MACHINE LEARNING APPROACHES TO SAFEGUARDING CONTINUOUS WATER SUPPLY IN THE ARID AND SEMI-ARID LANDS OF NORTHERN KENYA

Fred Otieno, Timothy Nyota, Isaac Waweru Wambugu, Celia Cintas, Samuel Maina, William Ogallo & Aisha Walcott-Bryant
Future of Climate
IBM Research | Africa
Nairobi, Kenya

ABSTRACT

Arid and semi-arid lands (ASALs) in developing countries are heavily affected by the effects of global warming and climate change, leading to adverse climatic conditions such as drought and flooding. This proposal explores the problem of fresh-water access in northern Kenya and measures being taken to safeguard water access despite these harsh climatic changes. We present an integrated water management and decision-support platform, eMaji Manager, that we developed and deployed in five ASAL counties in northern Kenya to manage waterpoint access for people and livestock. We then propose innovative machine learning methods for understanding waterpoint usage and repair patterns for sensor-instrumented waterpoints (e.g., boreholes). We explore sub-sequence discriminatory models and recurrent neural networks to predict waterpoint failures, improve repair response times, and ultimately support continuous access to water.

1 INTRODUCTION

Achieving sustainable access to fresh water for low- and middle-income countries with an ever-expanding population requires multifaceted approaches, particularly in arid and semi-arid lands (ASALs). Resilience to climate variability and change in ASALs is greatly dependent on sustainable access to groundwater Taylor et al. (2013). Changes in climatic patterns are likely to increase the spatio-temporal variability in precipitation rates and surface water that makes it difficult to rely on these sources of water. Mitigating solutions such as rainwater harvesting, while important, may not be sustainable in ASALs which are often characterized by prolonged spells of drought followed by rainstorm flash floods that cause heavy loss of life and property Lin (1999).

Kenya has experienced a rise in the frequency of drought and floods in recent years. An analysis done in the country by Glew et al. (2010) indicates that climate change will likely increase drought incidences, temperatures, and water scarcity in northern Kenya which can exacerbate the poverty levels in a region that depends on rangeland grazing of livestock for their livelihood. Consequently, this region needs to adapt and manage its limited groundwater resources differently. However, to date, policy- and decision-makers lack the necessary tools for improved evidence-based decision support necessary for the planning and management of groundwater resources. Fortunately, recent advancements in technologies such as remote sensors and IoT (Internet of Things) devices combined with new deep learning methods can be leveraged to understand challenges and generate insights pertaining to access and use of groundwater.

We present a novel water supply management and decision-support platform using a case study of five counties in northern Kenya. We then propose two key areas for machine learning to provide new insights into waterpoint management and decision-support: (1) the use of discriminatory sub-sequence mining of waterpoint data to detect recurrent behaviours or unknown trends on deployed waterpoint sensors, and (2) waterpoint failure prediction using long short-term memory (LSTM) recurrent neural networks.

Figure 1: Sensors dashboard in eMaji Manager indicating sensor and site information.

2 eMaji Manager Platform: inventory, issue reporting and remote sensing for waterpoints

The eMaji Manager is an integrated mobile and web waterpoints management system that has been implemented across 5 counties in northern Kenya through the KRAPID (Kenya Resilient Arid Lands Project for Integrated Development Project) initiative mwa. One of the goals of KRAPID is to improve water access in the 5 northern counties of Kenya from 37% to more than 50%. Before the start of the project, county water officials relied on a community leader to inform them about an issue at a waterpoint such as a pump failure at a borehole. A team from the county would then conduct a site visit to confirm the issue and plan for repair. The repair work would then commence, and once completed and the waterpoint is functional, the issue was closed. This process would take weeks to months for a single waterpoint and could involve multiple waterpoints spread out in different geographical locations. As a result, many communities were often left without continuous access to water. To mitigate such technical water inaccessibility challenges, KRAPID aims at using technology to significantly reduce the time of reporting and repair of waterpoints.

