WiNGS-Ops
AI Assurance Report

May 26, 2023

Bringing Ingenuity to Life.
paconsulting.com

Denver Office

PA Consulting Group Inc.
Suite 3550
1700 Lincoln Street
Denver
CO 80203
USA
+1 720 566 9920

paconsulting.com

Prepared by: PA Consulting, Inc.

Version: FINAL

Figures

Tables

Executive Summary

PA Consulting (PA) was engaged by San Diego Gas & Electric's (SDG&E) Wildfire Mitigation team to conduct an independent review of the Wildfire Next Generation (WiNGS) models. In this report, PA covers the findings and recommendations for the WiNGS-Ops model. WiNGS-Ops utilizes Artificial Intelligence (AI) and Machine Learning (ML) algorithms to forecast wildfire and Public Safety Power Shutoff (PSPS) risks and consequences to help inform real time operational decisions during PSPS activation events. This model has involved utilizing industry and domain-specific assumptions to generate inputs and derive calculations to drive an optimized output. This model also uses a wide array of disparate data sources to ensure a thorough and encompassing view of the network is included.

PA finds that the WiNGS-Ops model is being developed using more formalized and standardized approaches than prior versions. WiNGS-Ops has not been extensively used in an operational setting, but the forecasts it produces are reasonably documented and well understood by the current set of end users. The approaches WiNGS-Ops takes meet the regulatory requirements outlined in the California Office of Energy Infrastructure Safety 2023-2025 Wildfire Mitigation Plan Technical Guidelines. SDG&E teams working on the WiNGS-Ops model are well aligned and have built and maintain WiNGS-Ops with a level of competence in keeping with the quality of the end-product.

PA applied its proprietary AI Assurance Framework in this independent review to assess the end-to-end process utilized for model initiation, development through operation. This review focused on establishing if industry best practices for the deployment of AI and ML have been adhered to and recommend improvements to be made. Review outputs are captured in this report, which is structured in the same manner as the AI Assurance Framework. This Framework is results focused, and as such, we associated severity levels on the recommendations, from the perspective of potential of impact to output should recommendation not be implemented.

The latest version of WiNGS-Ops is being actively developed ahead of the 2023 fire season, so our assessment was focused on the approaches taken in developing and operating the WiNGS-Ops model and its sub-model components. We found that as WiNGS-Ops matured, additional standardized approaches were developed (e.g., model set up and code repository structure, general model selection, cross validation, and a standardized list of model features) and utilized by the team. Approaches used for model training generally follow industry best practices, however, improvements may be made in the way that some of the sub-models (e.g., vehicle contact) utilize unbalanced data sets. Common features between models have been standardized and limited at the current time to ensure that the features used in model development are not too complex or costly and can be utilized in production runs. At times, other less important features should be excluded to decrease risk of model overfitting.

The teams that operate WiNGS-Ops are split into two groups, each with a clear focus: WiNGS-Ops Data Science team conducts model training and development, and the Digital Innovations Advanced Analytics (MLOps) Team runs the production inference pipeline. The WiNGS-Ops Data Science team conducts data, feature, and algorithm explorations to develop a suitable risk and consequence forecasting sub-model which then is transferred to the Advanced Analytics team to run and maintain in the production environment to generate wildfire and PSPS risk and consequence forecasts. The two teams work collaboratively under an operating agreement that clearly outlines areas of responsibilities and how the teams work together.

The WiNGS-Ops Data Science team tends to work on their sub-models individually and using their own judgements on what features to include and decide when their sub-models can be pushed to production. This approach favors speed to development, but there are limitations in standardization of approaches and potentially challenges for prioritization that could be missed. We found instances where the modeling team are resource constrained and would prefer to have a larger team focusing on a standardized development and testing approach but are limited by what can be realistically achieved to have the models ready for the upcoming fire season. Finally, there are limitations to WiNGS-Ops model predictions (e.g., normal system configurations and connectivity, infraction data not considered in model) which are understood by the current business users. However, as WiNGS-Ops matures and more end users are identified (e.g., potentially within the Emergency Operations Center), the need to have more individuals who understand what the model predictions mean and what the limitations are will be greater.

1 Introduction

San Diego Gas & Electric's (SDG&E) Wildfire Mitigation teams have been utilizing Data Science techniques and technologies to create advanced models for risk and investment planning purposes. These models are data assets in use for Capital Planning and Operational purposes. Each of these models have involved utilizing industry and domain-specific assumptions to generate inputs and derive calculations to drive an optimized output. These models also use a wide array of disparate data sources to ensure a thorough and encompassing view of the network is included. These models are both referred to as the Wildfire Next Generation System (WiNGS) models. SDG&E engaged PA Consulting to perform the independent 3rd-Party review on two models. The two main models discussed for initial review are mentioned below with this report forming the output of the review of the WiNGS-Ops model:

- WiNGS-Planning: This model, hosted in AWS, is utilized for investment planning purposes. One of the main sources of data involved is a historic Ignition Model, based on mileage of overhead conductors. Other variables are included based on wind probability assumptions. A Decision Matrix is utilized for final output of model for use by investment teams.
- WiNGS-Ops: This is a risk-based model for more real-time determination of the risk posed to the network in wildfire season. This model has a strong Machine Learning component and is time-series based. The model has been released and is operating in AWS Sagemaker.

With the exponential increase in data, organizations are leveraging Artificial Intelligence (AI) and Machine Learning (ML) to identify insights and help make better informed, data driven decisions. These data science projects can be a catalyst for organizational strategy and objectives. However, it is also vital that these projects prove to be reliable and trustworthy. PA's AI Assurance Framework follows industry best practices for the deployment of AI and ML by reviewing and providing evidence on how they are governed and managed. PA's general assurance framework is designed to provide assurance that the models are well documented, has the trust of its business owners, inputs and limitations are well understood, algorithms are secure (against unapproved or un-intentional changes), and they work as intended. The PA Consulting team applied this framework to the independent review of SDG&E's WiNGS models.

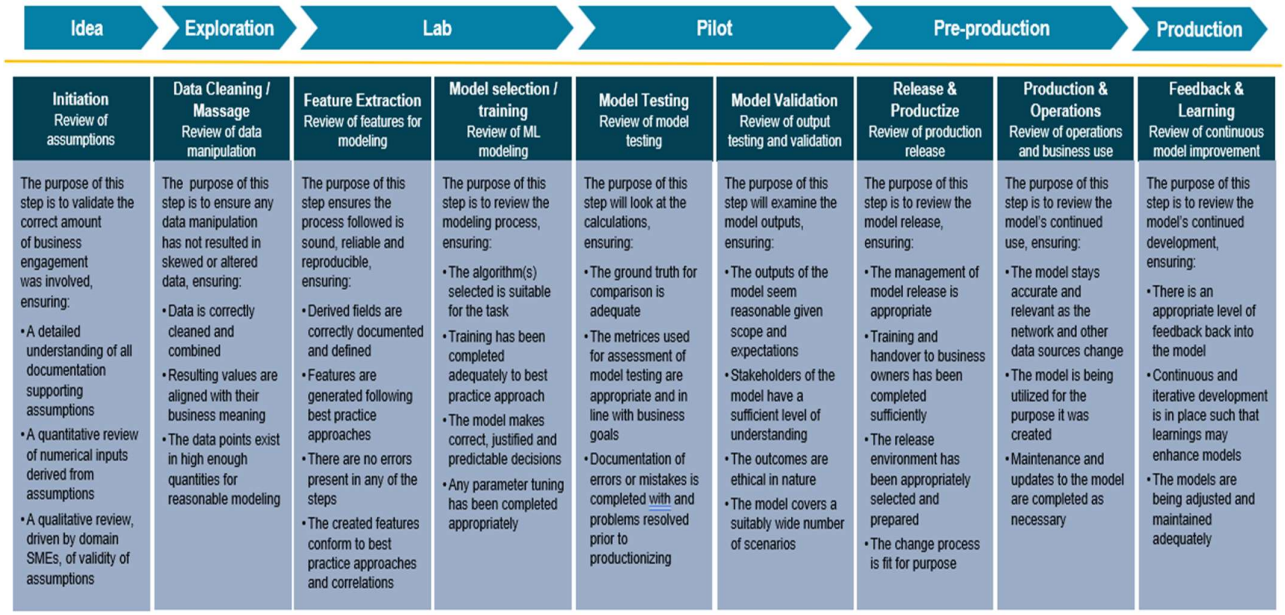
1.1 AI Assurance Framework and Approach

The PA AI Assurance Framework is designed to review an organization's AI system for best practices and risks, ensuring trust in the end solution. The results of the evaluation are aimed to help foster the responsible design, development, deployment, and use of those AI systems over time.

The framework steps through each stage of the algorithm lifecycle to investigate common underlying risks, identify controls to mitigate these risks, and detect evidence required to prove these controls are in place. At each stage, discussions with the relevant stakeholders are performed to understand the approach and methods in place, with the stakeholder answer questions and providing proof while walking through pertinent documentation, datasets, code, repositories, final deliverables, and IT environments.

Figure 1 below shows the stages of PA's AI Assurance framework which were followed during the review of the WiNGS-Ops model. These modular areas of focus look to cover the full model development lifecycle from Initiation and problem formulation through to release, production runs and use by the end business users.

Figure 1: PA's AI Assurance Framework



Each of these stages aims to focus on a specific part of the modelling process. Table 1 found on the next page describes each stage's process objectives and the overall risk that can be introduced when not following best practice:

Table 1: AI Assurance Framework stages and objectives

Stage	Potential Risk	Overall Process Objective
Initiation	Lack of business engagement	Ensure the business takes full sponsorship of the project: provide necessary resources during the project and after go-life
Data Cleaning	Inaccurate data cleaning	Ensure that data is correctly cleaned and combined, such that the relevant data is appropriately formatted, has values that are aligned with their business meaning, and exists in high enough quantities for the prospective model
Feature Extraction	Inadequate feature extraction	Ensure that features are extracted in a sound (reliable, accurate and reproducible) way and operates within the boundaries of applicable rules and regulations
Model Selection/ Training	Inadequate selection/training	Ensure that the model is selected and trained in an adequate way so that it makes correct, justified and predictable decisions
Model testing	Inadequate testing	Ensure that testing is adequately performed and documented so that errors and mistakes are identified and solved before the solution is set to production
Model Validation	Model not validated	Ensure that the model is validated and authorized by key decision makers
Release & Productize	Not fit for purpose	Ensure that the solution fits its purpose and continuously adds value to the business
Production & Operations	Model becomes inaccurate (drift)	Ensure that the solution stays accurate, predictable, without drift
Feedback & Learning	No feedback & learning to improve the model	Ensure that feedback and continuous learning loops are in place so that the model improves continuously

Throughout the independent review, findings are captured, and recommendations are made. The criticality of the potential impact which a recommendation aims to remediate is defined in

Table 2 below:

Table 2: Impact criticality definitions

Potential Criticality	Definition
H	Significant impact on accuracy of risk and consequence forecasts. These represents immediate concerns that needs to be addressed
M	May impact WiNGS-Ops outputs / predictions, potentially over/under stating the risks and / or consequences
L	Unlikely to change WiNGS-Ops outputs / forecasts

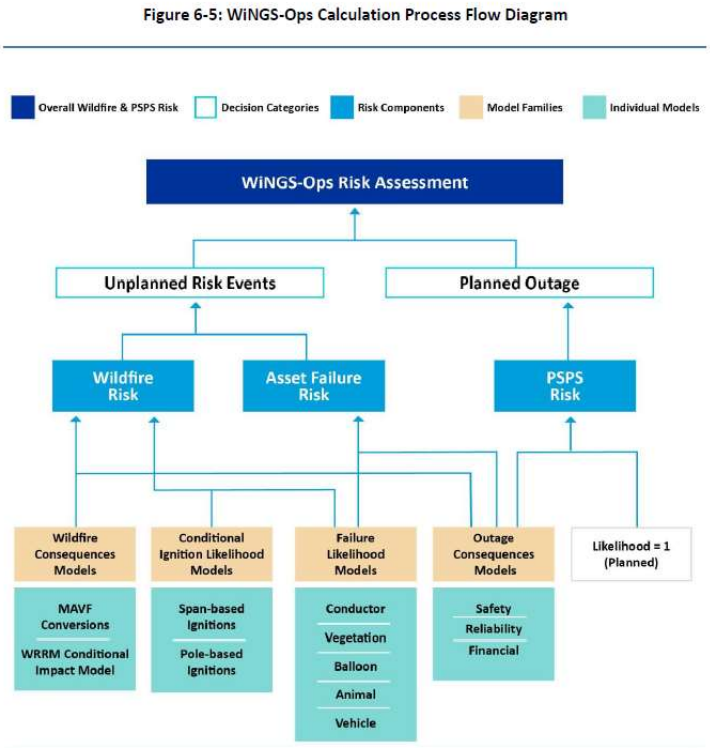
1.2 Regulatory requirements

This independent review of WiNGS-Ops is conducted to meet requirements outlined in Section 6.6 Quality Assurance and Control and Appendix B – Model Substantiation, of the Office of Energy Infrastructure Safety's (OEIS) 2023-2025 Wildfire Mitigation Plan Technical Guidelines (WMP Technical Guidelines) Documentation. This review of model substantiation is aimed to ensure that a model is correct and suitable for its purposes, understood by the users, and validated. We conclude the model is used and useful and follows the requirements of the WMP Technical Guidelines.

