An Automated Workflow for Satellite-based Monitoring of Field Flooding

Xinyi Wang[1], Wan-Chun Liao[1],

Kirk Klausmeyer[2], Nathaniel Rindlaub[2], Cody Carroll[1,3]

[1] Data Institute, University of San Francisco

[2] The Nature Conservancy

[3] Department of Mathematics and Statistics, University of San Francisco

Abstract

Every year, over one billion birds migrate along the Pacific Flyway and travel through California. Many of these birds need wetlands for food and rest to support their journey, but over 95% of the historical wetlands in the Central Valley have been drained and developed. The Nature Conservancy and partners launched the BirdReturns program to pay farmers to flood their fields to create a makeshift habitat to support migratory wetland birds on their journeys.

To ensure farmers flood their fields for the full duration of their contracts, experiments were carried out to estimate the extent and duration of flooding on the enrolled fields using free images captured by satellites. As the program scales up, more efficient tools are needed to ingest the satellite data, generate flooding extent estimates, and send weekly reports to the shareholders.

Based on the previous experiments, we created a cost-effective and automated data pipeline with Python scripts in a GitHub repository. We incorporated the following techniques: Application Programming Interface (API) data acquisition, image computation, GitHub Action for scheduled workflow, an interactive web application for report visualization, and a GMAIL/SMTP client for report sharing.

The BirdReturns program brings numerous ecological and social benefits to migratory birds, farmers, wetland managers, and communities across the Central Valley. The successful implementation of the data pipeline improves efficiency and reduces the workload of flood status monitoring for the BirdReturns program. Additionally, it provides a valuable approach for building an automated data pipeline for small-scale and start-up companies, where efficiency and costs are the primary concerns.

Introduction

Background

Every year, millions of ducks and shorebirds migrate along the Pacific Flyway and travel through California.[1] Many of these birds need wetlands for food and rest to support their journey. However, over 90% of Central Valley's historical wetlands (light green and blue in Fig. 1) have been drained and developed.[2] Fortunately, some of these wetlands have been converted to crops like rice (yellow in Fig. 1) that can support migratory wetland birds if managed appropriately.