The eMaji Manager ensures efficiency and transparency in the management and operation of waterpoints. The platform has the following key capabilities: ticketing and issue tracking for monitoring and repair of boreholes; descriptive analytics for reporting at the county and sub-county levels; and visualization and analytics for decision support. For each waterpoint, the platform captures information about the waterpoint name, geolocation, functional status, cost of water for livestock and people, and infrastructural details. Additionally, we collect time-series data from sensor-instrumented boreholes such as operating hours, yield, and operational status. The data from the sensors is visualized on a dashboard that is used by county government officials for monitoring. Thus far, 114 boreholes have been instrumented with sensors that have generated sensor data since 2018. Fig. 1 illustrates how the dashboard is used to monitor sensor uptime, sensor status, and site uptime. "Sensor Uptime" is the proportion of time that a given sensor is active and transmitting data: low, high and unknown. "Sensor Status" refers to the proportions of sensors on water points within an administrative region in any one of the following states: normal use, low use, no use, offline, repair, or seasonal disuse. "Site Uptime" refers to the proportion of boreholes categorized by unknown, low, medium and high, where uptime refers to the period of time that water is extracted. Fig. 1 also shows the geolocations of all instrumented waterpoints and the "Sensor Status".

3 Pattern mining and failure prediction for waterpoints

County and sub-county water stakeholders have to address several key questions to ensure there is continuous access to water for their residents. One of their major concerns is the ability to forecast potential areas where there may be water scarcity and to develop mitigation strategies to minimize community disruption to water access. Such challenges can be addressed by understanding functional patterns of instrumented waterpoints, detecting anomalies in their operation, and predicting failures. We propose to study "waterpoint behavior" using two different techniques. First, we pro-

pose an unsupervised method to analyze sequential patterns from waterpoint sensor data that lead to failures, detect recurring behaviours, and automatically generate rules and unknown trends associated with a waterpoint. Second, we propose to train time-series models that predict events such as failure, or high yield, that will trigger targeted interventions to the communities potentially affected.

We can find several examples in literature that use sequence mining techniques and recurrent networks to predict failure on sensor data and industrial processes. For example, Hajiaghayi & Vahedi (2019) present a solution based on recurrent neural networks to find sessions that are prone to code failure in applications that rely on telemetry data for system health monitoring. In the manufacturing space, Lim et al. (2017) propose a method to construct a predictive model of failure based on event sequences observed at the wire bonding process step. Kuzin & Borovicka (2016) shows different approaches that deal with early failure detection of sensor parts.

Given the sequential nature of waterpoint sensor data, we use Discriminatory Subsequence Mining to compare two subgroups of instrumented waterpoints (boreholes) and mine patterns that appear predominantly in one of the subgroups. For example, the set of waterpoint sensor data that contains a failure, compared to the rest of the waterpoints. Sequence mining provides the domain expert with a series of common sequential patterns of waterpoints and associated metrics including *coverage*- what percentage of one group contains a specific pattern, and *lift*- how likely a pattern is to appear in one group and not in the other one. For example, we can find a discriminatory sequence $\{low \rightarrow normal \rightarrow normal \rightarrow normal\}$, which refers to one month of low use followed by three months of normal use, with $coverage_{left} = 0.3388$, $coverage_{right} = 0.0892$, and $lift = 3.795$. The $coverage_{left}$ value means that 33% of the sensors that contain this sequence end with failure status. The $coverage_{right}$ value refers to a low frequency in the rest of the population data, thus, the pattern found is discriminatory regarding failure class in sensors. Furthermore, the lift value shows that this pattern is 3 times more likely to appear on sensors with failure events than functional waterpoint sensors. This insight would be valuable for developing automated waterpoint repair regime.

For failure prediction, our strategy is to map the input sequence of events from sensor_status to a fixed-sized vector representation using a recurrent neural network (RNN), and then feed the vector to a softmax layer for classification on the type of failure status: {offline, repair, seasonal disuse}. Given a series of status events $e = \{e_0, e_1, e_2, \cdots, e_T\}$ we first use a lookup layer to get the representation vector for each status event e_i for examples of status types. The output at the last moment h_T can be regarded as the representation of the full sequence of status events during the last n days prior to the failure. This has a fully connected layer followed by a softmax non-linear layer that predicts the probability of a particular type of failure for that time window of events. We can process, in real-time, new sensor data and use the trained model to infer what is the probability, $P(s_i)_x = fail$ for an online sensor, s_i, installed at a waterpoint, to fail in the next x days. When the probability exceeds a given threshold we can update the dashboard on eMaji Manager and automatically add the waterpoint to a "repair watchlist". In the big picture, this would contribute to the water supply crisis alarm that has a spectrum of (low - medium - intense) water crisis based on the projection of impacted waterpoints in a region. Example properties that would factor into triggering the alarm are waterpoint relevance for drought emergencies and the population served.