1.2.1 Risk assessment framework and calculation schematic

WiNGS-Ops is a tool aimed at providing data and AI / ML driven projections of forecasted wildfire risks and consequences as defined by the risk assessment framework (Figure 2). it follows the calculation schematic in Figure 3, which represents SDG&E's approach to meeting the risk category component requirements as outlined in Section 6 of the WMP Technical Guidelines.

Figure 2: SDG&E WiNGS-Ops Risk Assessment Framework



To satisfy regulatory risk assessment outlined in the Wildfire Technical Guide, WiNGS-Ops uses separate components to derive the likelihoods of failure for equipment (e.g., overhead conductors), contacts (which comprises of contacts from animals, vehicles, vegetation, as well as unclassified outages). There are separate Probability of Ignition (PoI) for each of the categories of failures, however, there is a plan in place to amalgamate them this year. The consequences of the wildfires follow the identical approach and methods as WiNGS-Planning and uses the MAVF approach to quantify risks around safety, reliability, and financial consequences, with the appropriate weights assigned.

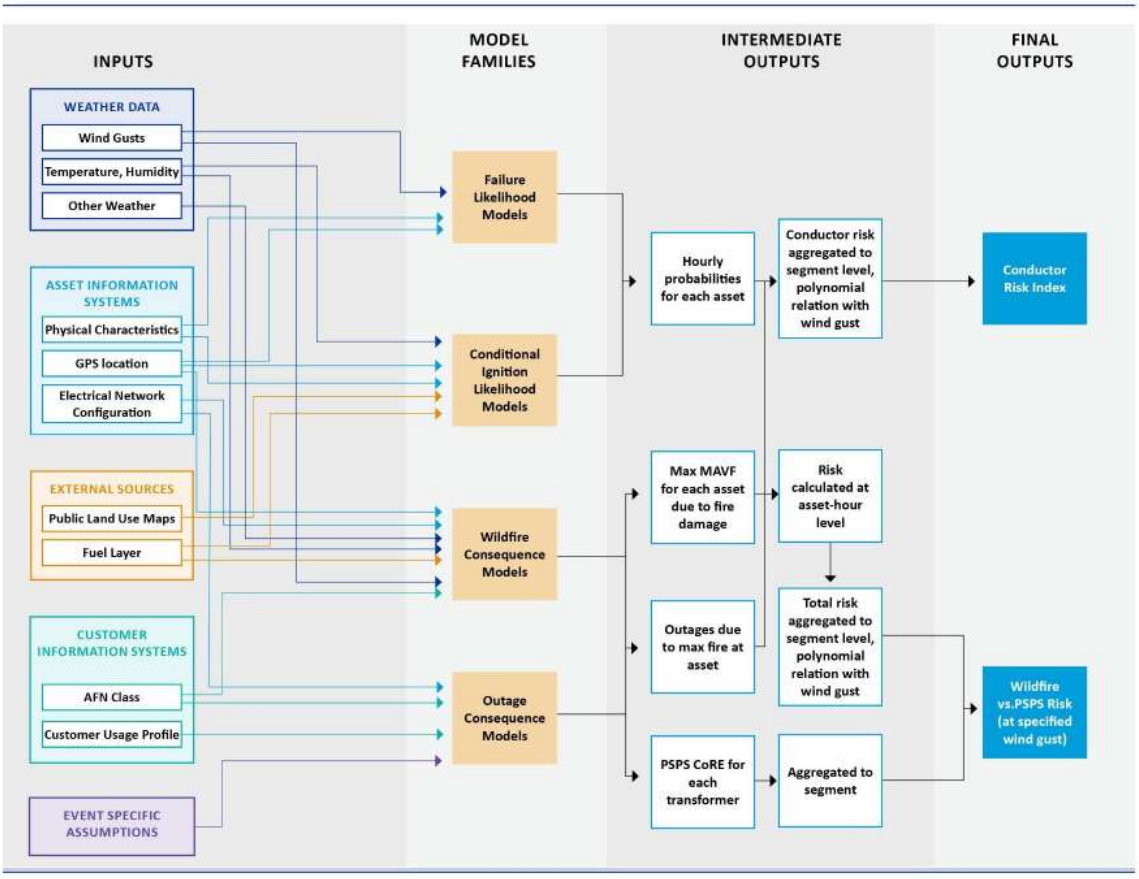
PSPS risk calculates the consequences of an event assuming there will be an PSPS activation (i.e., the likelihood of this event is 1), and the consequences are calculated using similar approaches and methods as WiNGS-Planning using the MAVF approach.

The WiNGS-Ops model, unlike the WiNGS-Planning model is less deterministic, and aimed at informing real time risks to be addressed during potential PSPS activation events. WiNGS-Ops differs slightly from WiNGS-Planning as WiNGS-Ops uses a conditional probability of failure approach to assess the short-term risks of wildfire. That is, instead of using the Likelihood of Risk Event (LoRE), the risk is assessed

as a product of the Probability of Failure (PoF) and POI given a failure. Further, WiNGS-Ops also leverages machine learning algorithms to determine the various PoF and PoIs.

WiNGS-Ops calculation schemes are presented in Figure 3. The wildfire conditional risk is computed as the conditional probability of ignitions given outages that fall into select categories (i.e., conductor, vegetation, balloons, animals, vehicle and other). PoFs and PoIs for the different components are calculated separately, and then aggregated to develop the overall wildfire risks. The Wildfire Consequences are computed to address the various impact factors such as safety, reliability, financial consequences resulting from wildfires (consequences are weighted to address their relative importance). The consequence approach is the same as the one used by WiNGS-Planning.

Figure 3: SDG&E WiNGS-Ops Calculation Schematic



PA's review of the WiNGS-Ops model is conducted in accordance with our AI Assurance Framework (described in more detail in Section 2). WiNGS-Ops is currently used by the Wildfire Mitigation Strategy (WMS) team to provide insights as to wildfire risk and PSPS scores during potential events. As the model matures, more business end users are envisioned in the future.

In the case of WiNGS-Ops, there are two sets of code repositories: one for data exploration and training, and one for production use. The models are first trained in the training environment by the WiNGS-Ops Data Science team and transferred to the Advanced Analytics (MLOps) team to optimize and maintain the tools for use in a production environment. There is a clear delineation of responsibilities covered in an Operations Agreement, with frequent check-in sessions to discuss and address issues.

1.2.2 Prior independent assessments

As WiNGS-Ops is a still maturing model, there have been internal developer and Advanced Analytics checks of the code to ensure quality and fit for purpose. However, WiNGS-Ops models have not been independently assessed prior to this effort.

2 WiNGS-Ops Assessment

To be compliant with the Office of Energy Infrastructure Safety Wildfire Management Program guidelines, SDG&E initiated an independent 3rd-party review on the WiNGS-Ops model. This model evaluates both wildfire and PSPS impacts at the sub-circuit and segment level to forecast both PSPS risks as well as wildfire risks. The key decisions driven from this model provide quantified risks to inform decisions around activation of PSPS events (e.g., timing and duration).

2.1 Limitations

This section details any limitations of the independent review, while considering their impacts on the findings. This review was not intended to verify the veracity of the assumptions, or the correctness of the calculations carried out, but is more meant to validate the approaches and methodologies used. As the WiNGS-Ops model is maturing and several component pieces are still under development as of the writing of this report, our assessments are limited to reviewing existing code and the processes followed by the WiNGS-Ops development effort (e.g., common approaches to data cleaning, feature extraction, model selection, training, testing, and production).

The final production of the next version of WiNGS-Ops code is expected to be finalized by Q3 2023. As this model is not yet fully mature as of the writing of this report, our assessments are focused on the various model components and algorithms, rather than a deep dive into the algorithms themselves. As such, the findings, conclusions, and recommendations are generally focused around the general approaches for WiNGS-Ops rather than the detailed algorithm functions and capabilities. Further, as the WiNGS-Ops model components and algorithms are stored in disparate systems, we have not had an opportunity to assess how all the WiNGS-Ops model component algorithms function as a whole, and instead only assessed the individual components.

2.1.1 Models reviewed

The WiNGS-Ops model is comprised of the model families (sub-models) as outlined Figure 3. As the model is currently under development, we were limited to reviewing the following sub-model repositories due to on-going development work. Table 3 lists the sub model repositories that were reviewed.

Table 3: Sub-model Repositories Reviewed

Sub-Model Repositories Reviewed
Vehicle Contact (In Dev)
Foreign Object (In Dev)
Balloon Contact*
Vegetation
Weather Sanitization
MLOps Repository
Span Based Ignitions
Pole Based Ignitions
PSPS Model
Conductor Model ** ‡

* Foreign Object model is replacing the Balloon Contact Model

** Conductor model is stored on a local computer and full in-depth review was not possible

‡ Conductor model is intended to be replaced by Asset Wind Driven model

2.1.2 Inputs-related Limitations

The documentation for many of the inputs (for example: the pole restoration time) is missing key information and detail.[1]. Some of these assumptions are used for consistency purposes (e.g., MAVF safety weights) were provided by SME and are consistent with the SDG&E RAMP reports but may not be as fully documented as other assumptions with detailed calculations, rationales, or justifications.

As directed by stakeholders, WiNGS-Ops does not take into consideration inspection information (e.g., if there were alternative configurations, presence of infractions) on the circuit. However, it assesses the projected risks and consequences from the category outages covered in its Risk Assessment Framework.

WiNGS-Ops PSPS risk is calculated based on the number and types of customers who would be potentially impacted should a PSPS event be activated (default length is 24 hours). WiNGS-Ops uses customer classifications that include Medical Baseline (MBL), and various categories of Access and Financial Needs (AFN) to understand the different socioeconomic segments that would be impacted by a PSPS decision (both from lost consumption and incurred expenses perspectives).

2.1.3 Usage-related Limitations

WiNGS-Ops is meant to be an assessment of forecasted risks. During a PSPS activation event, WiNGS-Ops would use the forecasted wind gusts to calculate the forecasted wildfire and PSPS risks[2] as a function of wind gusts. WiNGS-Ops generates two respective WiNGS-Ops score curves based on wind gust. The two WiNGS-Ops score curves intersect at the “red dot”, where the risks and consequences of wildfire and PSPS are identical. Above the intersection (e.g., high wind gust), the risk of wildfires tends to exceed the PSPS risk and consequences, and below the red dot, the risk of wildfire tends to be below the PSPS risk and consequences. This output is generated based on forecasted wind gusts to help inform decision makers’ PSPS activation decisions. WiNGS-Ops has only been used once in a PSPS activation event[3], and as such, work is still underway to determine how best to use the recommendations from the WiNGS-Ops forecast.

2.2 Assessment Outcomes

Based on PA’s assessment, the WiNGS-Ops model is still maturing, but is being built to be in compliance with requirements outlined in the Wildfire Mitigation Technical Guidelines. The approaches followed by each component are reasonable, sound, and documented. The existence of two separate code repositories (one for developing and training the models, and one for production model) does present certain risks in terms of assumptions and computation methods may be slightly different (e.g., training code base is not optimized for production runs, and production code base does not train models on new data). However, to address these issues related to separate outcomes, the WiNGS-Ops Data Science and Advanced Analytics teams developed the Operating Agreement to delineate responsibilities and formalize handoffs, meetings, and documentation requirements. The team is regularly maintaining and working to address items on the development roadmap to carry out further enhancements.

There may be development resource constraints that limit how quicky various additions or improvements can be addressed, however. The WiNGS-Ops Data Science does not solely focus on model development and enhancement, and the use of the WiNGS-Ops model from an operational perspective

[1] Either due to model developer or SME turn over and the original creators of these assumptions and inputs did not provide full documentation as to how the input was generated / determined.

[2] Risk in this context is defined as the product of likelihood and consequence

[3] A slightly older, less mature (and non-cloud based) version of WiNGS-Ops was in the 2022 PSPS event. SDG&E’s PSPS reports are available at: https://www.sdge.com/wildfire-safety/psps-more-info#reports

is still in the initial stages. As the model uses become more well established during PSPS events, and more improvements / requirements become identified, the current development team may not have sufficient capacity to meet all future improvement demands. Industry best practices typically will have at a minimum two capable individuals focusing on each model during the development phase. This can help provide peer reviews, consistency checks, and help validate the work done and updates.

3 WiNGS-Ops Model Assurance Details

3.1 Process Introduction and High-level Recommendations

Our independent WiNGS-Ops review captured a number of recommendations, with severity ranging from Low to High. These have been documented in the sections below, split for each pillar of PA's AI Assurance methodology. Each of these recommendations aim to individually address a part or process component, suggesting an improvement in line with Data Science and modelling best practice. This section also touches on some larger recommendations which span multiple pillars of review and would have a greater positive impact if implemented. These more impactful recommendations also would require greater levels of planning and effort in their implementation to reap the full value.

3.1.1 WiNGS-Ops Data Science Process

An effective, efficient, and well-aligned data science process is critical for optimal modeling and decision support outputs. If the team executing the process is not operating in a similar fashion, or if the capacity of the team is sub-optimal, then the effectiveness of the modeling process and output could be at risk. The team must be aligned in the approaches they take to model creation, training, testing and validation such that each constituent part of WiNGS-Ops model meets the same level of statistical rigor and review. The team must also be close-knit such that an open forum exists for discussion and idea-generation which can enable a robust peer-review process. Finally, a strong data science team must be well managed with activities clearly tracked and prioritized such that the members of the team have capacity to drive optimal results in the modeling and work that they perform.

