RESEARCH REPORT

Semantic Early Warning of a Hydrate Production Line Event in Petrobras 3W WELL 25 using SF2 Suite R*T*

Francisco Webber | SF2 Systems GmbH | 18 May 2026

Data basis: Petrobras 3W Dataset, WELL-25, Class 8 Hydrate Production Line

Abstract

This report examines whether a Hydrate Production Line event in the Petrobras 3W dataset WELL-25 can be made visible with SF2 Suite RT as a semantic state shift before the manually annotated event timestamp. To do so, synchronized raw sensors from a single well were converted into a discrete state vector, semantically mapped in an SF2 Retina, and then evaluated via overlap measurements against baseline and fault fingerprints.

The investigated sequence comprises 722,192 records at 1 Hz sampling. The annotated fault status occurs on 2020-07-07 17:47:04. The SF2 overlap curve to the Hydrate Production Line fingerprint already shows a stable semantic deviation 6d 21:18:51 before the event; a strong signal with overlap ≥ 90 is stable for more than 60 seconds 13:40:34 before the event. At the same time, the baseline overlap curve remains high, indicating a locally continuous process development that is nevertheless drifting semantically toward the fault state.

Contents

1 Motivation and Delimitation from Conventional Methods 2
2 Description of the Input Data 3
3 Description of the Derived State Vector 5
4 Training of the Retina 7
5 Overlap Measurements 8
6 Analysis of the Overlap Curves 8
7 Advantages of the SF2 Method Compared with Traditional Methods 10
8 Further Procedure 11
9 Conclusion 11
References 11