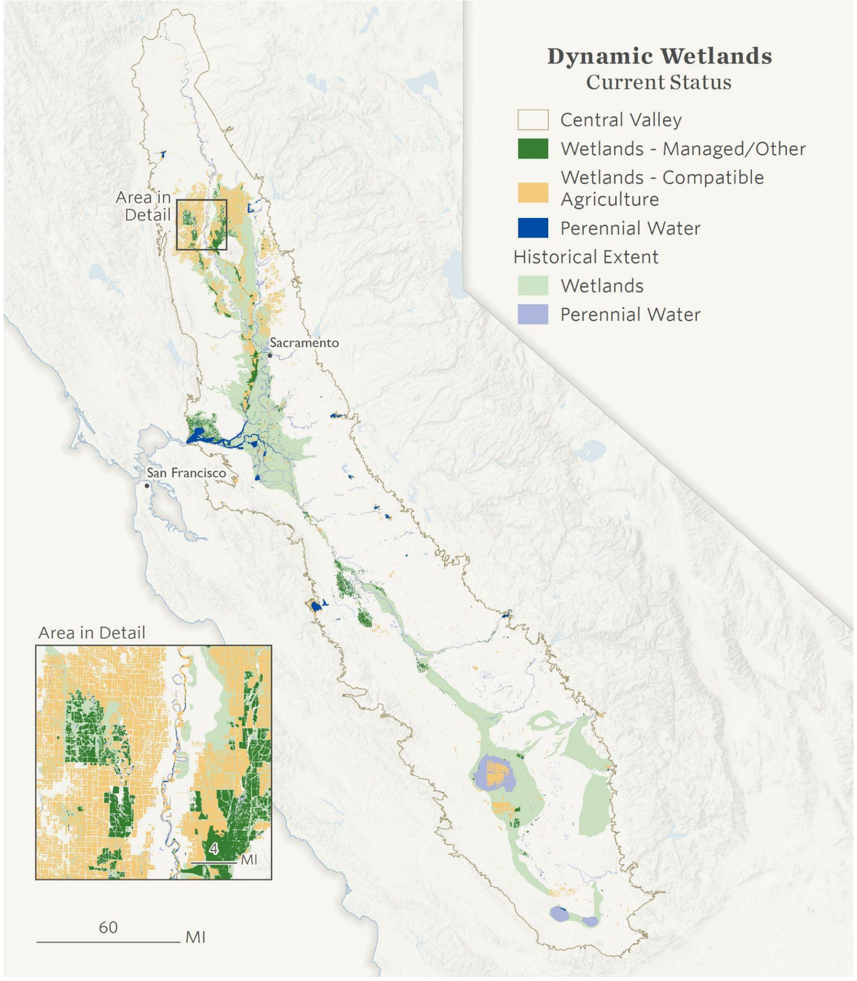


Figure 1: Wetland Status Map for California Central Valley

The Nature Conservancy (TNC) and partners launched the BirdReturns program in 2014 to pay farmers to flood their fields to support migratory wetland birds. Simultaneously, BirdReturns delivers multiple benefits for farmers, wetland managers, and communities across the Valley.[3]

The program created over 50,000 acres of flooded wetland habits in 2021-2022. The program aims to scale up and deliver an additional 100,000 acres and serve them as a rapid response program for maintaining habitat during drought.[3]

Problem Definition

The program is currently scaling up with government funds to combat the impacts of the drought.[1] In this case, new technology is needed to cost-effectively monitor the enrolled fields and ensure they are flooded. Besides ground-based monitoring using surveys, TNC has been experimenting with using free satellite images to estimate the extent and duration of flooding on the enrolled fields. A data workflow to extract and process satellite images has been established using Google Colab Python scripts to extract and process satellite images from Google Earth Engine (GEE) as shown below:

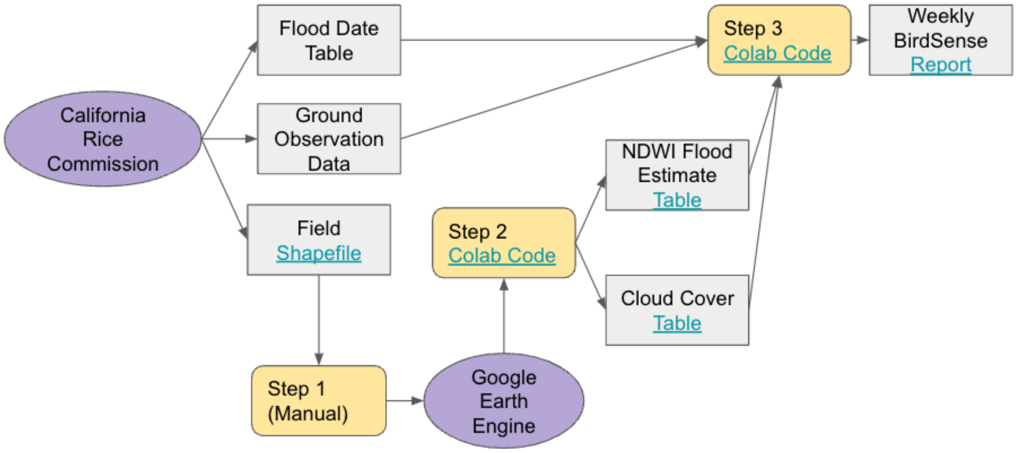


Figure 2: Existing BirdSense Workflow

However, there are several drawbacks associated with the existing workflow that need to be addressed. First, it requires human intervention to run Colab codes and send emails manually, which introduces the potential for errors and delays. This inconsistency with manual operation also leads to schedule uncertainty, making it difficult to ensure reporting timely. Furthermore, the current workflow lacks scalability as it is not adaptive to new fields or programs. This means that implementing changes or expanding the workflow to accommodate additional fields or programs would require significant manual adjustments. Additionally the non-intuitive report format, presented as a table, poses limitations in effectively communicating the information.