Industry best practices tend to have development teams working on each model, so that there can be a degree of code approvals, checks, and validation. This also helps to support continuity for the model development and maintenance. Additional resources could also support additional unit testing of code, which in turn will also provide some benefits to model documentation as unit testing often provides a description of how the code is supposed to work (i.e., what business objectives are) and what the outputs are supposed to be (i.e., what the expected outcomes are).

The WiNGS-Ops Data Science team which runs the build and training cycles of the WiNGS-Ops model is small but highly capable with great subject matter expertise which lends it to building effective and successful machine learning models. While the capability of the team is strong, the members often work independently in model development, with minimal support from each other. This, in part likely due to an overall capacity issue, could lead to an inconsistency in the approaches taken to build, test and validate models. A resolution for enabling greater alignment in this team would be enhancing capacity through improved management of the prioritized activities and size of the team. Potential improvements to the effectiveness and efficiency of the team may be attained through the following recommendations:

R1.1 Model Approach Standardization: As most of the model build completed by the team is independent, there is a potential lack of standardization for the development, training, testing and validations of models. There is some standardization in the use of repos and feature extraction, would recommend expanding this standardization to all aspects of model development so that all models are tested and validated to the same specification.

> *Severity Level: Low* – without a standardized approach, each model may not hold the same level of credibility given varying levels of testing and validation. This would improve consistency of model outputs.

R1.2 Internal Model Review Process: Best practice data science solution development must include a level of peer-review to validate the scripts that are developed and operated. Creation of a more formalized internal model review process would provide a forum through which ideas may be discussed and considered before implementation, and through which a robust and consistent approach to model review may be performed.

> *Severity Level: Medium* – this would enable potential improvements or ideas to be highlighted and discussed leading to more effective and efficient models.

R1.3 Model Documentation: As the team has been operating in a reactive state to changes in the WMP guidelines and recommendations, full documentation of each of the models is not complete. To reduce any risk due to reliance on the experience and knowledge of the individual team members, the recommendation would be for a concerted effort to ensure documentation is complete for each of the latest model versions to be released for fire season 2023.

> *Severity Level: Low* – without robust model documentation, there is a reliance on the experience and memory of the members of the team to explain the reasoning behind model decisions and changes.

R1.4 Team Enhancements: The team consistently faces capacity constraints due to the ever-changing landscape of the WMP guidelines and recommendations, coupled with continued regulatory requests for data and information. As such, the team operates reactively to requests and priorities, without a true backlog of tasks captured and delivered against. The recommendation would be to enhance the team in a couple of different ways. A scrum master can help generate and manage a backlog of tasks and activities such that activities may be prioritized, and a demand management process may be created. The addition of a data analyst to the team could assist with external regulatory data requests, alleviating some of the time demands of the WiNGS-Ops Data Science team.

> *Severity Level: Medium* – without changes to the team size and roles of team members, the full potential of the members of the team may not be realized. Improved team size, capability and demand management would allow for a more optimal environment, within which the greatest value may be generated.

3.1.2 Enterprise-wide Data Governance

Strong enterprise-wide Data Governance is critical for any successful data-driven organization. Without proper Data Governance procedures in place for managing the data assets in SDG&E, the value of the enterprise's data may not be fully realized. Proper ownership and management of the data ensures that the data is always clearly defined with definitions that are agreed and understood across the business. There must also be robust management of the information generated. This includes ownership and governance of the calculation methodologies, and definitions for the models, metrics and KPIs used across the business. Technology may be utilized to aid in a more robust management of data, allowing the mechanisms and processes to be digitally formalized in a solution such as Collibra or Informatica.

A Data Governance function exists for WMP and separate Enterprise Data Governance which covers the range of inputs used in WiNGS-Ops, however these functions are not well integrated which may pose a risk. This means that any changes in business definitions, standards for use or overall changes in the underlying data may occur without the WMP being made aware. This could lead to an unexpected change in the data, and an unexpected change in the output. There is no clear owner of the derived data and calculations involved in the model to determine their definition, principles of use or to make decisions on adjustments or changes that would be required. Without proper governance, there may not be a clear path for making decisions for changes of the model. The recommendation would be for a governance function, integrated with the enterprise-wide function which would cover the data inputs, information use, modeling, and the overall decision-making process for changes to WiNGS-Ops. This would enable correct management of the data and information assets across SDG&E, utilized by the Wildfire Mitigation team, ensuring:

R1.5 Enhance communication with data owners: Ensure that there is an integrated function, such that communication from specific business owners for each data input into the models is cohesive and timely. This would ensure definitions, use, bounds for validity and decisions on changes that would be needed are communicated. They would also be responsible for ensuring that the data is up to date and accessible to all who may need it.

> *Severity Level: Medium* – lack of communication from data owners may result in unexpected changes and diminished data integrity. The data owner is accountable for the use, quality, and protection of a dataset.

R1.6 Calculation ownership: Owners of the specific constants (e.g., PSPS risks) and calculation methodologies sometimes called “information" such that their definitions and approaches are agreed, documented and uniform across the business. This is to ensure that any colloquial terms used for aggregated data assets are consistent such that an output like “miles of span in HFTD in one group’s calculation is the same as another’s.

Severity Level: Low – a calculation owner will be accountable for ensuring calculation methodologies are clearly defined and are used appropriately and consistently.

R1.7 Broader model ownership in the form of a board / group with regular meeting cadence to agree higher-level changes and adjustments, reviewing output of sensitivity analysis and changes prior to implementation. This would ensure that the responsibility for driving the direction of overall model enhancements and improvements is agreed amongst the Developers, Wildfire Mitigation team and the Business users.

Severity Level: Low – without regular communication between all stakeholders, the direction and prioritization of model development and improvements can be missed.

3.1.3 Reliance on External Resources

The team responsible for development and iteration of the WiNGS-Ops model is highly capable and efficient. Due to the requirements of different activities and resource constraints present, external contractors have been utilized across SDG&E. There are two areas where use of external contractors has impacted the WiNGS-Ops team. These areas are around the Enterprise Asset Management Program (EAMP/Asset 360) data expertise and the Advanced Analytics team who release the models and operate them in production.

The recommendations below relate to steps which would reduce the level of risk introduced with reliance on external contactors. The recommendations have been split across the two identified areas:

R1.8 External EAMP/Asset 360 data experts: EAMP/Asset 360 provides a rich asset data source used in modeling. The data itself is a clean and curated version of GIS and Asset Management data. The program is operated by external contractors who also remain as the data source SMEs. The recommendation would be for an internal SDG&E team to be onboarded to share some of the responsibility as SME for the data source. The source, including all dictionaries and implemented manipulations, should also be fully documented such that any new user may easily gain a complete understanding of the data and its use.

Severity Level: Medium – with a continued reliance on external parties for this critical data source, the SDG&E team will not gain full ownership, understanding and control over the underlying data. Internal SDG&E SME expertise in the data source will ensure a robust and future-proof mechanism for data understanding, questions, and data updates.

R1.9 External Inference team: The development team responsible for the inference aspects of the WiNGS-Ops are a group of external contractors. Feedback has been very positive on the capability and success of the team and the models seem to operate sufficiently well. The team is effective in the conversion of models from training and test phase to inference phase but do not look to challenge the training team to improve the models. Given the reliance on external contractors in this team, the recommendation would be to integrate more SDG&E resources into the team so that some of the knowledge and experience is internal.

Severity Level: Low – as the WiNGS-Ops model continues to mature and gain complexity, the technical debt on the external development members of the Advanced Analytics team will grow, increasing this reliance.

3.2 Model Initiation

In the initiation step of the AI assurance framework, the purpose is to validate the correct amount of business engagement was involved in the early stage of model planning and development. This ensures that the problem has been appropriately defined and understood, and that the modeling exercise was commenced with the correct goals in mind.

The process involves understanding the initial problem formulation and workshopping process to devise the task and focus for the models. Additionally, a vision of an end-product model is clear and an accepted view as to what the value the model will be to the business with metric driven KPIs. It is also important to ensure the relevant business areas have taken full sponsorship of the project. This pillar will aim to establish whether proper engagement has been made and maintained with the business owners, ensure that the boundaries of the model are well-defined and understood by stakeholders and review the documented assumptions and requirements involved in the model.

3.2.1 Findings

F2.1 The WiNGS-Ops model is a product of guidelines put forth by energy safety in 2019 after three PSPS events occurred. The Order Institute Rulemaking (OIR) needed to have a model that looks at both fire risk and PSPS risk. At that time, SDG&E already had a robust PSPS decision making process. A team created a beta conductor model that predicted fire risk. This model was not embedded into the PSPS decision tree but was used as another data point in the decision process. With the first full version of WiNGS-Ops, to be deployed this year (2023), the model outputs will be provided as an additional information point to support the de-energization decision.

F2.2 Even though OIR required the model, there is currently no plan for OIR to review or audit the models. After a PSPS event, a PSPS report is created which will include details about the model and the generated results. The OIR can issue violations if it deems a requirement is missing.

F2.3 The Utility Incident Commander (UIC) will ultimately make the decision on whether to perform a PSPS event, and WiNGS-Ops will provide additional data to help inform that decision.

F2.4 The business owners pass on the requirements to the development team, who then work on detailed development. There are regular progress updates and information sharing sessions to document progress, but there is no formal approval process by the business owners to ensure that the WiNGS-Ops Data Science team has captured all the new requirements.

F2.5 There is currently no predetermined metric for determining the value the model brings to SDG&E.

F2.6 There was regular participation by SMEs, specifically the engineering team, fire science team, and meteorology team during the initial stages of model development. There is continued engagement with those SME as needed when new features are explored.

F2.7 Formal documentation that captured the critical elements of the initiation and planning stage for the WiNGS-Ops was not performed. Most of the original stakeholders who were involved in the early stages are still with SDG&E.

3.2.2 Recommendations

R2.1 Per finding F2.2, in order to ensure that all OIR requirements are met and prevent possible violations, we recommend building and maintaining a formalized report that tracks the requirements and how they were carried out. Having this existing documentation will not only confirm what the requirements are and if and how they were completed but will also be ready to pass along to the OIR as appropriate.

Severity Level: Low – this will help prevent potential violations from the OIR by tracking all the requirements and how they were completed.

R2.2 Per finding F2.4, create a formal process through which requirements for model changes are captured, tracked, and completed against. This will ensure that the changes have been understood and captured correctly and will allow a definition of done to be assessed against, by the end users in their approval of model changes. We suggest using a project tracking tool such as Boards in Azure DevOps since it is already used by the team.

Severity Level: Low – without a documented process, requirements and requested changes may be incorrectly implemented or the end users may not have an easy mechanism for change approval.

R2.3 Per finding F2.5, in order to determine the value the model is bringing to the business, we recommend establishing metric(s) to gauge the effectiveness. This will ensure that the impact of model improvements and developments over time may be quantified and tracked.

Severity Level: Low – this recommendation will increase end user buy in and understanding to the changes that are enacted in the model.

R2.4 Per finding F2.7, we recommend documenting the initiation stage in order to capture critical elements of the initial planning stage. This includes defining what problem this model will solve, what is the feasibility of this model, who are the end users and how do they want to ingest the model outputs, who are the subject matter experts and what is their ability to participate in the model development, who will be the business owner of the model, what are the initial assumptions and how were they determined, and confirmation that all relevant business areas have taken full sponsorship of the project. Additional details on why certain decisions were made with respect to the model generation are also critical to document in the initiation process. Going forward, with the initiation of new model versions, we recommend documenting these critical elements so they can be referenced by future developers and users of the model.

Severity Level: Low – without this documentation in place, future developers and end users may have a more difficult time understanding the decisions and assumptions that were made, which SMEs to turn to for input, how the model will be measured for success, or the original problem and objectives.

3.3 Data Cleaning

In the Data Cleaning stage, the purpose is to validate the data cleaning process of fixing or removing data is thorough and has not resulted in skewed or altered data. This ensures the data inputs are as reliable and accurate as possible. The process involves assessing the methods used to validate the input data. Additionally, to confirm the steps taken on detecting, correcting and documenting duplications, missing values, misspellings, lexical errors, irregularities, and mis-fielded entries are in line with expectation.

3.3.1 Findings

F3.1 Data cleaning steps are completed for all the sub-models, which include converting fields to the proper data type, searching and converting blank values to consistent 'NaN' values, and dropping duplicate records.

F3.2 There are many weather-related data validation checks built into the Weather Sanitization model. Additional data validation takes place during the exploratory data analysis, however, it is not automated. Additionally, there is no automated data validation checks built into the inference pipeline outside of Weather Sanitization.

F3.3 In the vehicle contact model, historical vehicle contact events on poles without latitude and longitude values are dropped from the dataset. This reduction in data may remove valuable data points, limiting how representative the data set is.

F3.4 For the rear occasions when span conductor type is missing in the historical record, the immediate surrounding spans are checked. If both connecting spans are the same, then the missing span is assumed to be the same. If not the same, or too many adjacent spans are missing, then the highest risk conductor type is assumed (Copper #6). Similarly, if the conductor bearing angle is also 'NaN', then the value is assumed the worst, which is 30 degrees, which is the maximum loading direction for an expected Santa Ana wind event.

F3.5 It is unclear to the WiNGS-Ops Data Science team who the electrical network data (i.e., "as-operated" state of grid connectivity or configuration) owner is, or who to go to with questions. This has resulted in a lack in knowledge of the data processing and data quality steps that take place, and even the root data source. The model is built on the as designed and built network, however, the network changes every day when switches are switched and there is no way to capture the network as switched state. This makes backcast analysis difficult. Having a different network configuration during a potential PSPS activation event that is different than one that the model is trained on may impact the number of customers and affect the PSPS risk. However, from an operations perspective, circuits are supposed to be returned to the normal configurations ahead of these events, which in some ways limit this risk.