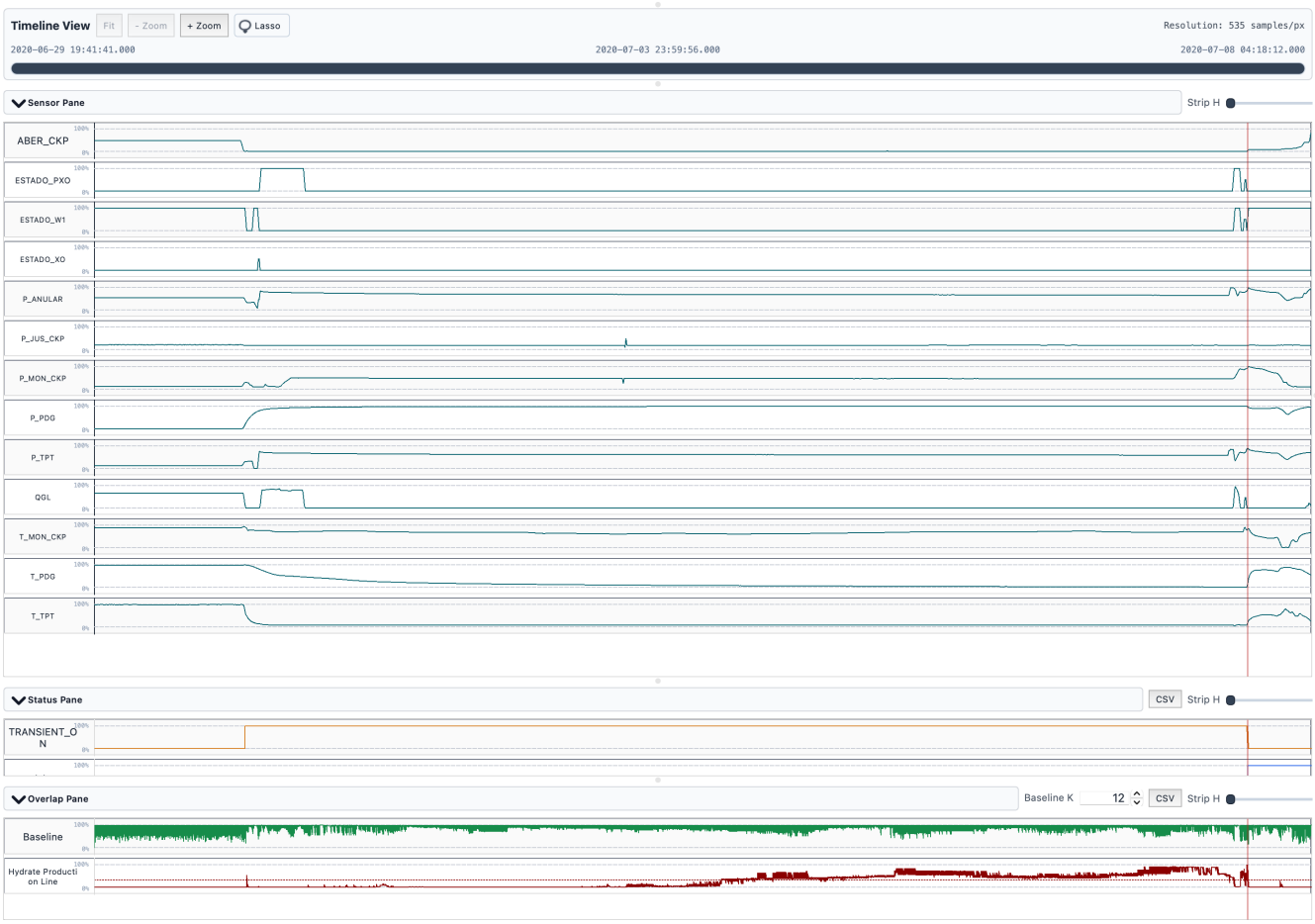


Figure 1 SF2 Studio view of the WELL-25 time series with synchronized raw sensors, expert annotation, and overlap curves. The red vertical line marks the transition to the annotated Hydrate Production Line special status. While the baseline similarity remains high over long stretches, the overlap with the hydrate fault fingerprint rises clearly already during the transient lead-in phase. This indicates a semantically detectable state drift before the actual event, without requiring domain interpretation of individual sensor channels.

1. Motivation and Delimitation from Conventional Methods

Hydrate formation is a classic flow-assurance problem in oil and gas production systems. Hydrates form under suitable pressure and temperature conditions in the presence of water and light hydrocarbons. This risk is especially relevant in offshore, subsea, and cold production systems, but it is not fundamentally limited to offshore wells.

Conventional technical approaches can be grouped roughly into three categories:

- **Threshold and SCADA logic:** operates with limits on individual sensors or simple combinations. These systems are robust and operationally established, but they often detect only states that are already technically conspicuous.
- **Physical flow-assurance models:** use thermodynamics, fluid data, geometry, heat transfer, and operating regime. They are strong from a domain perspective, but require expert parameters, model maintenance, and validation for each asset.
- **Data-driven ML/DL models:** learn patterns from historical data. They typically require a sufficient number of labeled examples, feature engineering, or robust cross-well generalization.

The SF2 method examined here follows a different approach. It does not model individual sensor values primarily; instead, it creates a semantic operating-state space from synchronized measurements. The relevant

research question is therefore not, “Can a classifier reproduce the expert status?”, but rather: Does the approach toward a critical state become visible as a change in system fingerprints before the status switch occurs?

2. Description of the Input Data

The experiment is based on the Petrobras 3W sequence WELL-00025_20200629194141.parquet from Class 8, Hydrate Production Line. From this, SF2-compatible CSV artifacts were generated:

Table 1 Input data artifacts

File	Role
well25_raw_sensor_in.csv	complete 1 Hz sensor stream, 722,192 records
well25_raw_sensor_train_200k.csv	completeness-filtered training set for Retina Core
PetroBras_status_timeseries.csv	synchronized annotations: status and TRANSIENT_ON
PetroBras_overlap_timeseries.csv	synchronized overlap curves: Baseline and Hydrate Production Line

Table 2 Core dataset metrics

Property	Value
Time range	2020-06-29 19:41:41 to 2020-07-08 04:18:12
Sampling	1 Hz
Full stream	722,192 records
Training set	200,000 records
Train status 0	162,131 records
Train status 1	37,869 records
Event start	2020-07-07 17:47:04
Event duration in file	10:31:09

The Petrobras 3W dataset was clearly assembled with a classic machine-learning focus: the goal was to make a large number of different anomaly periods and labeled fault examples available. For SF2, the weighting of data selection is different. The method benefits especially from sufficiently regular, consistent time series from a physical system, because they enable a stable semantic operating space to be built for each asset. WELL-25 was therefore selected for this use case: the sequence contains a large share of regular operating data in one continuous, temporally consistent segment and, in addition, the transition into a relevant Hydrate Production Line event.

