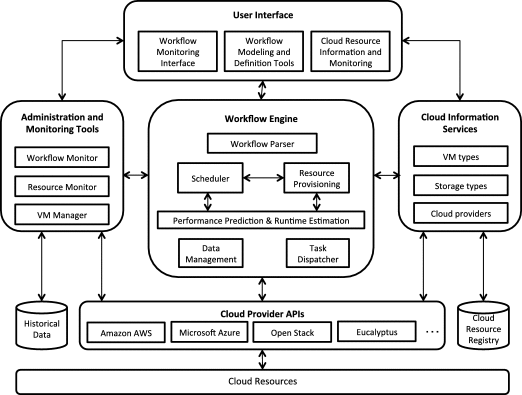


Figure 2. Reference architecture of a WMS (Rodriguez and Buyya, 2017).

### Class diagram for the WMS

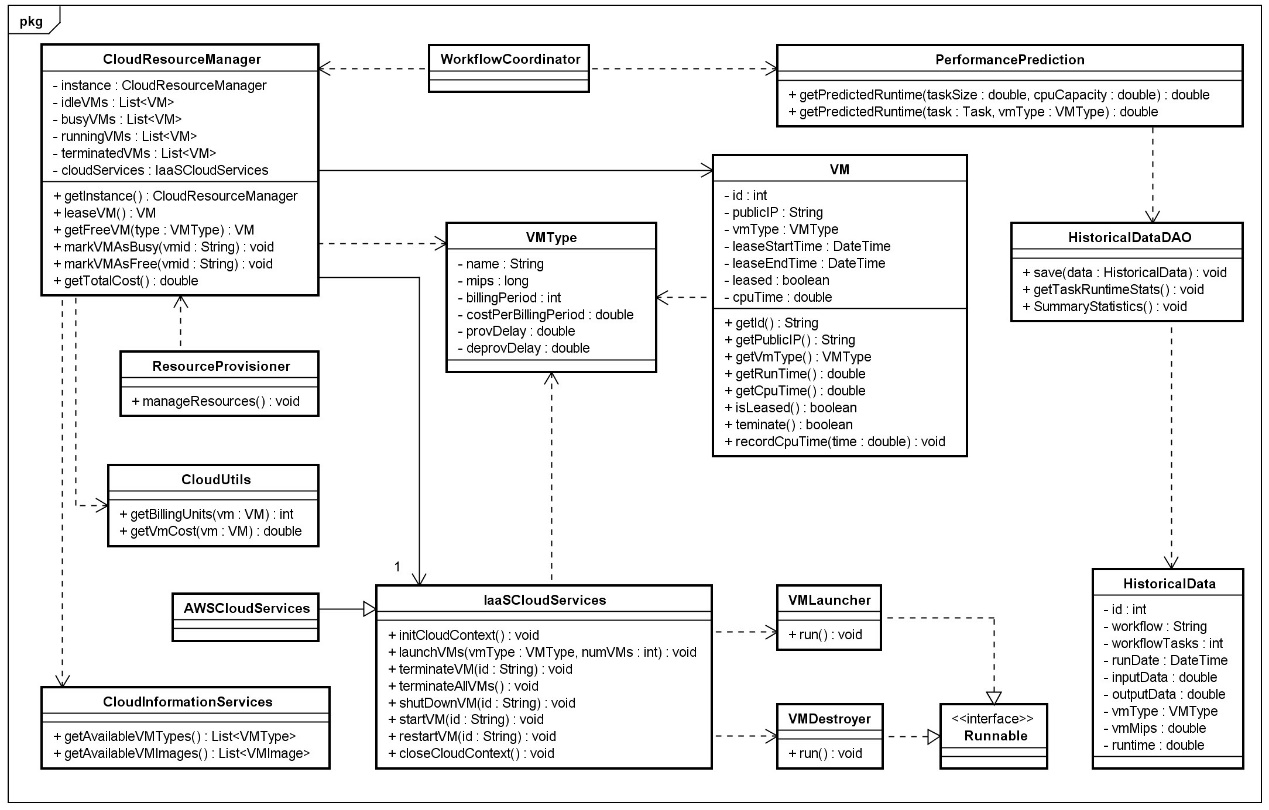


Figure 7. Classes for the workflow management system (adapted from: Rodriguez and Buyya, 2017).