F3.6 An imputer function is used in the Foreign Object model to autofill numerical columns that are NaN values with the mean values, and object and boolean columns with 'N/A'. No records are dropped. The numerical data is scaled using the standard scaler function, while the OneHotEncoder function is applied to the categorial columns. The target values for the foreign object are converted to log odds.

F3.7 Electric network configuration data (i.e., system connectivity) is utilized for the consequence models in the WiNGS-Ops models. It is also utilized by a wide number of other groups for different purposes, without the same level of education or understanding. This could pose a risk as the data may be used for a purpose which does not align with the purpose for which it was created.

F3.8 The technical subject matter experts for the EAMP/Asset 360 data source are all external contractors. This creates a risk as there is a knowledge gap for the SDG&E team.

F3.9 SAIDIDAT outage data is currently a manual data request from the reliability group. The reliability group emails audited copies of the data in an excel spreadsheet that has the requestors name stamped in the file name. The data is considered very sensitive and there is a high level of

control by the data owners. This data is utilized in a number of the different models which form WiNGS-Ops.

3.3.2 Recommendations

R3.1 Per finding F3.2, every data input should pass through some degree of automated data validation check to look for outliers, errors, text control, contradictions, etc. Each of these validation checks should have associated documentation that includes what to do when data is missing or anomalous. This should be implemented in the inference pipeline and should be consistent with data validation performed by the WiNGS-Ops data science team during their EDA process.

> *Severity Level: Medium* – there is currently a lot of reliance on source data owners to validate their data, however, if erroneous data makes its way into the model during the model development or inference pipeline, inaccurate outcomes will result. Poor data quality can therefore lead to poor model outcomes, which will result in a loss of trust in the model by the end users.

R3.2 Per finding F3.3, we recommend working with the GIS team to develop a logic-based solution for imputing pole location information using other fields when historical pole locations are missing. This may include utilizing an existing GIS redlining process for resolving these gaps.

> *Severity Level: Low* – this would ensure that the data used in modeling is most representative of SDG&E's network. It may also help reduce the number of minority class records that are dropped due to missing data.

R3.3 Per finding F3.5, we recommend noting this as a limitation of the model and prior to PSPS activation events that the systems are restored to the as-designed states wherever possible. However, we also recommend contacting Operations personnel to confirm the correct owner of the network as-operated electrical connectivity data since this data is a critical component of the WiNGS-Ops model. Additionally, seeking out information on the root data source, how it is validated, and the existing assumptions are critical for ensuring a complete understanding of the data and its correct use.

> *Severity Level: Low* – without knowing the correct data owner or who to reach out to with concerns or data issues, there will be continued uncertainty of the data and of the stewardship and accountability surrounding that data.

R3.4 Per finding F3.7, would recommend for greater governance and controls to be put in place for each of the data objects utilized by WiNGS-Ops such that none of the data created for and used in the models is inadvertently used for a different purpose, generating alternative and incorrect views of the landscape.

> *Severity Level: Low* – although this may not directly impact the output of the WiNGS-Ops model, it may affect the credibility of the data sources used if the source is used incorrectly elsewhere.

R3.5 Per finding F3.9, We recommend a direct query of SAIDIDAT data from its source database. This eliminates the reliance on an individual and prevents potential human error in the process.

> *Severity Level: Low* – manual data request and transfers are reliant on the requestor to ask for the information. Automating the request process may be a better way to obtain updated outage history data on a scheduled basis rather than on an as-requested basis.

3.4 Feature Extraction

In the Feature Extraction stage, the purpose is to validate the feature extraction process is sound, reliable, and reproducible. This ensures any derived values are informative, non-redundant, and appropriately reduces the number of resources required for evaluation. This process involves understanding the feature selection process to create a feature dataset, understand the feature extraction steps that were utilized to make the modelling more effective and responsive, and ensure the derived fields are correctly documented and defined.

3.4.1 Findings

F4.1 Feature Importance is evaluated to determine which ones are most important, but unimportant features are not removed. Keeping unimportant features may result in overfitting in certain models, where a combination of unimportant features allows a model to perform very well on a subset of the data, while not being fully applicable to the full, representative data set.

F4.2 The models utilize SANGIS public land use data which is updated annually. As this model is intended to provide a near real-time prediction, it must be based on the latest representation of the current landscape. As such an annually refreshed data source might introduce an inaccurate representation into the outputs.

F4.3 The SANGIS data set used in inference does not provide full coverage for assets located outside San Diego County, therefore values used for those assets are imputed using mean values.

F4.4 Transformations in the models have been converted from Python-based to SQL-based in the inference stage. This is beneficial as it allows for greater parallelization when dealing with large datasets.

F4.5 With this year's model developments, the focus lies within the amalgamation of sub-models, such as the pole ignition and span ignition model combing into one ignition model, and the animal contact, balloon contact and other contacts combining into one foreign contact model. There is no plan to add new feature sources due to resource constraints.

F4.6 When the WiNGS-Ops Data Science team looks to add new features, they balance the potential value the feature brings to the model with how difficult it would be to deploy the feature due to the data infrastructure or computational processing requirements.

F4.7 In new sub-model development repositories, the project structure has been standardized. This includes the standardization of five Python notebooks that perform the feature extraction process such that it will be consistent in each sub-model repository.

F4.8 The PSPS model utilizes weights in the categories of safety, financial and reliability, which are determined using SME input. The financial and reliability weights are all 1 (unweighted) while the safety weights vary from 1 to 20 based on the vulnerability of the customer. While these weightings are based on SME input, they are not documented and evidence for their explanation is unavailable.

3.4.2 Recommendations

R4.1 Per finding F4.1, for the models that do not have auto regularization, we recommend removing the less relevant features as measured by the feature importance function outputs. Removing less relevant features will help with the stability of the model, avoid overfitting, and reduce computation cost.

> *Severity Level: Medium* – it is unclear at this stage the impact that inclusion of these unimportant features has on the outputs. It has the potential to skew results which may have a large impact, so has been rated as such.

R4.2 Per findings F4.2 and F4.3, would recommend working closely with the SANGIS team to incorporate SDG&E territory currently not covered in their existing coverage data, as well as

request for a more frequent than annual data updates. This would ensure the models have access to the same information as the rest of San Diego county, and is most up to date during a red flag warning event.

Severity Level: Low – models run on data which has not been refreshed for a while or on imputed data based on mean values may provide inaccurate outputs. This may cause a model to under-represent the potential consequence of an ignition due to a missing at-risk land use.

R4.3 Per finding F4.5, due to the time pressures and resource constraints, the team does not have capacity to further improve models with development or incorporation of additional features.

Severity Level: Low – the impact of this would be minimal due to the models' existing satisfactory performance but might represent a missed opportunity for continued model improvements and enhancement.

R4.4 Per finding F4.8, would recommend creation of a documented framework to define the safety weights used in the PSPS model such that there is an explainable process through which they may be assessed and updated based on additional SME input. These weights must also be integrated into version control, so that any changes are managed and easily tracked, model version to model version. Documentation around this would help provide future model developers and users better understand why certain values were used, and what the historical justifications and rationale were.

Severity Level: Low – without a clearly documented process for suggesting changes to the weights and version control to track those changes, it may be difficult to provide explanatory evidence in support of decisions driven by this model.

3.5 Model Selection and Training

In the model selection and training stage, the purpose is to validate the techniques used for training and selecting the models. This ensures the resulting models were evaluated based on the necessary criteria and is making justified and predictable decisions. This involves walking through the process used for selecting and training the models, ensuring the algorithm selected is suitable for the task, training has been completed adequately to best practices approach, and review the metrics used for evaluating the various models tested.

3.5.1 Findings

F5.1 The Vehicle contact model currently handles the class imbalance (900 contacts in ten years or 87,000+ hours, i.e., many more non-events than events of interest) by down-sampling. At the time of review, no other techniques have been tested to address the imbalance in the data.

F5.2 The adopted model selection approach allows for discretionary judgement by the modeler to choose the best approach that balances accuracy and explainability.

F5.3 For each of the sub-models of WiNGS-Ops, there is one model owner. Currently there are Three WiNGS-Ops Data Science developers working on the various sub-models of WiNGS-Ops. Each of the sub-model owners will make the final decision on when to publish and push the sub-model to the production system. Sub-model owners seek feedback from their peers, but there is no formalized peer review or approval process in place.

F5.4 The conductor model is currently utilizing linear regression for predictions. Though linear regression is easily explainable, it may lack the elegance and complexity to fully represent the scope of the equipment failure resulting in ignition problem being modeled. A non-linear algorithm may better predict equipment failures due to ignitions so should be considered in future.

F5.5 The conductor model was trained on 12 years of data up to 2021.Starting in 2015, better wire down data was recorded and therefore different data quality exists prior to and after 2015. Also, the validation of the model outputs has been less accurate in the most recent two years that were tested (2020 and 2021), indicating that the model may be under-representing the conductor failure risk.

F5.6 The WiNGS-Ops Data Science team builds models with different data sources than the Advanced Analytics team. The WiNGS-Ops Data Science uses EAMP/Asset 360 data, which undergoes data cleaning and imputation resulting in the most accurate historical data. Equipment Failure Report (EFR) data is also used to further increase the accuracy of the historical data for training the models. There is a process to update the records should discrepancies be discovered (although historical discrepancies may not make it from EFR to the main EAMP/Asset 360 data). The Advanced Analytics team pulls data from GIS for the production model since it is considered the active data source of truth. Based on our interviews, training (using slightly updated historical data) and inference are done using slightly different datasets.

F5.7 Similar to finding F5.4, the foreign object model is trained using EAMP/Asset 360 data while the inference version in production, used during activation events, utilizes the GIS system of record to make predictions.

F5.8 The vehicle contact model uses XGBoost for its predictions. No other algorithms were tested to assess performance. At the time of the model review, no hyper-parameter tuning had taken place, however, it was verbally communicated that the GridSearchCV function will be used for tuning the model as part of the current development.

F5.9 The foreign object model uses a random forest regressor model, with a focus on hyper-parameter tuning to improve accuracy and less on testing other algorithms.

F5.10 Models are refreshed at least once per year. This also includes refreshing the training data, which only occurs once per year.

F5.11 To check for overfitting, k-fold cross validation is utilized during the training process. This method iterates through different test-train splits, evaluating how the model performs against each test split by comparing the variance of the scoring.

3.5.2 Recommendations

R5.1 Per finding F5.1, we recommend testing other approaches to handling class imbalanced data, including up-sampling, SMOTE, and ADASYN, in order to determine the most applicable method for each model.

> *Severity Level: Medium* – down-sampling excludes significant amounts of data which may result in an unrepresentative data sample being used for training and testing the model.

R5.2 Per findings F5.2, F5.4 and F5.8, we recommend testing other algorithms to ensure that the most suitable algorithm is used to solve the problem, balancing complexity of understanding and training with accuracy of modeling outputs.

> *Severity Level: Low* – without validating that there isn't a more suitable algorithm for the model, the team cannot be certain that they have built the most suitable model for the specific application. Testing alternative algorithms may help build greater trust in the end product.

R5.3 Per finding F5.3, we recommend a more collaborative approach towards model development and release. A peer-reviewed approval process (similar to the one used by WiNGS-Planning) can ensure consistency between sub-models and the best practices are followed.

> *Severity Level: Medium* – individual working may lead to inconsistencies between models resulting in models with differing levels of robustness being deployed.

R5.4 Per finding F5.5, we recommend retraining the conductor model based on data from 2015 to present, utilizing the 2022 data for testing and validation. This will ensure the most representative data is utilized in construction and training to create the most accurate and useful modeling outputs.

> *Severity Level: Medium* – based on the most recent data used for validation, the model under-represented the potential risk due to conductor failure. Re-training this model would generate a more representative output.

R5.5 Per findings F5.6 and F5.7, would recommend that the models are trained on the same data sources which would be utilized for inference in production such that the resulting outputs are most relevant and applicable.

> *Severity Level: Medium* – as the models were trained on different source data, the data relationships learnt may not be representative of what would be seen in the EOC. As a result, the outputs of the models may not be as accurate as if the data used for training was the same source as used in inference.

R5.6 Following on from R5.5 and in relation to finding F5.7, consider a larger program of GIS data cleaning, validating and improvement and investigate if existing GIS red lining processes can be leveraged to ensure the GIS system of record for assets represents the most accurate view of assets in SDG&E's territory. This would ensure that any modeling application or activation event would consider that most accurate understanding when making data-driven decisions.

> *Severity Level: Low* – it is critical that decisions in the EOC are made based upon the most accurate representation of the assets in the field. Ensuring this data source is accurate and up-to-date is crucial to enabling the EOC to operate from an informed position.

R5.7 Per finding F5.8, the foreign object model is using GridSearchCV for hyper-parameter tuning, however, it has not yet been implemented for the vehicle contact model. We recommend taking the same approach to tune the hyper-parameters.

Severity Level: Low – consistent use of techniques across models ensures that the quality and robustness of each model is uniform and contributes to an optimal output.

R5.8 Per finding F5.2, with explainability as a key driver for model selection, we recommend looking to incorporate SHAP to help explain the output of the models through calculating the contribution of each feature to the model output. These values can be used to understand the importance of each feature and to explain the results of the model.

Severity Level: Low – without a full understanding of the importance and contribution of the features in a model, the driving factors of the model's outputs are less explainable.