We need an automated data pipeline to perform the following tasks:

1. Data ingestion: collect the satellite data and the relevant information for further analysis
2. Flooding extent estimation: include a mechanism to estimate the extent of flooding based on the ingested data
3. Visualization: construct visualizations of flooding status and trends
4. Generation of weekly reports: generate comprehensive and informative weekly reports

By implementing an automated data pipeline that encompasses these tasks, the efficiency of data processing, report sharing, and field management can be significantly improved.

Methods

Data Sources

For the purposes of flood extent estimation and report generation, we use the following two primary data sources:

- **Copernicus Sentinel 2 Satellite Images**

  The SENTINEL-2 mission is an Earth observation program administered by the European Space Agency (ESA). It employs a constellation of two satellites to capture optical images of land and coastal regions. These satellite images offer high-resolution data with various spatial resolutions

(10, 20, and 60 meters) across 13 spectral bands. The mission's high revisit frequency, ranging from 2 to 3 days at mid-latitudes, enables effective land cover classification and monitoring of changes over time.[4] To access Sentinel-2 data, we utilize the Google Earth Engine (GEE) API, which is made available through a Google Cloud Service Account.

- **Google Drive for User-Provided Data**

  Users also have the option to include supplementary information, such as flooding periods, by storing relevant files in Google Drive. These files can be in formats such as Excel, CSV, and more. To retrieve these files from Google Drive, we employ the Google Drive Python API, which enables downloading of user-provided data.

Estimating Flooding Extent

Many techniques are currently available for estimating the extent of flooding using remote sensing, particularly with satellite data[5]. Two commonly used approaches are:

- **Threshold segmentation with Normalized Difference Water Index (NDWI)**
  The Normalized Difference Water Index (NDWI) is a spectral index that measures the presence of water in an image based on the differential absorption of light by water and other features. By applying a suitable threshold to the NDWI values, areas with high water content can be identified and segmented, providing an estimation of the flooding extent.

- **Image classification with machine learning algorithms**
  This approach involves training machine learning models using images labeled with land cover types. The models can learn to distinguish between different land cover types, including water bodies, and then be applied to satellite imagery to automatically classify pixels as either flooded or non-flooded to estimate the flooding extent for a given area.

Based on promising results from The Nature Conservancy's experiments for flood extent estimation, we elected to implement the NDWI approach proposed by Gao in 1996 [6] (equation below). It uses Near-Infrared (NIR) and Short Wave Infrared (SWIR) channels to calculate NDWI through the relation:

$$NDWI \ = \frac{NIR - SWIR}{NIR - SWIR}$$

where,

NIR: Band 8 - 0.83 µm wavelength, and

SWIR: Band 11 - 1.61 µm wavelength.

The given equation allows us to compute the NDWI and convert it to a binary mask for each pixel for each date from all the input satellite data. An NDWI threshold is applied to determine whether a pixel is considered flooded or not. Then, the computed NDWI results by pixel are aggregated by fields.

The field-level analysis involves calculating the mean NDWI value, which represents the field's overall water content. Additionally, we determine the flooding percentage by dividing the count of flooded pixels by the total count of pixels in the field. This provides us with a measure of the extent of flooding within each field.

At the same time, we collect cloud probabilities ranging from 0 to 100, which indicate the likelihood of cloud cover in the imagery. To identify cloud-free areas, we convert the cloud probabilities into a binary mask using a cloud-free threshold. The cloud-free percentage is then calculated as the ratio of cloud-free pixels to the total number of pixels within each field. It is necessary to exclude fields with a cloud-free percentage below the specified threshold, as cloud shadows can introduce inaccuracies into the imaging and subsequent calculations. Clouded fields are marked as unavailable and are not considered in the downstream analysis.

The next step involves generating a weekly flooding extent report by aggregating the flooding percentage for each field on a weekly basis. This aggregation is performed by calculating the average of available data for the respective weeks. The aggregated results are then compiled into a table that represents the flood probability values for all the fields over the monitoring period of the program.

Weekly Report Visualization

During the reporting process, we primarily use the Plotly Python library for visualizing the flood-related information. We employ a range of plot types, including bar charts, heatmaps, and interactive maps, to effectively represent the current flooding status and track the flooding trends over time. These plots provide insightful visualizations that aid in the understanding of the data.