### Sequence Diagrams for General use case analysis

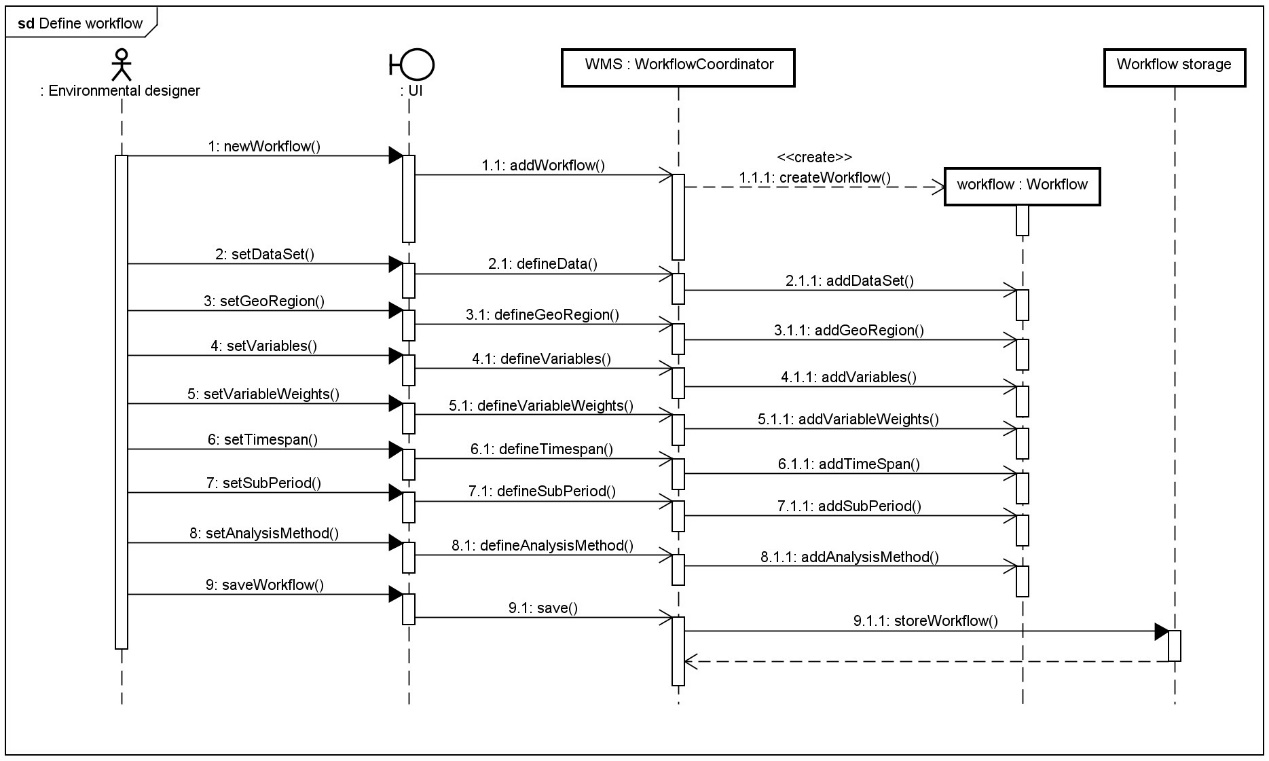


Figure 8. Sequence diagram for defining a workflow.