4 Discussion and Future work

Climate change and variability has had a direct impact on livelihoods in northern Kenya due to the increased uncertainty of drought seasons. This has necessitated the development of adaptation strategies to help cope with acute water shortages. We developed the eMaji Manager as a tool to assist in adapting to drought emergencies. eMaji Manager is an integrated waterpoint inventory, issues tracking, and decision support system for county officials. We propose to enhance the quality of data in eMaji by exploiting the waterpoint and sensor data to extract temporal patterns and predict non-functionality to shorten repair response time and provide an emergency alert system.

Further work needs to be done on cross-domain data linkage to integrate insights from other data sources such as weather, human migration patterns, and aquifers. Recent work in weather modeling, for instance in forecasting extreme events Rolnick et al. (2019), shows great promise for identifying weather patterns and predicting precipitation with a forecast horizon of between four to six weeks Hwang et al. (2019). We seek to augment weather, waterpoint and sensor data to enhance the understanding of temporal and spatial patterns on water availability for ASAL regions.

REFERENCES

Kenya rapid program. URL https://mwawater.org/programs/kenya-program-background/.

Glew, Louise, Hudson, Malcolm D., Osborne, and Patrick E. *Evaluating the effectiveness of community-based conservation in northern Kenya: A report to The Nature Conservancy*. Centre for Environmental Sciences, University of Southampton, https://rmportal.net/groups/cbnrm/cbnrm-literature-for-review-discussion/evaluating-the-effectiveness-of-community-based-conservation-in-northern-kenya-a-report-to-the-nature-conservancy, 2010.

Mahdi Hajiaghayi and Ehsan Vahedi. Code failure prediction and pattern extraction using lstm networks. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*, pp. 55–62. IEEE, 2019.

Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. Improving subseasonal forecasting in the western u.s. with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD âĂŹ19, pp. 2325âĂŞ2335, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016. doi: 10.1145/3292500.3330674. URL https://doi.org/10.1145/3292500.3330674.

Tomás Kuzin and Tomás Borovicka. Early failure detection for predictive maintenance of sensor parts. In *ITAT*, pp. 123–130, 2016.

Hwa Kyung Lim, Yongdai Kim, and Min-Kyoon Kim. Failure prediction using sequential pattern mining in the wire bonding process. *IEEE Transactions on Semiconductor Manufacturing*, 30(3): 285–292, 2017.

Xiao Lin. Flash floods in arid and semi-arid zones. *Technical documents in hydrology*, 1999.

David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *arXiv preprint arXiv:1906.05433*, 2019.

Richard G Taylor, Bridget Scanlon, Petra Döll, Matt Rodell, Rens Van Beek, Yoshihide Wada, Laurent Longuevergne, Marc Leblanc, James S Famiglietti, Mike Edmunds, et al. Ground water and climate change. *Nature climate change*, 3(4):322–329, 2013.