To ensure thorough analysis, a watch list is provided that highlights potential fields that have not fulfilled their contractual obligations. It serves as a means for further investigation into these specific fields and addressing any issues or concerns that may arise.

We employ DataPane to create a comprehensive, accessible, and interactive reporting dashboard. This powerful tool allows us to transform Jupyter Notebooks or Python scripts into interactive web applications. DataPane integrates with various data manipulation libraries such as Pandas DataFrame, visualization libraries like Matplotlib/Seaborn, and mapping libraries such as Plotly and Folium. Various kinds of DataPane blocks including, big numbers, tables, plots, and map formatting are considered to present the report in an intuitive and interactive manner.

Building an Automated Workflow

In order to build an automated workflow, the following techniques and tools are used for data acquisition, processing, data visualization, and report sharing:

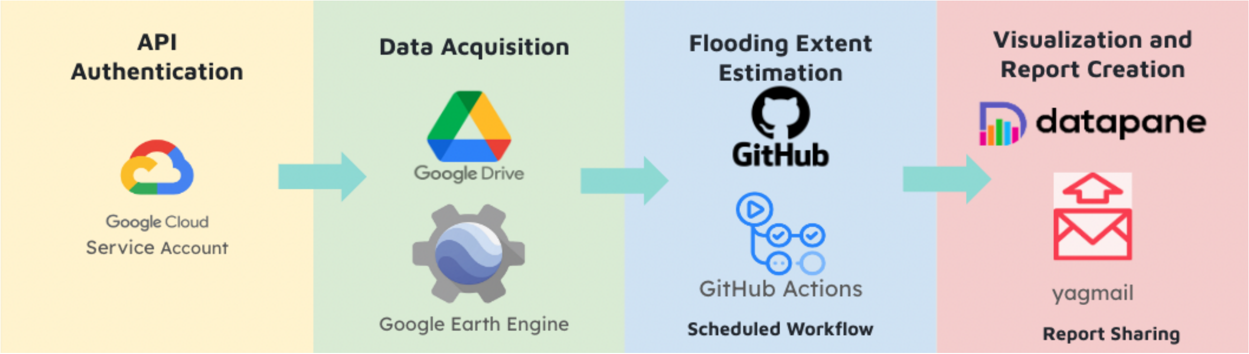


Figure 3: Techniques Integrated into the Automated Workflow

The automation of our workflow is achieved through GitHub Actions. We have implemented a YAML file that outlines a series of jobs to be executed step by step. This structured workflow allows for the seamless completion of tasks in a systematic manner.

The workflow can be triggered in many ways including manually or on a predefined schedule. For scheduled automation, we have incorporated the POSIX cron syntax, which provides flexibility in determining the timing of workflow execution. Whether on a daily, weekly, or specific day of the week/month schedule, our workflow can be configured accordingly to meet the desired frequency. When a triggering event occurs, GitHub sets up a virtual machine to execute the workflow. Each job in the workflow runs sequentially, with each step being executed within the job.[7]

GitHub Actions is primarily a continuous integration/continuous delivery (CI/CD) platform that allows developers to automate workflows and tasks. While it offers robust automation capabilities, it does have certain limitations when it comes to serving as a dedicated data pipeline solution. Other alternatives to GitHub Actions include Airflow, Stitch, FiveTran, etc.[8] These tools offer various features and capabilities for automating data pipelines, managing data workflows, and integrating diverse data sources.

For this project, we also explore Apache Airflow as an alternative data automation tool. Airflow is a popular open-source platform designed to programmatically author, schedule, and monitor complex data workflows. It uses a Directed Acyclic Graph (DAG) to represent and visualize the dependencies and relationships between tasks in a workflow.

Outcomes

Automated Workflow

The proposed workflow, depicted in Figure 4, encompasses several key steps to extract and analyze satellite images. First, the workflow retrieves satellite data from the Sentinel 2 source via the GEE API. These images are then filtered based on the fields' shapefiles and specific dates of interest. Subsequently, flooding estimates are calculated using the Normalized Difference Water Index (NDWI), and weekly summaries are generated.