3.6 Model Testing

In the model testing stage, the purpose is to validate the techniques used for testing the models are in line with the overall goals of the business. This ensures the performance of the model is tested to agreed-upon 'ground-truth' data and measured using appropriate metrics. This process involves understanding the methods and metrics used for testing the model, review sensitivity analysis to evaluate the model performance, assess if the outcome of the model meets the acceptance criteria, and identifying any potential problems and weaknesses.

3.6.1 Findings

F6.1 All models are tested on a 20% hold-out test-split of data, and are measured with various SciKit Learn's classification metrics, including accuracy, F1 Score, Recall, Precision, and ROC AUC.

F6.2 The foreign object model makes use of a modified Brier score. Where the traditional Brier score utilizes the square of the difference between predicted and actual observations, normalized for the sample size, the foreign object model uses that same calculation un-normalized. This means that the score is dependent on the size of the population used in modeling and may cause issues when comparing across model versions with different population sizes.

F6.3 For the vehicle contact model, only the down-sampled test data split was tested for model performance. No testing was performed on an imbalanced validation set.

F6.4 There is no set standard for "good enough" when it comes to model testing performance. The current goal utilized is if the model performs better than random, and if it is better than what they currently have.

F6.5 The ground truth data is determined by the relevant SMEs within the company. For example, the ignitions data is managed by the fire science team with their own set of processes to ensure data accuracy. These steps are not documented; however, the fire science team is responsive and accessible.

3.6.2 Recommendations

R6.1 Per finding F6.2, would recommend using the full Brier score such that the outputs are unaffected by population size. This will enable Brier scores to be compared across different versions of a model to allow model improvements to be validated.

> *Severity Level: Low* – the risk of not using full Brier score is that this modified score might be inadvertently used to compare models with different sample sizes. This would give an inaccurate view of the performance comparison so could result in an incorrect modeling decision.

R6.2 Per finding F6.3 and F7.3, for the vehicle contact model, we recommend incorporating a nested cross validation where one fold is an out-of-period imbalanced data split for the final validation, and other fold is split for training and testing on balanced sampled data set. This would provide an additional method for validating the accuracy of the model. Ensure the right metric is used for the evaluation, as some metrics are better for evaluation when there is class balance (ROC AUC) and others are better for when there is class imbalance (Precision-Recall AUC).

> *Severity Level: Medium* – validating the imbalanced data with this approach checks performance of the model on the real class distribution.

R6.3 Per finding F6.4, establish a consistent and agreed approach for model testing across the team such that each member may be sure of the optimal model and be agreed when training is complete. This will ensure consistency across models and build credibility with the end users.

> *Severity Level: Low* – models may have differing levels of robustness without a uniform defined and agreed approach to testing.

R6.4 Per finding F6.5, we recommend for all data that is ingested into the models that detailed documentation is provided by the data owners. The documentation should contain pertinent information such as the data owner, data collection methodology, data dictionary, structure of the data, data validation and quality assurance steps taken, data manipulations from the raw data, and confidentiality, access and use conditions. This will ensure a detailed understanding of the data that can be reference as needed, critical for ground truth data.

Severity Level: Low – without detailed documentation, there is a risk the data can be misinterpreted, or if there is turnover or new hires on the WiNGS-Ops Data Science or Advanced Analytics teams, they may have a more challenging time referencing and understanding the data inputs.

3.7 Model Validation

In the model validation stage, the purpose is to assess how the models are challenged, validated, and approved. This will ensure the models have been reviewed by all pertinent stakeholders and the outputs are correctly interpreted. This process involves reviewing the validation process and ensuring it accounts for a suitably wide number of scenarios, the outcomes are ethical in nature, and the model achieves its intended purpose. Also, to ensure the outputs of the model seem reasonable given the scope and expectation. Additionally, to confirm that the model is validated and authorized by key decision makers, and they have a sufficient level of understanding which is in line with the documented purpose of the model. Note: Our Model Validation stage differs slightly from SDG&E''s definition of model validation, which encompasses our Model Testing and Validation pillars as well.

3.7.1 Findings

F7.1 Back-casting has been completed to apply new models to historical PSPS events. This helps understand the efficacy of the model outputs, ensuring alignment with past decisions and data understanding. The exercise was time-consuming and therefore was only completed on one historical event. It was time consuming for multiple reasons, one being that the historical network system changes were challenging to get, and the other reason was that a lot of information was stored in spreadsheets. Because most models have migrated to AWS, better version control and data snapshotting is in place to increase efficiency in performing back-casting exercises.

F7.2 The end users perform domain-knowledge driven validations, but do not have a formalized validation process to challenge the outputs and ensure the results are within reason.

F7.3 Validation of the models with imbalanced data occurs only on the down-sampled data set.

F7.4 The Advanced Analytics team performs high-level validation on the production output data, such as checking to see if there are any missing feeders from what they were expecting.

F7.5 Detailed sense checks are performed by the WiNGS-Ops Data Science model owners. For example, for the PSPS model, 100’s of transformers are randomly selected and the risk scores are assessed based on the input values. After that, segment level checks are compared against each other. When a model doesn’t pass the sense check, they will not deploy the model and figure out where the issue is.

F7.6 For the conductor model, a script is run that plots and maps the model output results that helps the validation process.

F7.7 Model results are shared with SMEs to ensure the model outputs are in line with their domain knowledge.

F7.8 After the weather data is cleaned, a report can be generated to validate all the cleaning and data imputation. This report compares how many new data points were generated for each field by weather station and compares the field averages pre and post imputation for the various fields per weather station, such as wind direction, wind speed, and wind gusts. This script is not automatically run, it is run locally and stored locally on an analyst’s computer.

3.7.2 Recommendations

R7.1 Per finding F7.1, a more holistic and reliable model validation process may be created to allow automated back-casting for each model change. This would allow for greater confidence in the updated version of each model to be gained. Given the snapshots of data are now maintained in the cloud, this ensures that this process would be simpler to perform.

> *Severity Level: Low* – without an automated and uniform approach to model output validation, validating each new model release will be a time-consuming and inconsistent process.

R7.2 Per finding F7.1, we recommend ensuring that all necessary data and calculation components are captured, including the network configuration, at the time of a PSPS event to help streamline future back-casting exercises.

Severity Level: Low – implementing this would allow for the automated and uniform approach mentioned above could be enacted for model back-casting.

R7.3 Per finding F7.2, establishing a formalized validation process by the end users will establish consistency in the validation approach, and also build credibility with OEIS by demonstrating the results are reviewed in a specific and systematic way.

Severity Level: Low – without a formalized validation process, there is the potential for end users to validate the model differently every time a new model version is released. This may result in missing an important check or reviewing an output that differs from a previous model version.

3.8 Model Release and Productization

In the Model Release and Productization stage, the purpose is to validate if a formal release management process is in place, the quality of code is tested in a demonstratable and correct way, and a proper hand-over plan to the business stakeholders is in place. This process involves reviewing the code quality checks, checking the version control procedures, ensuring the solution fits its purpose and adds value to the business, and evaluating how data is consumed by the end users.

3.8.1 Findings

F8.1 In March 2023, an operational agreement between the WiNGS-Ops Data Science team and the Advanced Analytics team was established that resulted in the WiNGS-Ops Data Science team controlling the feature engineering and model training and testing, while the Advanced Analytics team controls inference.

F8.2 The full WiNGS-Ops model has yet to run completely on AWS with some sub-models still running locally on WiNGS-Ops Data Science team computers. This can impact version control of models as they are not integrated into the existing repositories.

F8.3 The current conductor model and PSPS model are developed locally on a computer and version control once it is deployed. It is planned for both models to be migrated to the cloud and will be stored in git-controlled Azure DevOps.

F8.4 Models are developed by the WiNGS-Ops Data Science team and stored in AWS model registry. The inference team pulls the models from the model registry for production.

F8.5 The Advanced Analytics team does not have full knowledge of how the various models were trained. This may mean that they aren't fully informed for effective review and critiquing while creating the production versions.

F8.6 Separate teams for model training and model inference (productization) ensures a greater level of error checking and validation. One example of this was the identification by the Advanced Analytics team of an inner join where 40% of spans were being dropped, which could have massively skewed the outputs of the models.

F8.7 The Span Ignition and Pole Ignition models are not perfectly independent; however, the plan is to combine them into one model this year to prevent any overlap.

F8.8 Flake8 is used for linting, based on PEP8 code standards with slight configuration options. Auto formatting is completed by Black, and Pyright is used as the static code checker. A profiler tool has not been used extensively.

F8.9 Currently, the team is not performing any unit testing on the model scripts. No operational unit testing may allow bugs or defects to remain in the model, unidentified until impacting the output data.

F8.10 The team is also not performing any integration testing to ensure all functions are scripts are working together as intended.

F8.11 Very few functions have docstrings that explain the overall functionality of the code.

3.8.2 Recommendations

R8.1 Per finding F8.2 and F8.3, Recommendation to migrate the conductor training model and PSPS model scripts to Azure DevOps Repos. This will ensure development on local machines are version controlled, tracked appropriately, and accessible by the rest of the team. This will also allow models to leverage cloud compute capabilities, meaning that more advanced models may be produced. Additionally, the PSPS model should be passed to the inference team such that the entire WiNGS-Ops model can be executed through the inference pipeline.

Severity Level: Medium – current processes limiting version control and access could introduce errors and confusion in the correct version that should be run in production. Full cloud migration would limit the risk of this issue.

R8.2 Per finding F8.5, the model training team should provide a more thorough explanation of the model training process and decisions which would enable the Advanced Analytics team to have a better grounding for implementing the code. As well as education sessions, thorough documentation would enable any new team members to be onboarded swiftly.

Severity Level: Low – without full understanding and knowledge of the model training process, the Advanced Analytics team may not be able to add as much value in critiquing and improving the models.

R8.3 Per finding F8.7, would recommend that the pole and span ignition models are combined to remove any overlaps which might exist in the separate models.

Severity Level: Medium – currently the models are not fully independent, which may skew the results. This should be rectified such that an accurate representation of risk may be generated.

R8.4 Per finding F8.8, consider running a profiler to help understand the resource consumption of the various operations in the model. This can potentially resolve performance bottlenecks and help the model execute faster.

Severity Level: Low – this recommendation does not affect the model output but may improve the runtime performance of the model.

R8.5 Per finding F8.9, incorporate unit testing to ensure all functions are performing as intended and more easily isolate errors when they occur. Unit tests also check that the code still functions as expected after making changes which builds code stability.

Severity Level: Medium – Without unit testing, there is no assurance that code will function correctly, and undiscovered bugs may exist. This can lead to poor quality modeling results and wasted time and resources spent debugging.

R8.6 Per finding F8.11, incorporate integration testing to ensure all functions and scripts are working together as intended and there are no conflicts or errors between different code units.

Severity Level: Medium – without integration testing, there is no assurance that all the functions and scripts are working together correctly. If any errors exist, without integration testing, the team will be less efficient at debugging and will spend time extra time and resources fixing errors.

R8.7 Ensure all python functions have docstrings – per finding F8.10. This will ensure that all functions are correctly documented, and definitions, descriptions and decision point reasoning are captured. Docstring best practice for a function include a brief description of what the function is and what it's used for, any arguments that are passed, labeling which are required and which are optional, any restrictions on when the function can be called, or any exceptions that are raised.

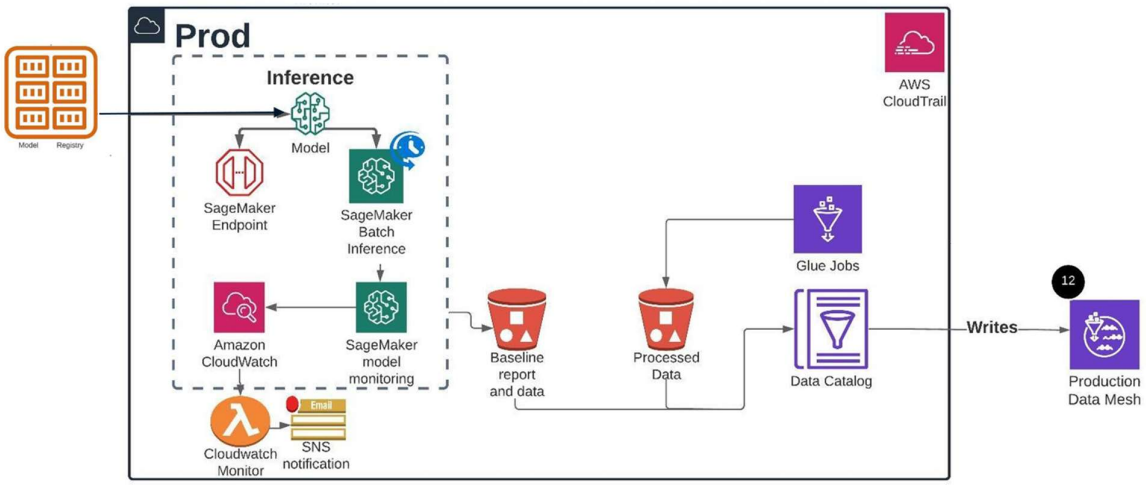
Severity Level: Low – this recommendation will not affect the model outputs but is a best practice to follow when writing code.

3.9 Production and Operations

In the production and operations stage, the purpose is to review the model's use in production by operational groups. This ensures that the model is utilized for the purpose that was intended such that maximum benefit is being realized. This process involves ensuring that the technical measures are in place, the solution stays accurate, predictable, and without drift. Also, maintenance and updates to the model are completed as necessary. Additionally, to confirm the model is being utilized for the purpose it was created.