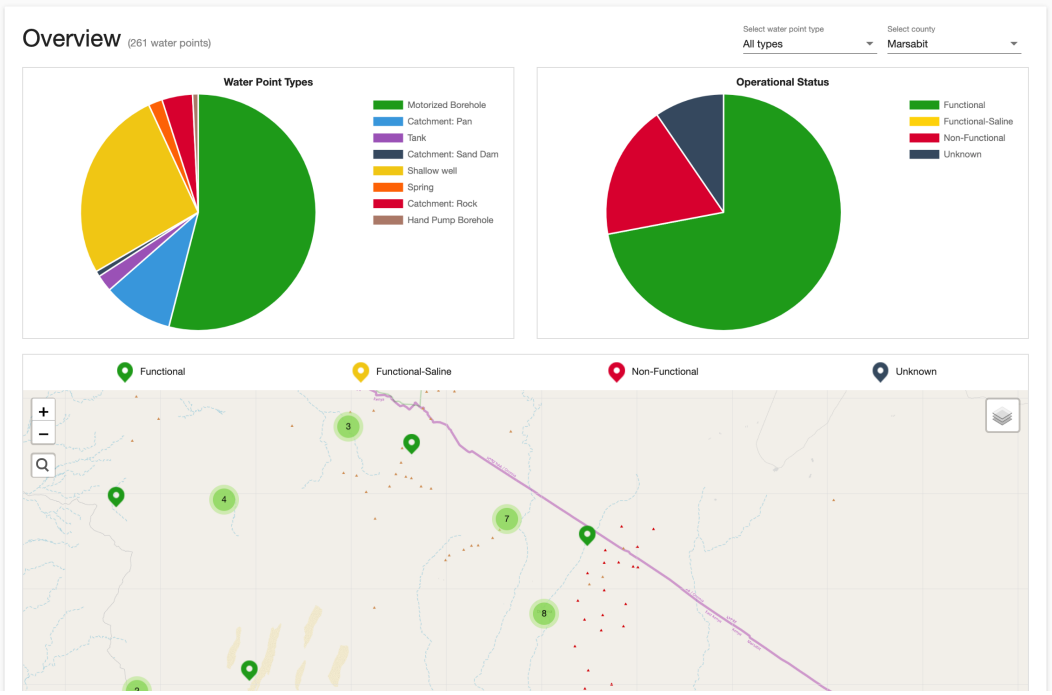


Figure 2: Overview dashboard for county official. The dashboard displays summary data about waterpoint types in the region, operational states summary of waterpoints and individual waterpoints on a map

A Appendix

A.1 Waterpoint inventory management system

Fig. 2 shows the overview page for waterpoint inventory management. This page helps the county officers to manage waterpoints in their regions. In addition to viewing a summary of waterpoint statuses, waterpoint types and locations of waterpoints on a map, the officers are able to view, edit and add waterpoint information including, infrastructure information on the waterpoints, sensor information, costs associated with the waterpoints, usages and even reports for a particular waterpoint as shown in Fig. 3

A.2 Decision support platform

Fig. 4 shows the overview page for the decision support functionality. An officer is given a snapshot of the status of waterpoints in their region. At a glance, they are able to know how many tickets are open, how many waterpoints are impacted, and the number of animals and people affected by the non-functional waterpoints. They also see which tickets are open and are able to act on them, or assign relevant officials to follow up on open tickets, thus ensuring responsive repairs. The officer can also drill down to see details of the particular non functional borehole. The system also generates recommendations for the officers on mitigation actions that they can take, as illustrated in Fig. 5 based on severity of reported issue.

Acknowledgments

We would like to acknowledge SweetSENSE Inc who instrumented waterpoints and made the sensor data available as a part of the Kenya RAPID project. We also would like to acknowledge Millennium Water Alliance and partners for their continued support on the Kenya RAPID project.

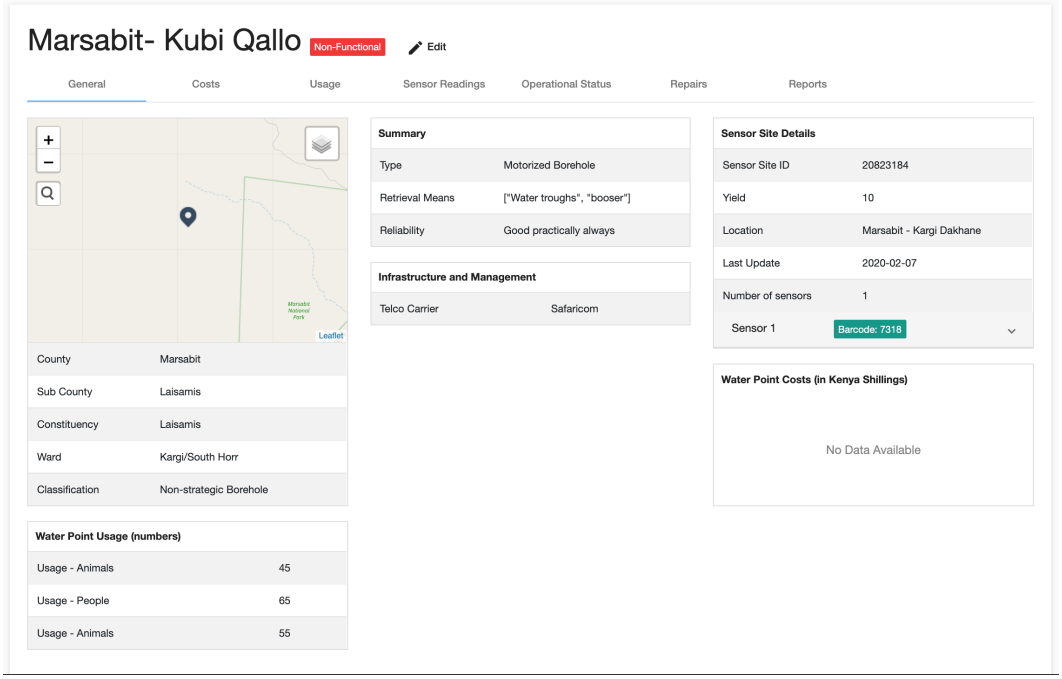


Figure 3: Dashboard showing all details of a particular waterpoint. The details include infrastructure information, usage, sensor, operational status and reports for the waterpoint