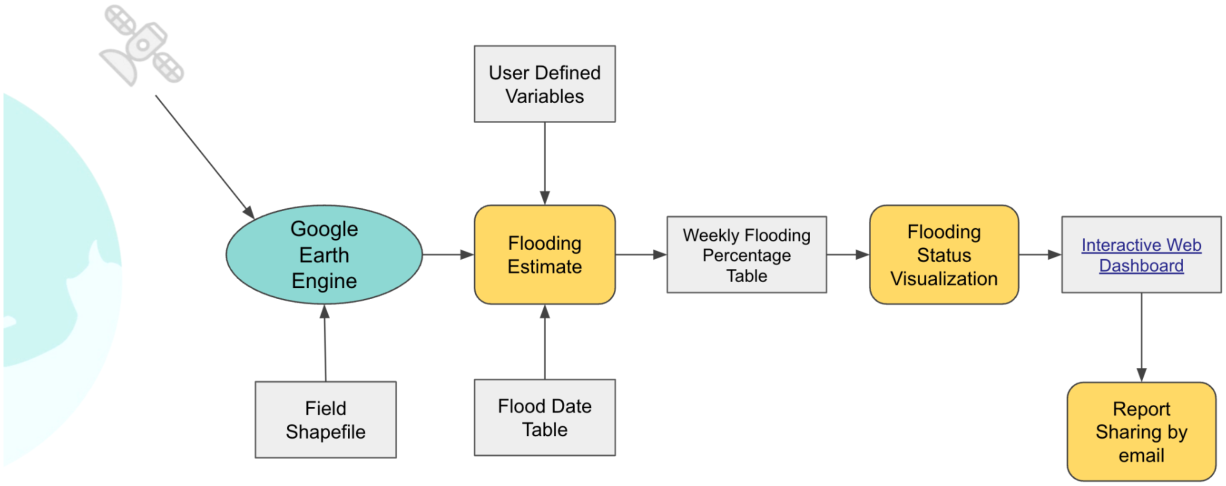


Figure 4: BirdSense Automated Workflow

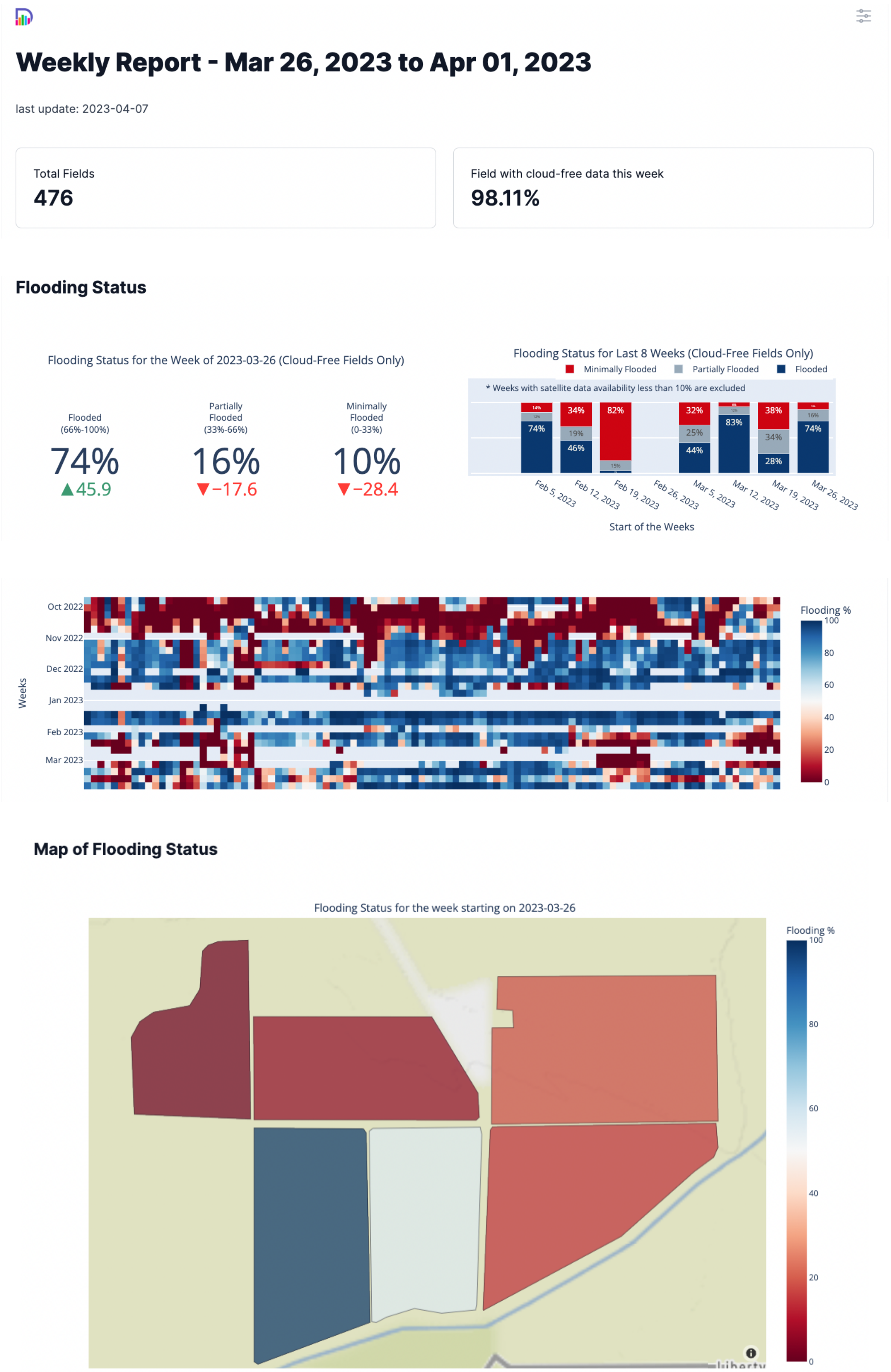
To provide insights on current and historical flooding patterns, we created an interactive web dashboard (Fig. 5). This dashboard incorporates summary statistics and visualizations of flood-related information. The dashboard is automatically shared with stakeholders via email, ensuring timely and regular dissemination of the results.