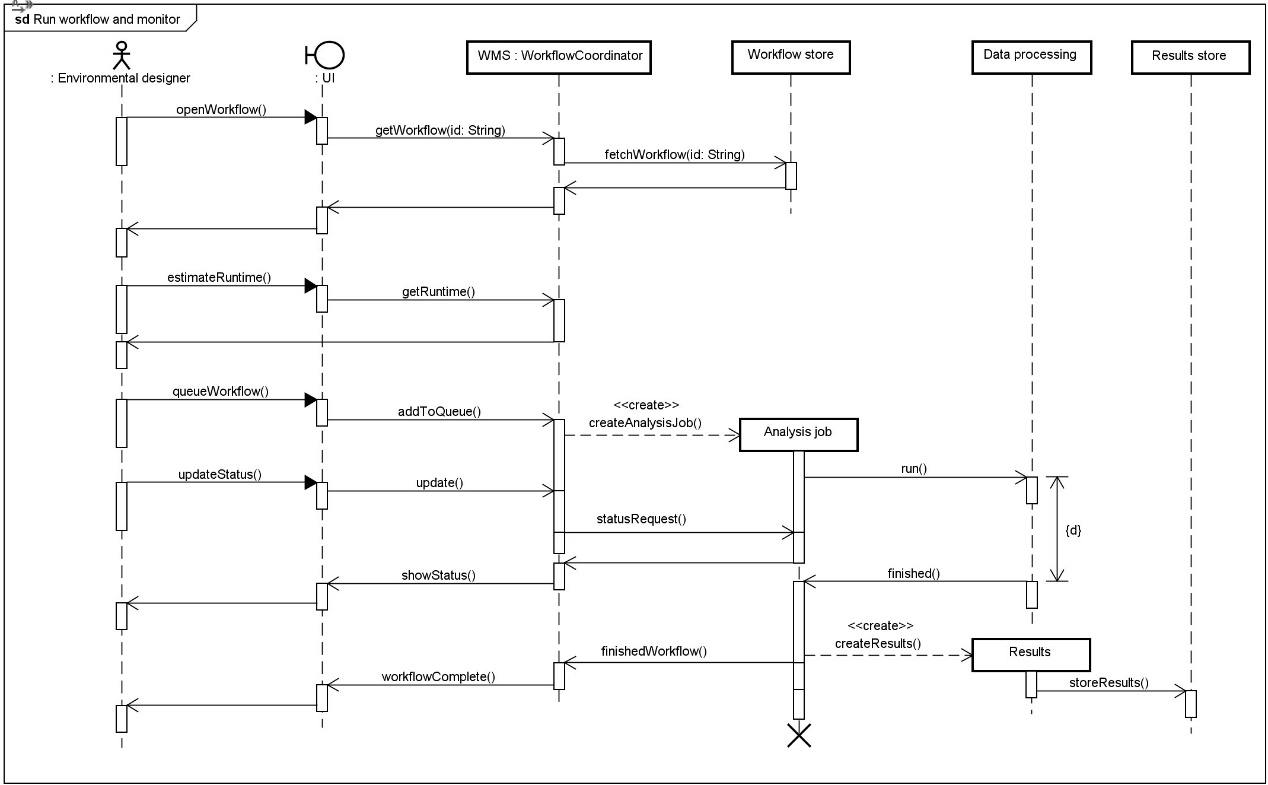


Figure 9. Sequence diagram for running and monitoring a workflow.