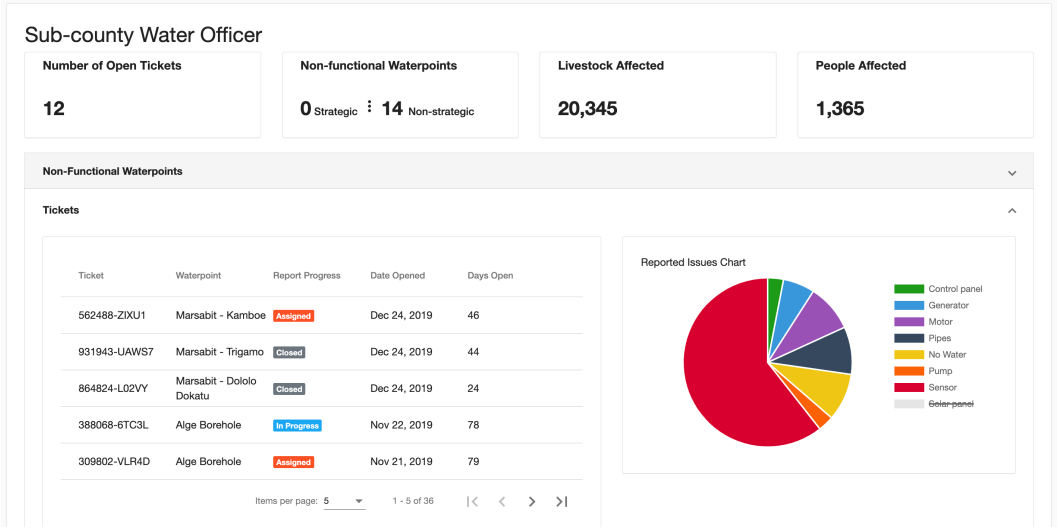


Figure 4: Decision support dashboard for county official. The dashboard displays summary data about waterpoints in the region, non-functional waterpoints, the tickets and issues requiring action, and a summary of reported issues.

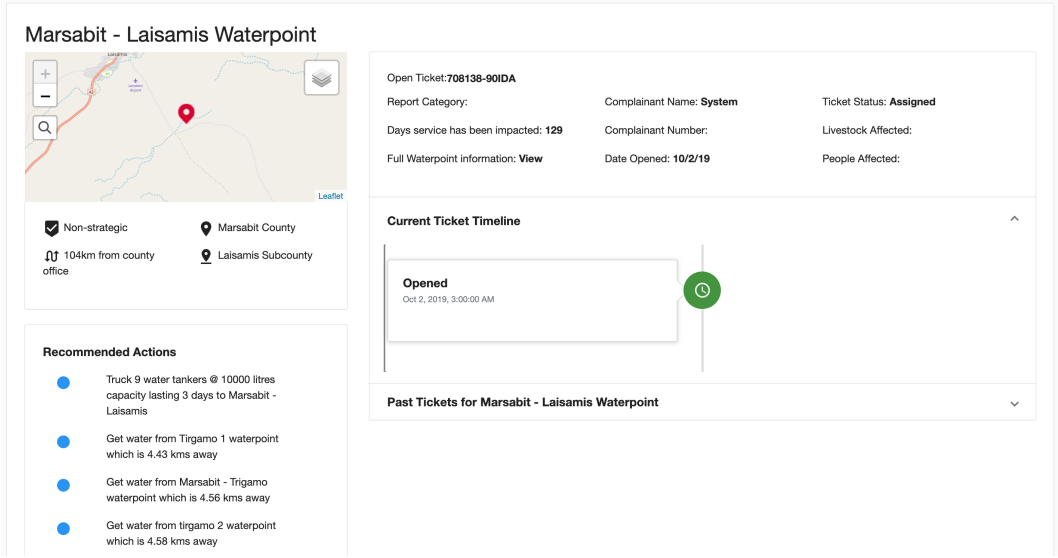


Figure 5: Decision support drill-down page for county official. The page displays waterpoint details of the non functional waterpoint, present and past ticket details to allow for tracking of the issues and their resolution, and a recommended mitigation actions based on severity of issue and context of waterpoint