The BirdSense Workflow incorporates the following features:

- Extraction of satellite data from Sentinel 2 from GEE API
- Data processing for the flooding percentage and cloud-free indicator for each field
- Retrieval of field enrollment data from Google Drive API
- Generation of a dashboard report through the DataPane app
- Scheduling workflows for multiple programs
- Automated sharing dashboard report through email

DataPane Dashboard

The dashboard is created using DataPane and allows us to transform a Jupyter Notebook or Python script into an interactive web app. The figure below is an example of a reporting dashboard established by the BirdSense automated workflow. It can be accessed through the link: https://github.com/XinyiWang-Jessica/TNC-BirdSense-Workflows. The dashboard provides an informative and intuitive report using real-time information.



Weekly Report - Mar 26, 2023 to Apr 01, 2023

last update: 2023-04-07

Total Fields
476

Field with cloud-free data this week
98.11%

Flooding Status

Map of Flooding Status

Figure 5: An Example Interactive DataPane Dashboard

BirdSense GitHub Repository

The BirdSense GitHub repository (github.com/XinyiWang-Jessica/TNC-BirdSense-Workflows) serves as a central location to store all the necessary programs for setting up the workflow. It is comprised of several components:

- Package Requirements: a list of all the Python packages required
- Action YAML file: outline of the step-by-step jobs to be executed and triggers the workflow.
- Python Scripts:
    - Main.py: the orchestration of the execution of workflow tasks including data acquisition and processing, report generation and sharing
    - Step2.py: additional functions to support data acquisition and processing
    - Step3.py: additional functions to support result visualization
- User Definitions: user-defined information related to thresholds and BirdReturn programs
- Readme file: a detailed instruction on how to use and recreate the workflow.

Also, the authentication required to access GEE API, Google Drive, DataPane, and Gmail are saved in the GitHub Repository secrets which are read as environment variables. These sensitive credentials are saved as secrets within the repository and are read as environment variables during the workflow execution.

Moreover, the repository serves as the central hub for executing the automated workflow. With the GitHub Action, the repo is able to trigger the workflow, initiate a self-hosted runner (typically Linux by default), execute the programs within the repository, and push any changes made during the workflow back to the repository.