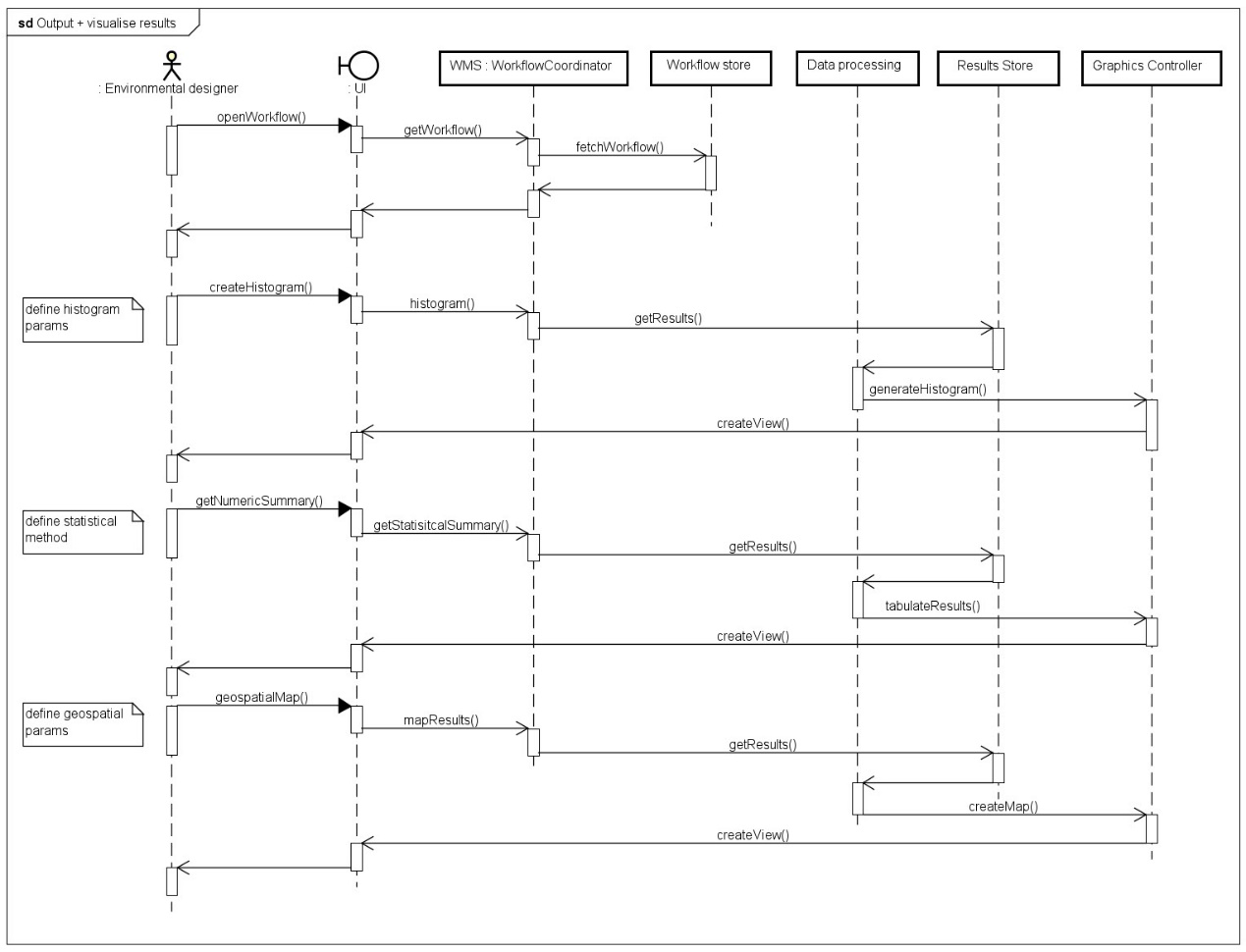


Figure 10. Sequence diagram for visualisation and output of results.

### Evaluation

The system is intended to produce recommendations for environmental design strategies based on the analysis of a set of weather data. The hypothesis to be tested is that design strategies can be linked with patterns discovered in weather data at various spatiotemporal scales and with different subsets of variables. It is proposed that this can be tested using a big data architecture that enables data analytics over large sets of weather data. Colombia is chosen for a test case as the tropical climate combined with the extreme topography and proximity to oceans.

Products of the system will be decomposed for evaluation; weather patterns will be assessed both quantitively and qualitatively and matching of recommended design strategies and the patterns evaluated qualitatively.

Several metrics exist to express the completeness and homogeneity of clusters discovered through analytics. Validity-measure or V-measure expresses how well both completeness and homogeneity are satisfied (Rosenberg and Hirschberg, 2007). Other metrics capture one or the other and include Purity, Entropy, Rand Index, misclassification index, f-measure, silhouette coefficient and cluster distortion of clusters can be calculated.

Delta-maps and self-organising maps can be quantitatively evaluated using average quantization errors and topographic errors (Liu, Weisberg and Mooers, 2006). Network metrics (Fountalis, Bracco and Dovrolis, 2014) such as link-maps represent edge strength between areas, strength of areas can be defined as the weighted sum of the number of links, network significance can be defined by s-core decomposition by progressively removing lower strength areas

For results of all analysis methods visual inspection of graphical output will play an important role in evaluation of the system. Including interactive graphical representations will allow dynamic exploration of spatiotemporal results. Use of correlation matrices, 2D scatter plots and plotting georeferenced zones on maps will amplify knowledge discovery and allow products of the artefact to be presented to domain experts. Qualitative evaluation of the artefact and the products (patterns and recommendations) will be undertaken by identifying very specific use cases and developing walkthroughs.



Figure 11. Sketched Interface design.

### Data structures

Document based noSQL storage will be used for initial prototypes, mongoDB and couchDB are potential candidates as they provide geospatial commands. Example documents for a weather station data and a weather record would look like:

|  |  |
| --- | --- |
| Weather station | Weather record |
| {  "\_id": ObjectId("5acffd17ff5df66734f6464b"),  "stationCode": 20502345,  "stationName": "el dorado airport",  "latitude": 10.991427,  "longitiude": -74.063284,  "elevation": 345  } | {  "\_id": ObjectId("7ff5df66734f6464b5acffd1"),  "stationCode": 20502345,  "time": ISODate("2005-08-31T14:40:00Z"),  "temperature": 24.5,  "relativeHumidity": 64.5,  "windDirection": 15,  "windSpeed": 2.4  } |

### Algorithms

The main algorithms for the proposed data analytics are SOMs, Delta-maps and clustering plus the methods for evaluating the results of these (discussed above).