Figure 4 shows the architecture diagram for the Production instance of the WiNGS-Ops model in Amazon Web Services (AWS).

Figure 4: WiNGS-Ops Production Architecture Diagram



3.9.1 Findings

F9.1 The Advanced Analytics team is made up of external resources, which means that the WiNGS-Ops model inference is reliant on non-permanent employees. This poses a risk as the continued successful operation is contingent on contractual arrangements being followed and maintained.

F9.2 Cloud services are utilized from each of the three main cloud vendors. Azure DevOps is utilized for its file repository, version control, and project tracking. AWS provides the infrastructure for the solution itself. And the Google product Apigee is used for the API management services.

F9.3 Although each granular model has been documented by the Advanced Analytics team, there is no documentation of the overall inference pipeline. This means the knowledge of this deployment process lies with the team itself which poses a risk.

F9.4 The output of WiNGS-Ops is currently an overall number which does not allow for a clear understanding of the specific drivers of that modeling output. This may lead to confusion as to what the model score represents and could lead to loss of trust or credibility in the model.

F9.5 Limitations of the WiNGS-Ops model are not well communicated to the business users. This means they may not be best equipped to fully understand and utilize the output of the WiNGS-Ops model.

F9.6 The SDG&E members of the WiNGS-Ops Data Science and IT teams hold significant experience and knowledge of the models, data, and systems. This knowledge base and experience is not captured in documentation to ensure consistency and maintainability into the future.

F9.7 The current owner of the Weather Sanitization repository does not have the admin rights to approve and push changes to the master branch. They are currently using a sub-branch for production.

F9.8 In the Inference pipeline, the spans are mapped to the nearest weather station on Euclidean distance. This does not align with the Meteorology team's weather stations to device associations which aims to accommodate key drivers such as topography.

F9.9 Some segments in the production model do not have a value output from the WiNGS-Ops model available for use in the production version. There is no communication or explanation provided for these segments at this time.

F9.10 All snapshots and stored data are stored using the default cloud storage type.

F9.11 Monitoring is completed using email alerts and ad-hoc log analysis. There is currently no monitoring dashboard or advanced monitoring in place.

F9.12 Up to ten percent of feeders will not have a result in production due to data quality issues, such as invalid Global IDs. The impact of this is not communicated to end users.

F9.13 There is currently no model representation in the EOC to aid understanding of the WiNGS-Ops outputs and limitations in an activation event.

3.9.2 Recommendations

R9.1 Per finding F9.1, recommendation to ensure there is a skilled and knowledgeable base of internal resources involved in each aspect of the WiNGS-Ops modeling process such that a full reliance on external parties is not maintained.

> *Severity Level: Low* – the Advanced Analytics team is skilled and knowledgeable so there is minimal risk to the model outputs at this stage.

R9.2 In future, it may be preferential to consolidate services under one cloud provider for ease of use, integration, and billing, per finding F9.2. This can ensure that future updates to any of the cloud services are always made in a way to keep compatibility and seamless integration with the other developed components.

> *Severity Level: Low* – this recommendation has no impact on the output of the WiNGS-Ops model but would allow for greater efficiency in use of cloud services. Although cloud services may work together across different vendors, they are optimized to work most effectively when combined with services belonging to one single cloud provider.

R9.3 Per finding F9.3, would recommend creating robust and granular documentation of the deployment pipeline which would ensure a lower reliance on the experience of resources.

> *Severity Level: Medium* – without this documentation, a continued reliance on external resources would be mandatory as there would be no straightforward mechanism through which SDG&E resource could inform themselves on the finer details of the inference pipeline.

R9.4 Per finding F9.4, would recommend for the key drivers of the modeling output to be exposed to the users, such that they may gain a greater understanding of the outputs and some indication on how this output should be viewed and utilized.

> *Severity Level: Low* – this detail may allow for greater understanding and trust in the WiNGS-Ops output.

R9.5 Per finding F9.5, would recommend creating documentation of the limitations of the models which underpin the WiNGS-Ops outputs and ensure that these are fully understood by the business users. This will ensure that any decisions being made on the result of the WiNGS-Ops model are made from the most informed position.

> *Severity Level: Medium* – without this understanding of the limitations of the model, sub-optimal decisions may be made due to a misinterpretation of the results.

R9.6 Per finding F9.6, would recommend that documentation is completed for the full lifecycle of each model in training and in inference such that the knowledge, skills and experience of the team is captured for future use. This would also enable training and onboarding of new resources in future to be more straightforward and regulatory filings to be completed more swiftly. Example

pieces to include in this documentation are the problem formulation process, documenting all decision points and reasonings and documenting future plans and intentions.

Severity Level: Low – the team is knowledgeable in the models they have constructed so any risk is reduced. In most cases there is only one team member with discrete knowledge of the specific model.

R9.7 Per finding F9.7, we recommend updating the technical ownership of the weather sanitization repository, and any other repositories that may have changed ownership.

Severity Level: Medium – the script is well understood by multiple parties so there is minimal risk with this, however there is no single owner to drive decisions or improvements.

R9.8 Per finding F9.8, On the inference side, implement the device to weather station associations that the Meteorology team determined based on topographical features into the weather station mapping. This will ensure the most suitable weather station data is used for each segment.

Severity Level: Medium – this has the potential to produce quite skewed results if there is a significant topographical impact on certain spans.

R9.9 Per finding F9.9, would recommend that data issues are corrected such that all segments have an outputted value from the WiNGS-Ops model. Failing that, there should be full communication and explanation to be provided to the end users for those segments where a WiNGS-Ops output was unable to be generated. This would ensure that awareness of these missing values is gained, and decisions are not based on the omission of those segments in the model outputs.

Severity Level: Medium – while the PSPS de-energization decision takes other inputs aside from WiNGS-Ops, without a complete model output for every segment, it is conceivable that the decision maker will lose trust with WiNGS-Ops model if a PSPS decision would need to be made for a segment that has no WiNGS-Ops output.

R9.10 Per finding F9.10, would consider use of cold storage for long-term storage of snapshots or model runs which do not need to be accessed regularly. This would reduce the overall costs of the cloud infrastructure and become more important as the models and data sets mature and grow in size.

Severity Level: Low – as the size of files being stored currently is not large, use of cold storage would have a minimal effect on the cost of cloud services though remains a best practice recommendation.

R9.11 Per finding F9.11, would recommend developing a monitoring dashboard which may provide real-time error monitoring and view of the model runs such that issues may be highlighted and resolved in a timely manner.

Severity Level: Low – existing monitoring allow for errors to be identified and not be missed, advanced monitoring would allow a more streamlined process to error identification and remediation.

R9.12 Per finding F9.12, would recommend cleaning the data such that all Global IDs are valid and the number of feeders without output results due to invalid global IDs decreases. This will prevent situations where the WiNGS-Ops model is unable to produce risk scores.

Severity Level: Medium – having up to 10% of feeders without risk scores could cause the WMP to lose credibility within the organization when the model is needed to provide data driven insights for PSPS decision making.

R9.13 Per finding F9.13, would recommend a new role be created in the EOC to provide WiNGS-Ops model support. This person would be knowledgeable about all aspects of the model, outputs, limitations, and the impact on the other components utilized in EOC decision-making.

Severity Level: Low – without this role in the EOC, the model may not be fully understood so model outputs may be interpreted incorrectly. This could lead to sub-optimal decisions being made.

3.10 Feedback & Learning

In the feedback and learning stage, the purpose is to assess how model errors or faults are communicated and updates are executed in the model's continued development, maintenance, and feedback loop. This ensures model adjustments and maintenance plans are correctly implemented. This process involves assessing whether there is an appropriate level of feedback back into the models, continuous and iterative development is in place to improve the models and understanding how models are adjusted and maintained.

3.10.1 Findings

F10.1 Sub-models are only retrained and adjusted once per year because of the nature of the data used to train the models. Ignitions are not common occurrences, so retraining during the fire season is viewed as not needed.

F10.2 The operational model is still new and therefore has not had any issues, however, there is no formalized process on reporting issues.

F10.3 The Advanced Analytics team thoroughly tracks their model adjustments and edits with Azure DevOps Wiki pages.

F10.4 Static log analysis is built into the pipeline and can be viewed in Sagemaker or CloudWatch.

F10.5 Feedback between end users and the development team is verbally communicated during weekly meetings. There is no formal documentation process to capture discussions, feedback and track ownership and status of tasks.

F10.6 Any questions that the end users have about the model outputs are generally covered through informal emails and regular check-ins between the teams. This means that the end users have ready access to the model experts and have their questions answered in a timely manner, however these communications and decisions are not formally captured.

F10.7 There is no overall model versioning present which ensures any minor change to a sub-model is reflected as a larger overall model version change.

3.10.2 Recommendations

R10.1 Per finding F10.2, create a formalized process for issue reporting from the end users to the development teams. This should be simple and streamlined such that any issues may be raised, quantified, and remediated quickly.

> *Severity Level: Low* – there is no prescribed process currently which could lead to confusion as to the point of escalation for issues. This may result in a delay to any remediation activity and impact the quality of outputs.

R10.2 Per finding F10.5, each meeting should be formally documented, and actions / tasks should be added to a backlog that may be prioritized, tracked, and completed against. This will ensure that all tasks are captured and implemented as intended and miscommunication is avoided.

> *Severity Level: Low* – without a formalized process of documentation and action tracking, there may be more instances of misunderstanding of intention between teams, which might result in a sub-optimal outcome or re-work in remediating the concern.

R10.3 Per finding F10.6, create a formalized process for questions and model changes / tweaks ahead of each activation event. Changes to model code and outputs should be tracked through formal version control. This will mean that the decision points and actions taken are formally documented and easily explainable if a reference is required. This may aid answering regulatory questions or post-event report preparation.

> *Severity Level: Low* – the current process will result in a more time-consuming post-activation event reporting process. This may mean a period of potential re-work to

establish the reasoning behind certain tweaks and decisions taken in the model pre-event.

R10.4 Per finding F10.7, create an overall WiNGS-Ops model versioning process such that any change or update to any component of WiNGS-Ops results in a version iteration. This ensures that users have a clear indication of when a model methodology has changed. This may help the users understand which models may be easily compared.

Severity Level: Low – the current versioning methodology may result in inaccurate comparisons being made by end users across models.

Appendix A: Table of Recommendations

ID	Recommendation Name	Description	Severity Level
R1.1	Model Approach Standardization	Model Approach Standardization: As most of the model build completed by the team is independent, there is a potential lack of standardization for the development, training, testing and validations of models. There is some standardization in the use of repos and feature extraction, would recommend expanding this standardization to all aspects of model development so that all models are tested and validated to the same specification.	*Severity Level: Low* – without a standardized approach, each model may not hold the same level of credibility given varying levels of testing and validation. This would improve consistency of model outputs.
R1.2	Internal Model Review Process	Internal Model Review Process: Best practice data science solution development must include a level of peer-review to validate the scripts that are developed and operated. Creation of a more formalized internal model review process would provide a forum through which ideas may be discussed and considered before implementation, and through which a robust and consistent approach to model review may be performed.	*Severity Level: Medium* – this would enable potential improvements or ideas to be highlighted and discussed leading to more effective and efficient models.
R1.3	Model Documentation	Model Documentation: As the team has been operating in a reactive state to changes in the WMP guidelines and recommendations, full documentation of each of the models is not complete. To reduce any risk due to reliance on the experience and knowledge of the individual team members, the recommendation would be for a concerted effort to ensure documentation is complete for each of the latest model versions to be released for fire season 2023.	*Severity Level: Low* – without robust model documentation, there is a reliance on the experience and memory of the members of the team to explain the reasoning behind model decisions and changes.
R1.4	Team Enhancements	Team Enhancements: The team consistently faces capacity constraints due to the ever-changing landscape of the WMP guidelines and recommendations, coupled with continued regulatory requests for data and information. As such, the team operates reactively to requests and priorities, without a true backlog of tasks captured and delivered against. The recommendation would be to	*Severity Level: Medium* – without changes to the team size and roles of team members, the full potential of the members of the team may not be realized. Improved team size, capability and demand management would allow for a more optimal environment, within which the greatest value may be generated.

ID	Recommendation Name	Description	Severity Level
		enhance the team in a couple of different ways. A scrum master can help generate and manage a backlog of tasks and activities such that activities may be prioritized, and a demand management process may be created. The addition of a data analyst to the team could assist with external regulatory data requests, alleviating some of the time demands of the WiNGS-Ops Data Science team.	
R1.5	Data Owner Communication	Enhance communication with data owners: Ensure that there is an integrated function, such that communication from specific business owners for each data input into the models is cohesive and timely. This would ensure definitions, use, bounds for validity and decisions on changes that would be needed are communicated. They would also be responsible for ensuring that the data is up to date and accessible to all who may need it.	*Severity Level: Medium* – lack of communication from data owners may result in unexpected changes and diminished data integrity. The data owner is accountable for the use, quality, and protection of a dataset.
R1.6	Calculation Ownership	Calculation ownership: Owners of the specific constants (e.g., PSPS risks) and calculation methodologies sometimes called “information" such that their definitions and approaches are agreed, documented and uniform across the business. This is to ensure that any colloquial terms used for aggregated data assets are consistent such that an output like “miles of span in HFTD in one group’s calculation is the same as another’s.	*Severity Level: Low* – a calculation owner will be accountable for ensuring calculation methodologies are clearly defined and are used appropriately and consistently.
R1.7	Model Ownership	Broader model ownership in the form of a board / group with regular meeting cadence to agree higher-level changes and adjustments, reviewing output of sensitivity analysis and changes prior to implementation. This would ensure that the responsibility for driving the direction of overall model enhancements and improvements is agreed amongst the Developers, Wildfire Mitigation team and the Business users.	*Severity Level: Low* – without regular communication between all stakeholders, the direction and prioritization of model development and improvements can be missed.