Ecological and Social Impact

California's Central Valley plays a crucial role in addressing water scarcity, maintaining the freshwater ecosystem, and safeguarding migratory species. It stands as one of the largest agroecosystems globally, serving vital purposes for California. [9] The BirdReturns program based in Central Valley, brings numerous ecological and social benefits, including:

- **Creating Habitat for Migratory Birds:** The program establishes habitats and food sources that support the thriving of over 1 million migratory birds across 50+ species.[10]
- **Enhancing Shorebird Density, Richness, and Diversity**[1]
- **Cost Reduction in Water Management:** The expense of renting land for creating migratory bird habitats is only a fraction (0.5%-1.5%) of the purchasing and maintenance costs to create migratory bird habitats.[9]
- **Benefits to the local community:** The program offers additional benefits to the local community by aiding in groundwater recharge and enhancing water security.[10]
- **Benefits to Farmers:** By flooding irrigated crops, the program helps in flushing out salts and decomposing stubble, preparing farmlands for subsequent agricultural activities.[9] Moreover, participating farmers receive direct payments, over $2 million in total, which further supports their economic viability.[3]

Discussion

Ground-Based Survey vs. Satellite Images

When comparing ground-based surveys to estimation using satellite images, several advantages become apparent. First, using satellite images over in-person surveyors offers the benefit of low cost for data collection. Additionally, satellite imaging allows for the oversight of all fields simultaneously, providing a comprehensive view of all the fields enrolled in the BirdReturn Program. The ability to generate weekly reports automatically ensures up-to-date monitoring of fields and enables quick identification of problematic patches which are not being flooded as intended.

At the same time, it is important to acknowledge the limitations associated with satellite images. One is the presence of data gaps that occur during cloudy periods. This can hinder the accuracy and completeness of the data obtained. Cloud shadows can introduce inaccuracies into the images, affecting the reliability of the estimation. Another limitation is that, with only satellite data, it is challenging to estimate water depth, which is another important flooding quality measurement.

In general, though flooding-extent estimates using satellite images can not replace ground-based surveys entirely, they are a valuable complementary resource, especially in the effort of managing and scaling the BirdReturn program.

Performance of the BirdSense Automated Workflow

The BirdSense Workflow has been implemented across four distinct BirdReturn programs and is currently in production, continuously distributing weekly reports and supporting field flooding monitoring. A summary of the advantages of the BirdSense Automated Workflow includes:

- **Workload Reduction:** The workflow operates without the need for human intervention, significantly reducing manual efforts and streamlining the monitoring process.
- **Cost Savings for Flooding Monitoring**: By utilizing free satellite data instead of traditional ground-based surveys, the workflow reduces the expenses associated with data collection and monitoring.
- **Cost-Effective Workflow Infrastructure:** GitHub Actions provides 2,000 minutes of run time per month, enabling no-cost execution of the workflow. Additionally, services like GEE, Google Drive, and DataPane are freely available, contributing to cost savings.
- **Ease of Setup and Modify:** Setting up the workflow or adding new programs is straightforward and requires minimum coding experience. There is no need to configure the environment from scratch.

- **Local Machine Agnostic:** The workflow runs on GitHub-hosted runners, making it independent of the local machine setup. This flexibility ensures seamless execution across various computing environments.
- **Adaptability and Flexibility:** The workflow is adaptable to new programs and functions, allowing for future enhancements and modifications.
- **Easy Collaboration:** GitHub provides a seamless environment for collaborating with other team members. Also, the GitHub repository can be easily shared and recreated on GitHub
- **Security:** Sensitive information, such as tokens required for accessing services like GEE, Google Drive, and DataPane, is securely stored in the repository's Secrets.

Comparison with Other Automation Tools

As an alternative to running the BirdSense Automated Workflow with GitHub Actions, we also established a workflow using Airflow for data acquisition and processing. A comparison of Airflow and GitHub Actions for the purpose of creating an automated workflow is summarized in Table 1. Generally speaking, the size of the task and organization dictates which tool is more appropriate. GitHub Actions provides a centralized, streamlined, and lightweight solution for simple automation tasks. It can be easily integrated into existing workflows and set up with minimal effort. On the other hand, Airflow is a powerful tool for managing complex workflows. It requires significant infrastructure and resources to set up and maintain.