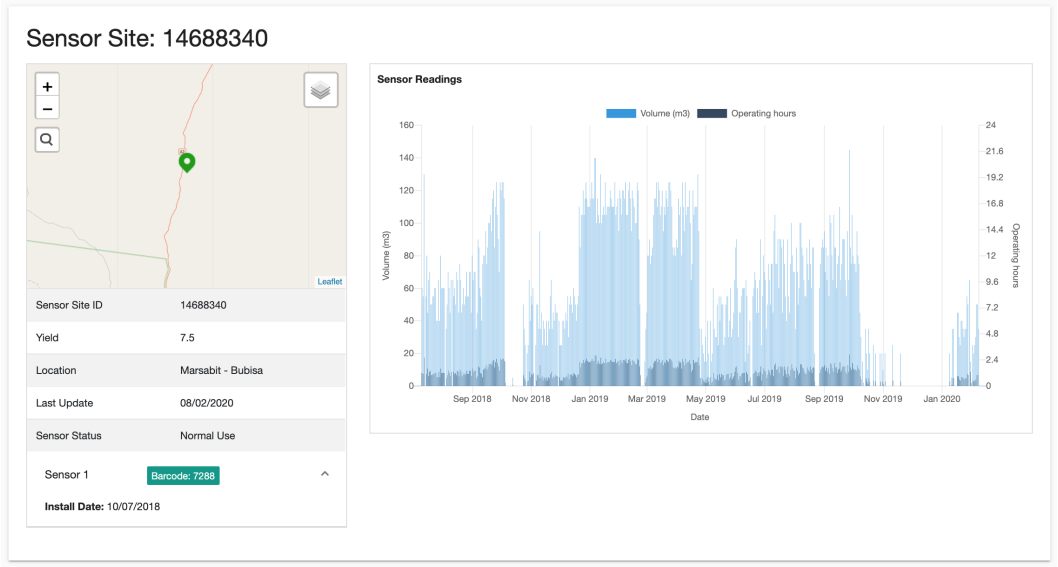
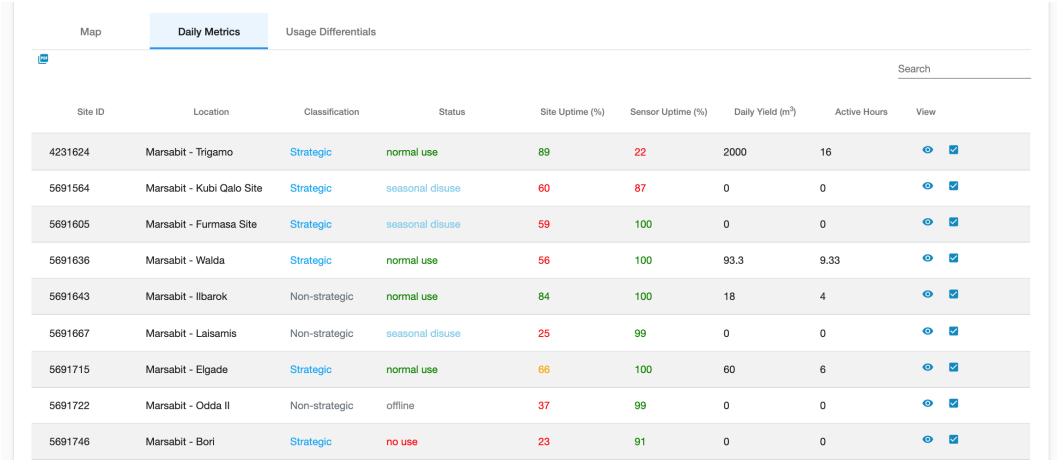


Figure 6: Dashboard showing sensor information: yield, location sensor status and a graph displaying daily sensor readings



Site ID	Location	Classification	Status	Site Uptime (%)	Sensor Uptime (%)	Daily Yield (m^3)	Active Hours	View
4231624	Marsabit - Trigamo	Strategic	normal use	89	22	2000	16	
5691564	Marsabit - Kubi Qalo Site	Strategic	seasonal disuse	60	87	0	0	
5691605	Marsabit - Furmasa Site	Strategic	seasonal disuse	59	100	0	0	
5691636	Marsabit - Walda	Strategic	normal use	56	100	93.3	9.33	
5691643	Marsabit - Ilbarok	Non-strategic	normal use	84	100	18	4	
5691667	Marsabit - Laisamis	Non-strategic	seasonal disuse	25	99	0	0	
5691715	Marsabit - Elgade	Strategic	normal use	66	100	60	6	
5691722	Marsabit - Odda II	Non-strategic	offline	37	99	0	0	
5691746	Marsabit - Bori	Strategic	no use	23	91	0	0	

Figure 7: Table showing daily data received from remote sensors including sensor id, sensor classification, status, up-time, yield and active hours.