ID	Recommendation Name	Description	Severity Level
R1.8	EAMP Data Experts	External EAMP/Asset 360 data experts: EAMP/Asset 360 provides a rich asset data source used in modeling. The data itself is a clean and curated version of GIS and Asset Management data. The program is operated by external contractors who also remain as the data source SMEs. The recommendation would be for an internal SDG&E team to be onboarded to share some of the responsibility as SME for the data source. The source, including all dictionaries and implemented manipulations, should also be fully documented such that any new user may easily gain a complete understanding of the data and its use.	*Severity Level: Medium* – with a continued reliance on external parties for this critical data source, the SDG&E team will not gain full ownership, understanding and control over the underlying data. Internal SDG&E SME expertise in the data source will ensure a robust and future-proof mechanism for data understanding, questions, and data updates.
R1.9	External Inference Team	External Inference team: The development team responsible for the inference aspects of the WiNGS-Ops are a group of external contractors. Feedback has been very positive on the capability and success of the team and the models seem to operate sufficiently well. The team is effective in the conversion of models from training and test phase to inference phase but do not look to challenge the training team to improve the models. Given the reliance on external contractors in this team, the recommendation would be to integrate more SDG&E resources into the team so that some of the knowledge and experience is internal.	*Severity Level: Low* – as the WiNGS-Ops model continues to mature and gain complexity, the technical debt on the external development members of the Advanced Analytics team will grow, increasing this reliance.
R2.1	OIR Requirements	Per finding F2.2, in order to ensure that all OIR requirements are met and prevent possible violations, we recommend building and maintaining a formalized report that tracks the requirements and how they were carried out. Having this existing documentation will not only confirm what the requirements are and if and how they were completed but will also be ready to pass along to the OIR as appropriate.	*Severity Level: Low* – this will help prevent potential violations from the OIR by tracking all the requirements and how they were completed.

ID	Recommendation Name	Description	Severity Level
R2.2	Model Change Documentation	Per finding F2.4, create a formal process through which requirements for model changes are captured, tracked, and completed against. This will ensure that the changes have been understood and captured correctly and will allow a definition of done to be assessed against, by the end users in their approval of model changes. We suggest using a project tracking tool such as Boards in Azure DevOps since it is already used by the team.	*Severity Level: Low* – without a documented process, requirements and requested changes may be incorrectly implemented or the end users may not have an easy mechanism for change approval.
R2.3	Model Value	Per finding F2.5, in order to determine the value, the model is bringing to the business, we recommend establishing metric(s) to gauge the effectiveness. This will ensure that the impact of model improvements and developments over time may be quantified and tracked.	*Severity Level: Low* – this recommendation will increase end user buy in and understanding to the changes that are enacted in the model.
R2.4	Initiation Stage Documentation	Per finding F2.7, we recommend documenting the initiation stage in order to capture critical elements of the initial planning stage. This includes defining what problem this model will solve, what is the feasibility of this model, who are the end users and how do they want to ingest the model outputs, who are the subject matter experts and what is their ability to participate in the model development, who will be the business owner of the model, what are the initial assumptions and how were they determined, and confirmation that all relevant business areas have taken full sponsorship of the project. Additional details on why certain decisions were made with respect to the model generation are also critical to document in the initiation process. Going forward, with the initiation of new model versions, we recommend documenting these critical elements so they can be referenced by future developers and users of the model.	*Severity Level: Low* – without this documentation in place, future developers and end users may have a more difficult time understanding the decisions and assumptions that were made, which SMEs to turn to for input, how the model will be measured for success, or the original problem and objectives.

ID	Recommendation Name	Description	Severity Level
R3.1	Automated Data Validation	Per finding F3.2, every data input should pass through some degree of automated data validation check to look for outliers, errors, text control, contradictions, etc. Each of these validation checks should have associated documentation that includes what to do when data is missing or anomalous. This should be implemented in the inference pipeline and should be consistent with data validation performed by the WiNGS-Ops data science team during their EDA process.	*Severity Level: Medium* – there is currently a lot of reliance on source data owners to validate their data, however, if erroneous data makes its way into the model during the model development or inference pipeline, inaccurate outcomes will result. Poor data quality can therefore lead to poor model outcomes, which will result in a loss of trust in the model by the end users.
R3.2	Pole and Span Imputation	Per finding F3.3, we recommend working with the GIS team to develop a logic-based solution for imputing pole location information using other fields when historical pole locations are missing. This may include utilizing an existing GIS redlining process for resolving these gaps.	*Severity Level: Low* – this would ensure that the data used in modeling is most representative of SDG&E's network. It may also help reduce the number of minority class records that are dropped due to missing data.
R3.3	Network As Switched Limitation	Per finding F3.5, we recommend noting this as a limitation of the model and prior to PSPS activation events that the systems are restored to the as-designed states wherever possible. However, we also recommend contacting Operations personnel to confirm the correct owner of the network as-operated electrical connectivity data since this data is a critical component of the WiNGS-Ops model. Additionally, seeking out information on the root data source, how it is validated, and the existing assumptions are critical for ensuring a complete understanding of the data and its correct use.	*Severity Level: Low* – without knowing the correct data owner or who to reach out to with concerns or data issues, there will be continued uncertainty of the data and of the stewardship and accountability surrounding that data.
R3.4	Data Object Governance	Per finding F3.7, would recommend for greater governance and controls to be put in place for each of the data objects utilized by WiNGS-Ops such that none of the data created for and used in the models is inadvertently used for a different purpose, generating alternative and incorrect views of the landscape.	*Severity Level: Low* – although this may not directly impact the output of the WiNGS-Ops model, it may affect the credibility of the data sources used if the source is used incorrectly elsewhere.

ID	Recommendation Name	Description	Severity Level
R3.5	SAIDIDAT Data Ingestion	Per finding F3.9, We recommend a direct query of SAIDIDAT data from its source database. This eliminates the reliance on an individual and prevents potential human error in the process.	*Severity Level: Low* – manual data request and transfers are reliant on the requestor to ask for the information. Automating the request process may be a better way to obtain updated outage history data on a scheduled basis rather than on an as-requested basis.
R4.1	Feature Removal	Per finding F4.1, for the models that do not have auto regularization, we recommend removing the less relevant features as measured by the feature importance function outputs. Removing less relevant features will help with the stability of the model, avoid overfitting, and reduce computation cost.	*Severity Level: Medium* – it is unclear at this stage the impact that inclusion of these unimportant features has on the outputs. It has the potential to skew results which may have a large impact, so has been rated as such.
R4.2	Alternative Land Use Data Source	Per findings F4.2 and F4.3, would recommend working closely with the SANGIS team to incorporate SDG&E territory currently not covered in their existing coverage data, as well as request for a more frequent than annual data updates. This would ensure the models have access to the same information as the rest of San Diego county, and is most up to date during a red flag warning event.	*Severity Level: Low* – models run on data which has not been refreshed for a while or on imputed data based on mean values may provide inaccurate outputs. This may cause a model to under-represent the potential consequence of an ignition due to a missing at-risk land use.
R4.3	Model Improvement Limitations	Per finding F4.5, due to the time pressures and resource constraints, the team does not have capacity to further improve models with development or incorporation of additional features.	*Severity Level: Low* – the impact of this would be minimal due to the models' existing satisfactory performance but might represent a missed opportunity for continued model improvements and enhancement.
R4.4	Safety Weights Documentation	Per finding F4.8, would recommend creation of a documented framework to define the safety weights used in the PSPS model such that there is an explainable process through which they may be assessed and updated based on additional SME input. These weights must also be integrated into version control, so that any changes are managed and easily tracked, model version to model version. Documentation around this would help provide future model developers and users better understand why certain values were used, and what the historical justifications and rationale were.	*Severity Level: Low* – without a clearly documented process for suggesting changes to the weights and version control to track those changes, it may be difficult to provide explanatory evidence in support of decisions driven by this model.

ID	Recommendation Name	Description	Severity Level
R5.1	Class Imbalance Approaches	Per finding F5.1, we recommend testing other approaches to handling class imbalanced data, including up-sampling, SMOTE, and ADASYN, in order to determine the most applicable method for each model.	*Severity Level: Medium* – down-sampling excludes significant amounts of data which may result in an unrepresentative data sample being used for training and testing the model.
R5.2	Algorithm Testing	Per findings F5.2, F5.4 and F5.8, we recommend testing other algorithms to ensure that the most suitable algorithm is used to solve the problem, balancing complexity of understanding and training with accuracy of modeling outputs.	*Severity Level: Low* – without validating that there isn't a more suitable algorithm for the model, the team cannot be certain that they have built the most suitable model for the specific application. Testing alternative algorithms may help build greater trust in the end product.
R5.3	Collaborative Model Development and Release	Per finding F5.3, we recommend a more collaborative approach towards model development and release. A peer-reviewed approval process (similar to the one used by WiNGS-Planning) can ensure consistency between sub-models and the best practices are followed.	*Severity Level: Medium* – individual working may lead to inconsistencies between models resulting in models with differing levels of robustness being deployed.
R5.4	Conductor Model Retrain	Per finding F5.5, we recommend retraining the conductor model based on data from 2015 to present, utilizing the 2022 data for testing and validation. This will ensure the most representative data is utilized in construction and training to create the most accurate and useful modeling outputs.	*Severity Level: Medium* – based on the most recent data used for validation, the model under-represented the potential risk due to conductor failure. Re-training this model would generate a more representative output.
R5.5	Same Data Sources	Per findings F5.6 and F5.7, would recommend that the models are trained on the same data sources which would be utilized for inference in production such that the resulting outputs are most relevant and applicable.	*Severity Level: Medium* – as the models were trained on different source data, the data relationships learnt may not be representative of what would be seen in the EOC. As a result, the outputs of the models may not be as accurate as if the data used for training was the same source as used in inference.
R5.6	GIS Cleaning	Following on from R5.5 and in relation to finding F5.7, consider a larger program of GIS data cleaning, validating and improvement and investigate if existing GIS red lining processes can be leveraged to ensure the GIS system of record for assets represents the most accurate view of assets in SDG&E's territory. This would ensure that any modeling application or activation event	*Severity Level: Low* – it is critical that decisions in the EOC are made based upon the most accurate representation of the assets in the field. Ensuring this data source is accurate and up to date is crucial to enabling the EOC to operate from an informed position.

ID	Recommendation Name	Description	Severity Level
		would consider that most accurate understanding when making data-driven decisions.	
R5.7	Hyper-parameter Tuning	Per finding F5.8, the foreign object model is using GridSearchCV for hyper-parameter tuning, however, it has not yet been implemented for the vehicle contact model. We recommend taking the same approach to tune the hyper-parameters.	*Severity Level: Low* – consistent use of techniques across models ensures that the quality and robustness of each model is uniform and contributes to an optimal output.
R5.8	SHAP	Per finding F5.2, with explainability as a key driver for model selection, we recommend looking to incorporate SHAP to help explain the output of the models through calculating the contribution of each feature to the model output. These values can be used to understand the importance of each feature and to explain the results of the model.	*Severity Level: Low* – without a full understanding of the importance and contribution of the features in a model, the driving factors of the model's outputs are less explainable.
R6.1	Brier Score	Per finding F6.2, would recommend using the full Brier score such that the outputs are unaffected by population size. This will enable Brier scores to be compared across different versions of a model to allow model improvements to be validated.	*Severity Level: Low* – the risk of not using full Brier score is that this modified score might be inadvertently used to compare models with different sample sizes. This would give an inaccurate view of the performance comparison so could result in an incorrect modeling decision.
R6.2	Class Imbalance Validation Methodology	Per finding F6.3 and F7.3, for the vehicle contact model, we recommend incorporating a nested cross validation where one-fold is an out-of-period imbalanced data split for the final validation, and other fold is split for training and testing on balanced sampled data set. This would provide an additional method for validating the accuracy of the model. Ensure the right metric is used for the evaluation, as some metrics are better for evaluation when there is class balance (ROC AUC), and others are better for when there is class imbalance (Precision-Recall AUC).	*Severity Level: Medium* – validating the imbalanced data with this approach checks performance of the model on the real class distribution.
R6.3	Uniform Model Testing	Per finding F6.4, establish a consistent and agreed approach for model testing across the team such that each member may be sure of the optimal model and be agreed when training is complete. This will	*Severity Level: Low* – models may have differing levels of robustness without a uniform defined and agreed approach to testing.