Future Applications

Currently the automated workflow is being used as a complement to the expensive and time-consuming on-the-ground surveying for TNC BirdReturn Programs. Our goal is to eventually have it replace in-person monitoring to reduce costs and improve consistency and efficiency in field assessment.

Features	**GitHub Actions**	**Apache Airflow**
Use Cases	CI/CD platform. Simple data pipelines.	Complex data pipelines with many tasks and dependencies.
Complexity	Relatively low. Easy to set up.	High. Comparatively steep learning curve.
Customization	Limited options.	Highly customizable and extendable.
Scalability	Relatively low. May not handle large-scale data pipelines.	High. Scalable solution for large data pipelines and growing data volume.
Integration	Integrates well with GitHub repositories.	Widely integrated with data stores, other services, machine learning frameworks, etc.
Infrastructure Requirement	Local machine agnostic. Run on GitHub-hosted runners (Linux).	Local or virtual machine.
Reproducibility	Easy to share and reproduce.	Has environment prerequisites to reproduce.
Cost	Free usage with a run-time limit.	Free and open source.
Security	Can save tokens in repo secret.	Requires third-party services to store and access tokens and passwords.

Table 1: Comparison of GitHubActions and Apache Airflow for Automated Data Pipeline

The BirdSense workflow provides an easy and cost-effective way to build a data pipeline for tasks like data acquisition through API, data processing, dashboard development, or report sharing. This data pipeline approach to other projects requires automation, and this approach is especially beneficial for small-scale projects and start-up companies, for which efficiency and costs are the primary concerns.

Recommendations

Since the implementation of the BirdSense automated workflow, we are continuously improving its efficiency, accuracy, and adaptability by applying it to different BirdReturn programs and scenarios. Future directions and areas of the pipeline that can be improved include:

- **User Feedback:** Gathering feedback from new users and applications for different programs will help identify areas for improvement and ensure its usability across different scenarios.
- **Enhanced Interactive Dashboard:** Expanding the functionality of the interactive dashboard to include features like field filtering and historical data investigation will empower users to explore and analyze data more effectively.
- **Accuracy Improvement with Machine Learning**: Investigating and researching machine learning algorithms for flooding extent estimation might lead to accuracy improvement.
- **Test Program for Continuous Integration and Deployment (CI/CD):** Developing a comprehensive test program to validate the output of the workflow will ensure its stability and reliability.

References

1. Golet, Gregory H., et al. "Using ricelands to provide temporary shorebird habitat during migration." Ecological Applications 28.2 (2018): 409-426.

2. Reiter, Matthew E., et al. "Local and landscape habitat associations of shorebirds in wetlands of the Sacramento Valley of California." Journal of Fish and Wildlife Management 6.1 (2015): 29-43.

3. The Nature Conservancy, "BirdReturns - Creating flexible bird habitat in California's Central Valley," 2023. [Online]. Available: https://birdreturns.org/

4. The European Space Agency, "SENTINEL-2 MISSION GUIDE," [Online]. Available: https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2

5. Li, J., Ma, R., Cao, Z., Xue, K., Xiong, J., Hu, M., & Feng, X. (2022). Satellite Detection of Surface Water Extent: A Review of Methodology. Water, 14(7), 1148. https://doi.org/10.3390/w14071148

6. Gao, B. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment, 58(3), 257-266. https://doi.org/10.1016/S0034-4257(96)00067-3

7. GitHub Docs, “Understanding GitHub Actions” , [Online]. Available: https://docs.github.com/en/actions/learn-github-actions/understanding-github-actions

8. TrustRadius, "Data Pipeline Tools", [Online]. Available: https://www.trustradius.com/data-pipeline

9. Rohde, Melissa M., Mark Reynolds, and Jeanette Howard. "Dynamic multibenefit solutions for global water challenges." Conservation Science and Practice 2.1 (2020): e144.

10. Sesser, Kristin A., et al. "Waterbird response to variable-timing of drawdown in rice fields after winter-flooding." Plos one 13.10 (2018): e0204800.