**Statement of Deliverables:**   
Literature review

UML documentation of the system development process:

* Requirements list
* Use case diagrams
* Package diagrams
* Class, object and component diagrams
* Deployment diagram
* Sequence diagrams
* Test plans
* Test cases

Climate big data analytic system that enables knowledge-discovery.

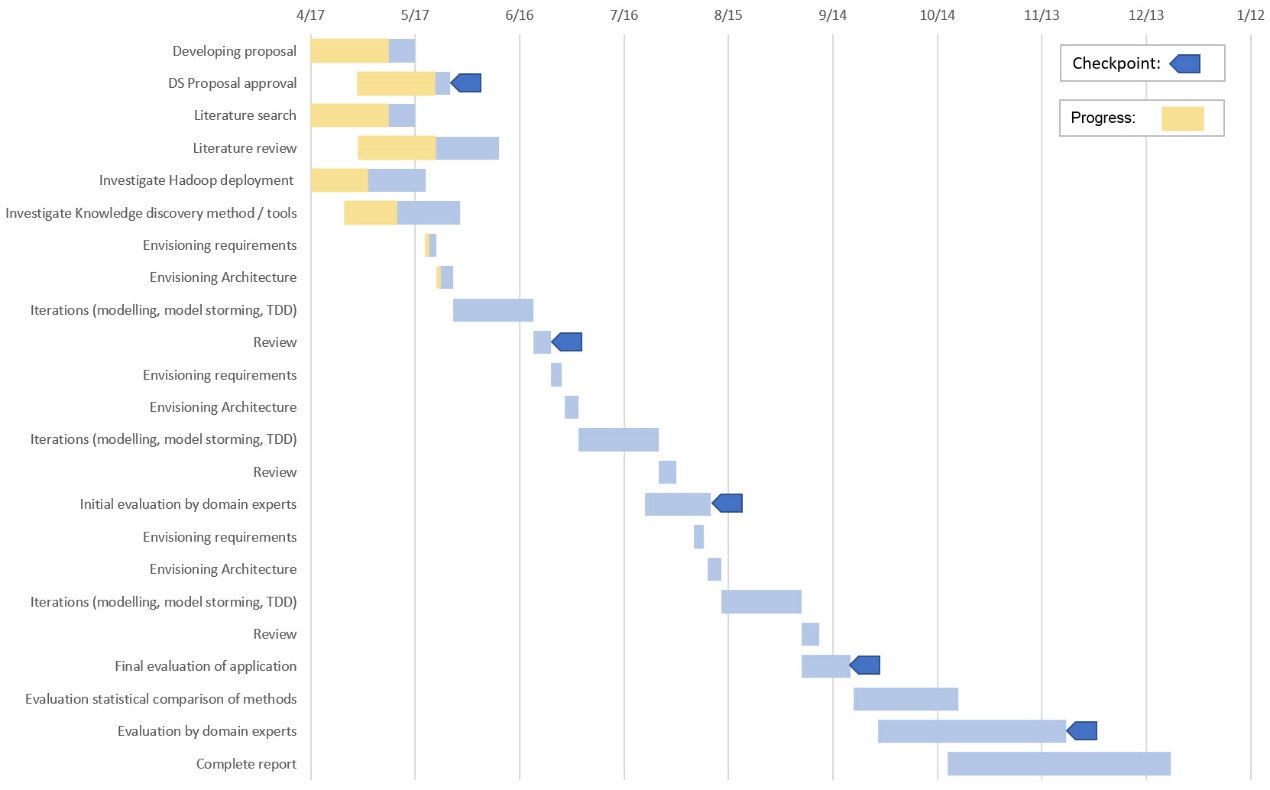
Example recommendations for construction strategies given a geographical location and the associated weather data.

Visualisation of results / output.

Set of statistical test results of outputs from the application.

Analysis of results and system by domain expert(s).

Generalised, transferable description of system architecture.

***Plan****:   
*

*Figure 12. Gantt chart of proposed activities*

**References**

Ambler, S. W. (2004) *The object primer : agile modeling-driven development with UML 2.0*. Cambridge University Press.

ASHRAE (2013) *2013 ASHRAE Handbook: Fundamentals*, *ASHRAE*. doi: 10.1163/ej.9789004155947.i-937.23.