ID	Recommendation Name	Description	Severity Level
		ensure consistency across models and build credibility with the end users.	
R6.4	Data Documentation	Per finding F6.5, we recommend for all data that is ingested into the models that detailed documentation is provided by the data owners. The documentation should contain pertinent information such as the data owner, data collection methodology, data dictionary, structure of the data, data validation and quality assurance steps taken, data manipulations from the raw data, and confidentiality, access and use conditions. This will ensure a detailed understanding of the data that can be reference as needed, critical for ground truth data.	*Severity Level: Low* – without detailed documentation, there is a risk the data can be misinterpreted, or if there are turnover or new hires on the WiNGS-Ops Data Science or Advanced Analytics teams, they may have a more challenging time referencing and understanding the data inputs.
R7.1	Back-casting Model Validation Process	Per finding F7.1, a more holistic and reliable model validation process may be created to allow automated back-casting for each model change. This would allow for greater confidence in the updated version of each model to be gained. Given the snapshots of data are now maintained in the cloud, this ensures that this process would be simpler to perform.	*Severity Level: Low* – without an automated and uniform approach to model output validation, validating each new model release will be a time-consuming and inconsistent process.
R7.2	Back-casting Data Capture	Per finding F7.1, we recommend ensuring that all necessary data and calculation components are captured, including the network configuration, at the time of a PSPS event to help streamline future back-casting exercises.	*Severity Level: Low* – implementing this would allow for the automated and uniform approach mentioned above could be enacted for model back-casting.
R7.3	End User Formalized Validation Process	Per finding F7.2, establishing a formalized validation process by the end users will establish consistency in the validation approach, and also build credibility with OEIS by demonstrating the results are reviewed in a specific and systematic way.	*Severity Level: Low* – without a formalized validation process, there is the potential for end users to validate the model differently every time a new model version is released. This may result in missing an important check or reviewing an output that differs from a previous model version.

ID	Recommendation Name	Description	Severity Level
R8.1	Centralize Models	Per finding F8.2 and F8.3, Recommendation to migrate the conductor training model and PSPS model scripts to Azure DevOps Repos. This will ensure development on local machines are version controlled, tracked appropriately, and accessible by the rest of the team. This will also allow models to leverage cloud compute capabilities, meaning that more advanced models may be produced. Additionally, the PSPS model should be passed to the inference team such that the entire WiNGS-Ops model can be executed through the inference pipeline.	*Severity Level: Medium* – current processes limiting version control and access could introduce errors and confusion in the correct version that should be run in production. Full cloud migration would limit the risk of this issue.
R8.2	Model Training Process Explanation	Per finding F8.5, the model training team should provide a more thorough explanation of the model training process and decisions which would enable the Advanced Analytics team to have a better grounding for implementing the code. As well as education sessions, thorough documentation would enable any new team members to be onboarded swiftly.	*Severity Level: Low* – without full understanding and knowledge of the model training process, the Advanced Analytics team may not be able to add as much value in critiquing and improving the models.
R8.3	Combine Pole and Span Ignition Models	Per finding F8.7, would recommend that the pole and span ignition models are combined to remove any overlaps which might exist in the separate models.	*Severity Level: Medium* – currently the models are not fully independent, which may skew the results. This should be rectified such that an accurate representation of risk may be generated.
R8.4	Profiler	Per finding F8.8, consider running a profiler to help understand the resource consumption of the various operations in the model. This can potentially resolve performance bottlenecks and help the model execute faster.	*Severity Level: Low* – this recommendation does not affect the model output but may improve the runtime performance of the model.
R8.5	Unit Testing	Per finding F8.9, incorporate unit testing to ensure all functions are performing as intended and more easily isolate errors when they occur. Unit tests also check that the code still functions as expected after making changes which builds code stability.	*Severity Level: Medium* – Without unit testing, there is no assurance that code will function correctly, and undiscovered bugs may exist. This can lead to poor quality modeling results and wasted time and resources spent debugging.

ID	Recommendation Name	Description	Severity Level
R8.6	Integration Testing	Per finding F8.11, incorporate integration testing to ensure all functions and scripts are working together as intended and there are no conflicts or errors between different code units.	*Severity Level: Medium* – without integration testing, there is no assurance that all the functions and scripts are working together correctly. If any errors exist, without integration testing, the team will be less efficient at debugging and will spend time extra time and resources fixing errors.
R8.7	Docstrings	Ensure all python functions have docstrings – per finding F8.10. This will ensure that all functions are correctly documented, and definitions, descriptions and decision point reasoning are captured. Docstring best practice for a function include a brief description of what the function is and what it's used for, any arguments that are passed, labeling which are required and which are optional, any restrictions on when the function can be called, or any exceptions that are raised.	*Severity Level: Low* – this recommendation will not affect the model outputs but is a best practice to follow when writing code.
R9.1	Internal Resources Embedded into Each Team	Per finding F9.1, recommendation to ensure there is a skilled and knowledgeable base of internal resources involved in each aspect of the WiNGS-Ops modeling process such that a full reliance on external parties is not maintained.	*Severity Level: Low* – the Advanced Analytics team is skilled and knowledgeable so there is minimal risk to the model outputs at this stage.
R9.2	Cloud Consolidation	In future, it may be preferential to consolidate services under one cloud provider for ease of use, integration, and billing, per finding F9.2. This can ensure that future updates to any of the cloud services are always made in a way to keep compatibility and seamless integration with the other developed components.	*Severity Level: Low* – this recommendation has no impact on the output of the WiNGS-Ops model but would allow for greater efficiency in use of cloud services. Although cloud services may work together across different vendors, they are optimized to work most effectively when combined with services belonging to one single cloud provider.
R9.3	Pipeline Deployment Documentation	Per finding F9.3, would recommend creating robust and granular documentation of the deployment pipeline which would ensure a lower reliance on the experience of resources.	*Severity Level: Medium* – without this documentation, a continued reliance on external resources would be mandatory as there would be no straightforward mechanism through which SDG&E resource could inform themselves on the finer details of the inference pipeline.

ID	Recommendation Name	Description	Severity Level
R9.4	Modeling Key Drivers	Per finding F9.4, would recommend for the key drivers of the modeling output to be exposed to the users, such that they may gain a greater understanding of the outputs and some indication on how this output should be viewed and utilized.	*Severity Level: Low* – this detail may allow for greater understanding and trust in the WiNGS-Ops output.
R9.5	Limitations Documentations	Per finding F9.5, would recommend creating documentation of the limitations of the models which underpin the WiNGS-Ops outputs and ensure that these are fully understood by the business users. This will ensure that any decisions being made on the result of the WiNGS-Ops model are made from the most informed position.	*Severity Level: Medium* – without this understanding of the limitations of the model, sub-optimal decisions may be made due to a misinterpretation of the results.
R9.6	Full Model Lifecycle Documentation	Per finding F9.6, would recommend that documentation is completed for the full lifecycle of each model in training and in inference such that the knowledge, skills and experience of the team is captured for future use. This would also enable training and onboarding of new resources in future to be more straightforward and regulatory filings to be completed more swiftly. Example pieces to include in this documentation are the problem formulation process, documenting all decision points and reasonings and documenting future plans and intentions.	*Severity Level: Low* – the team is knowledgeable in the models they have constructed so any risk is reduced. In most cases there is only one team member with discrete knowledge of the specific model.
R9.7	Weather Sanitization Ownership Update	Per finding F9.7, we recommend updating the technical ownership of the weather sanitization repository, and any other repositories that may have changed ownership.	*Severity Level: Medium* – the script is well understood by multiple parties so there is minimal risk with this, however there is no single owner to drive decisions or improvements.
R9.8	Weather Station Imputation Mapping	Per finding F9.8, On the inference side, implement the device to weather station associations that the Meteorology team determined based on topographical features into the weather station mapping. This will ensure the most suitable weather station data is used for each segment.	*Severity Level: Medium* – this has the potential to produce quite skewed results if there is a significant topographical impact on certain spans.

ID	Recommendation Name	Description	Severity Level
R9.9	Missing Data Outputs	Per finding F9.9, would recommend that data issues are corrected such that all segments have an outputted value from the WiNGS-Ops model. Failing that, there should be full communication and explanation to be provided to the end users for those segments where a WiNGS-Ops output was unable to be generated. This would ensure that awareness of these missing values is gained, and decisions are not based on the omission of those segments in the model outputs.	*Severity Level: Medium* – while the PSPS de-energization decision takes other inputs aside from WiNGS-Ops, without a complete model output for every segment, it is conceivable that the decision maker will lose trust with WiNGS-Ops model if a PSPS decision would need to be made for a segment that has no WiNGS-Ops output.
R9.10	Cold Storage	Per finding F9.10, would consider use of cold storage for long-term storage of snapshots or model runs which do not need to be accessed regularly. This would reduce the overall costs of the cloud infrastructure and become more important as the models and data sets mature and grow in size.	*Severity Level: Low* – as the size of files being stored currently is not large, use of cold storage would have a minimal effect on the cost of cloud services though remains a best practice recommendation.
R9.11	Error Monitoring Dashboard	Per finding F9.11, would recommend developing a monitoring dashboard which may provide real-time error monitoring and view of the model runs such that issues may be highlighted and resolved in a timely manner.	*Severity Level: Low* – existing monitoring allow for errors to be identified and not be missed, advanced monitoring would allow a more streamlined process to error identification and remediation.
R9.12	Global ID Cleaning	Per finding F9.12, would recommend cleaning the data such that all Global IDs are valid and the number of feeders without output results due to invalid global IDs decreases. This will prevent situations where the WiNGS-Ops model is unable to produce risk scores.	*Severity Level: Medium* – having up to 10% of feeders without risk scores could cause the WMP to lose credibility within the organization when the model is needed to provide data driven insights for PSPS decision making.
R9.13	WiNGS-Ops Support Position	Per finding F9.13, would recommend a new role be created in the EOC to provide WiNGS-Ops model support. This person would be knowledgeable about all aspects of the model, outputs, limitations, and the impact on the other components utilized in EOC decision-making.	*Severity Level: Low* – without this role in the EOC, the model may not be fully understood so model outputs may be interpreted incorrectly. This could lead to sub-optimal decisions being made.
R10.1	Issue Reporting Process	Per finding F10.2, create a formalized process for issue reporting from the end users to the development teams. This should be simple and streamlined such that	*Severity Level: Low* – there is no prescribed process currently which could lead to confusion as to the point of escalation for issues. This may result in a delay to any

ID	Recommendation Name	Description	Severity Level
		any issues may be raised, quantified, and remediated quickly.	remediation activity and impact the quality of outputs.
R10.2	Action & Tasks Log	Per finding F10.5, each meeting should be formally documented, and actions / tasks should be added to a backlog that may be prioritized, tracked, and completed against. This will ensure that all tasks are captured and implemented as intended and miscommunication is avoided.	*Severity Level: Low* – without a formalized process of documentation and action tracking, there may be more instances of misunderstanding of intention between teams, which might result in a sub-optimal outcome or re-work in remediating the concern.
R10.3	Questions and Model Changes Tracking	Per finding F10.6, create a formalized process for questions and model changes / tweaks ahead of each activation event. Changes to model code and outputs should be tracked through formal version control. This will mean that the decision points and actions taken are formally documented and easily explainable if a reference is required. This may aid answering regulatory questions or post-event report preparation.	*Severity Level: Low* – the current process will result in a more time-consuming post-activation event reporting process. This may mean a period of potential re-work to establish the reasoning behind certain tweaks and decisions taken in the model pre-event.
R10.4	WiNGS-Ops Overall Versioning Process	Per finding F10.7, create an overall WiNGS-Ops model versioning process such that any change or update to any component of WiNGS-Ops results in a version iteration. This ensures that users have a clear indication of when a model methodology has changed. This may help the users understand which models may be easily compared.	*Severity Level: Low* – the current versioning methodology may result in inaccurate comparisons being made by end users across models.

Appendix B: Stakeholders Engaged

Name	Team	AI Assurance Stage Engagements
Joaquin Sebastian Peral	Enterprise Risk Management	Initiation, Data Cleaning, Feature Extraction, Model Selection & Training, Model Testing, Model Validation, Release & Productize, Production & Operations
Phi Nguyen	Enterprise Risk Management	Initiation, Data Cleaning, Feature Extraction, Model Selection & Training, Model Testing, Model Validation
Fatemeh Aarabi	Enterprise Risk Management	Data Cleaning, Feature Extraction, Model Selection & Training, Model Testing, Model Validation
Ashely Llacuna	Wildfire Mitigation Strategy	Initiation, Model Validation
Nisha Menon	Wildfire Mitigation Strategy	Initiation, Model Validation
John Blatchford	Advanced Analytics	Initiation, Release & Productize, Production & Operations, Feedback & Learning
Alex Johnson	Advanced Analytics	Release & Productize, Production & Operations, Feedback & Learning
Payam Kavousi Ghahfarokhi	Advanced Analytics	Release & Productize, Production & Operations, Feedback & Learning
Kaitlyn Petronglo	Advanced Analytics	Release & Productize, Production & Operations, Feedback & Learning
Patrick Liebel	Wildfire Mitigation Strategy	

About PA.

We believe in the power of ingenuity to build a positive human future.

As strategies, technologies, and innovation collide, we create opportunity from complexity.

Our diverse teams of experts combine innovative thinking and breakthrough technologies to progress further, faster. Our clients adapt and transform, and together we achieve enduring results.

We are over 4,000 strategists, innovators, designers, consultants, digital experts, scientists, engineers, and technologists. And we have deep expertise in consumer and manufacturing, defense and security, energy and utilities, financial services, government and public services, health and life sciences, and transport.

Our teams operate globally from offices across the UK, Ireland, US, Nordics, and Netherlands.

PA. Bringing Ingenuity to Life.

Discover more at paconsulting.com and connect with PA on LinkedIn and Twitter.

Denver Office
PA Consulting Group Inc.
Suite 3550
1700 Lincoln Street
Denver
CO 80203
USA
+1 720 566 9920

This report has been prepared by PA Consulting Group on the basis of information supplied by the client, third parties (if appropriate) and that which is available in the public domain. No representation or warranty is given as to the achievability or reasonableness of future projections or the assumptions underlying them, targets, valuations, opinions, prospects or returns, if any, which have not been independently verified. Except where otherwise indicated, the report speaks as at the date indicated within the report.

paconsulting.com