Atluri, G., Karpatne, A. and Kumar, V. (2017) ‘Spatio-Temporal Data Mining: A Survey of Problems and Methods’, *ACM Comput. Surv*, 1(1). doi: 10.1145/nnnnnnn.nnnnnnn.

Avci Salma, C., Tekinerdogan, B. and Athanasiadis, I. N. (2017) ‘Chapter 4 – Domain-Driven Design of Big Data Systems Based on a Reference Architecture’, in *Software Architecture for Big Data and the Cloud*, pp. 49–68. doi: 10.1016/B978-0-12-805467-3.00004-1.

Begoli, E. and Horey, J. (2012) ‘Design principles for effective knowledge discovery from big data’, in *Proceedings of the 2012 Joint Working Conference on Software Architecture and 6th European Conference on Software Architecture, WICSA/ECSA 2012*, pp. 215–218. doi: 10.1109/WICSA-ECSA.212.32.

Bracco, A. *et al.* (2017) ‘Advancing climate science with knowledge-discovery through data mining’, *npj Climate and Atmospheric Science*, 1(1), p. 4. doi: 10.1038/s41612-017-0006-4.

Cgiar-csi.org. (2012) *CRU-TS v3.10.01 Historic Climate Database for GIS | CGIAR-CSI.* Available at: http://www.cgiar-csi.org/data/uea-cru-ts-v3-10-01-historic-climate-database (Accessed: 5 November 2017).

Chouksey, P. and Chauhan, A. S. (2017) ‘Weather Data Analytics using MapReduce and Spark’, *International Journal of Advanced Research in Computer and Communication Engineering*, 6(2). doi: 10.17148/IJARCCE.2017.6210.

Dagade, V. *et al.* (2015) ‘Big Data Weather Analytics Using Hadoop’, *International Journal of Emerging Technology in Computer Science & Electronics*, 14(2), pp. 976–1353. Available at: https://pdfs.semanticscholar.org/f2e4/918444be9b30f29132e93ce02d29ccf26eda.pdf (Accessed: 29 May 2018).

Dimoudi, A. and Tompa, C. (2008) ‘Energy and environmental indicators related to construction of office buildings’, *Resources, Conservation and Recycling*, 53(1–2), pp. 86–95. doi: 10.1016/j.resconrec.2008.09.008.

Faghmous, J. H. and Kumar, V. (2014) ‘Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities’, in Chu, W. (ed.) *Springer Berlin Heidelberg*, pp. 83–116. doi: 10.1007/978-3-642-40837-3\_3.

Fang, W. *et al.* (2014) ‘Meteorological data analysis using MapReduce.’, *The Scientific World Journal*, 2014, p. 646497. doi: 10.1155/2014/646497.

El Fazziki, A. *et al.* (2015) ‘A multi-agent framework for a hadoop based air quality decision support system’, in *CEUR Workshop Proceedings*, pp. 45–59.

Forsythe, N., Blenkinsop, S. and Fowler, H. J. (2015) ‘Exploring objective climate classification for the Himalayan arc and adjacent regions using gridded data sources’, *Earth System Dynamics*, 6(1), pp. 311–326. doi: 10.5194/esd-6-311-2015.

Fountalis, I., Bracco, A. and Dovrolis, C. (2014) ‘Spatio-temporal network analysis for studying climate patterns’, *Climate Dynamics*, 42(3–4), pp. 879–899. doi: 10.1007/s00382-013-1729-5.

Givoni, B. (1992) ‘Comfort, climate analysis and building design guidelines’, *Energy and Buildings*, 18, pp. 11–23. Available at: https://ac-els-cdn-com.liverpool.idm.oclc.org/037877889290047K/1-s2.0-037877889290047K-main.pdf?\_tid=67644909-7d5e-4718-b47f-80732d024251&acdnat=1527603110\_130b2a2e77cd97d48361202940532374 (Accessed: 29 May 2018).

Jain, H. and Jain, R. (2017) ‘Big data in weather forecasting: Applications and challenges’, in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, pp. 138–142. doi: 10.1109/ICBDACI.2017.8070824.

Jayanthi, D. and Sumathi, G. (2017) ‘Weather data analysis using spark — An in-memory computing framework’, in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, pp. 1–5. doi: 10.1109/IPACT.2017.8245142.

Jayaratne, M. *et al.* (2017) ‘Apache spark based distributed self-organizing map algorithm for sensor data analysis’, in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 8343–8349. doi: 10.1109/IECON.2017.8217465.

Karpatne, A. *et al.* (2017) ‘Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data’, *IEEE Transactions on Knowledge and Data Engineering*, 29(10), pp. 2318–2331. doi: 10.1109/TKDE.2017.2720168.

Liu, Y. and Weisberg, R. H. (2005) ‘Patterns of ocean current variability on the West Florida Shelf using the self-organizing map’, *Journal of Geophysical Research: Oceans*, 110(6), pp. 1–12. doi: 10.1029/2004JC002786.

Liu, Y. and Weisberg, R. H. (2011) ‘A Review of Self-Organizing Map Applications in Meteorology and Oceanography’, in Igadwa Mwasiagi, J. (ed.) *Self Organizing Maps - Applications and Novel Algorithm Design*. www.intechopen.com. doi: 10.5772/13146.

Liu, Y., Weisberg, R. H. and Mooers, C. N. K. (2006) ‘Performance evaluation of the self-organizing map for feature extraction’, *Journal of Geophysical Research: Oceans*, 111(5). doi: 10.1029/2005JC003117.

Lopes, D., Palmer, K. and O’Sullivan, F. (2017) ‘Chapter 10 – Big Data: A Practitioners Perspective’, in *Software Architecture for Big Data and the Cloud*, pp. 167–179. doi: 10.1016/B978-0-12-805467-3.00010-7.

Manzano-Agugliaro, F. *et al.* (2015) ‘Review of bioclimatic architecture strategies for achieving thermal comfort’, *Renewable and Sustainable Energy Reviews*. Pergamon, 49, pp. 736–755. doi: 10.1016/J.RSER.2015.04.095.

Milne, M., Liggett, R. and Benson, A. (2009) ‘Climate Consultant 4.0 develops design guidelines for each unique climate’, *American Solar Energy Society Meeting*. Available at: http://www.energy-design-tools.aud.ucla.edu/papers/ases09-milne.pdf (Accessed: 10 April 2018).

Netzel, P. *et al.* (2016) ‘On Using a Clustering Approach for Global Climate Classification’, *Journal of Climate*, 29(9), pp. 3387–3401. doi: 10.1175/JCLI-D-15-0640.1.

NOAA (no date) *Land-Based Station Data | National Centers for Environmental Information (NCEI) formerly known as National Climatic Data Center (NCDC)*. Available at: https://www.ncdc.noaa.gov/data-access/land-based-station-data (Accessed: 5 June 2018).

Olgyay, V. and Olgyay, A. (2015) *Design With Climate: Bioclimatic Approach to Architectural Regionalism*. Princeton University Press.

Omer, A. M. (2008) ‘Energy, environment and sustainable development’, *Renewable and Sustainable Energy Reviews*, 12, pp. 2265–2300. doi: 10.1016/j.rser.2007.05.001.

Rodenburg, B. and Maria Fiore, M. (2017) *Detecting Weather Twins using Apache Spark*, *LSDE: Large Scale Data Engineering 2017*. Available at: https://event.cwi.nl/lsde/2017/showcase\_n2.shtml (Accessed: 29 May 2018).

Rodriguez, M. A. and Buyya, R. (2017) ‘Chapter 18 – Scientific Workflow Management System for Clouds’, in *Software Architecture for Big Data and the Cloud*, pp. 367–387. doi: 10.1016/B978-0-12-805467-3.00018-1.

Rosenberg, A. and Hirschberg, J. (2007) ‘V-Measure: A conditional entropy-based external cluster evaluation measure’, pp. 410–420. Available at: http://www.aclweb.org/anthology/D07-1043 (Accessed: 6 June 2018).

Varghese, S. M. and Riyaz, P. . (2015) ‘Leveraging Map Reduce With Hadoop for Weather Data Analytics’, *IOSR Journal of Computer Engineering Ver. II*, 17(3), pp. 2278–661. doi: 10.9790/0661-17320612.

Wentz, F. . *et al.* (2015) *Remote Sensing Systems Cross-Calibrated Multi-Platform (CCMP) 6-hourly ocean vector wind analysis product on 0.25 deg grid, Version 2.0, [subset: CCMP V2.0 Level-3.5].* Santa Rosa, CA: Remote Sensing Systems. Available at: www.remss.com/measurements/ccmp (Accessed: 5 November 2017).