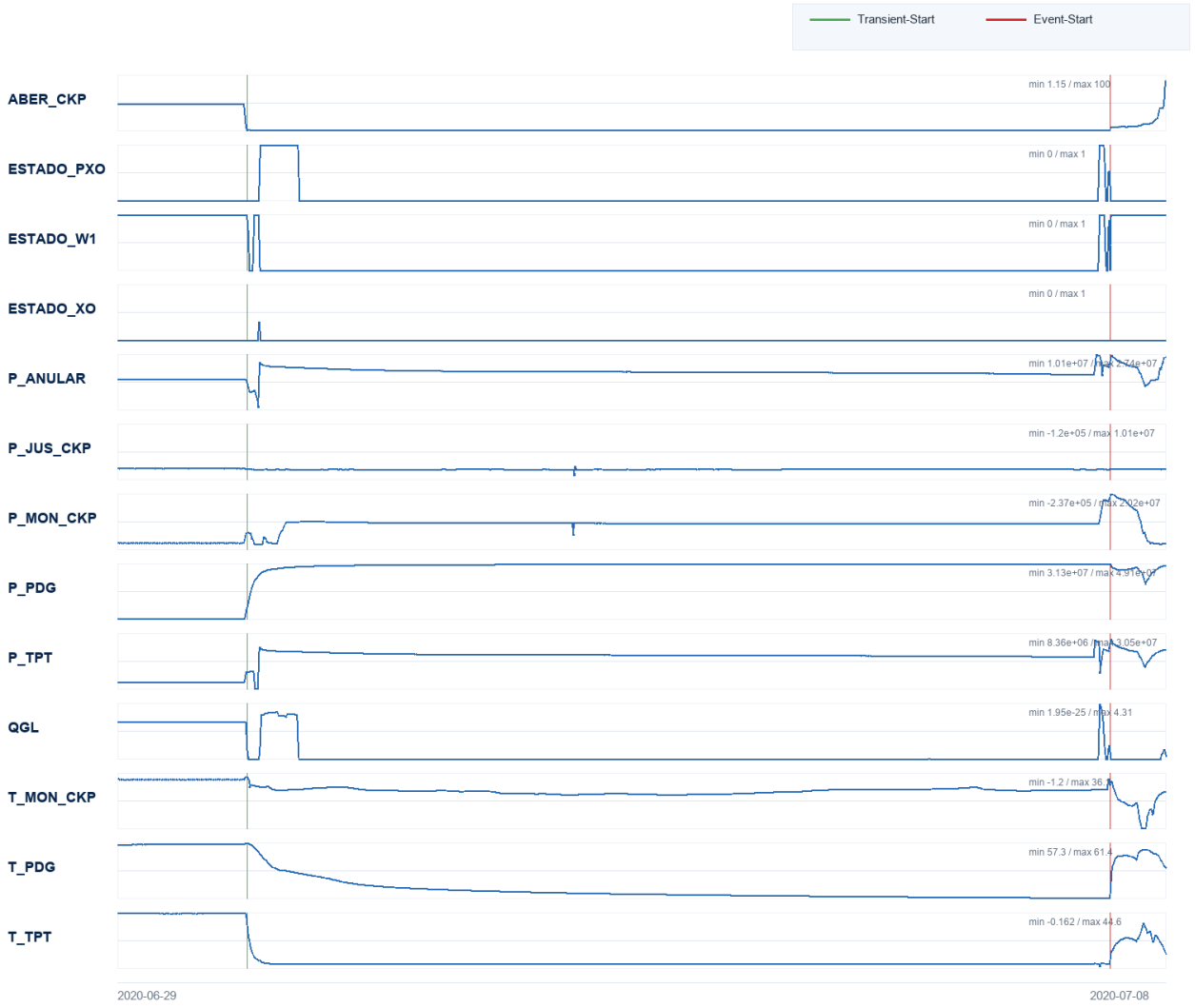


Figure 2 Normalized raw sensor channels in WELL-25. All operational sensor/state channels without status and without TRANSIENT_ON; min-max normalized per channel.

2.1 Separation of Observation Data and Annotations

For scientific interpretation, the separation between the observation vector and the annotation is essential. The operational observation vector consists of the raw sensor/state channels of the well. The variable status is the expert annotation of the later special state; TRANSIENT_ON marks the transient lead-in phase that was also annotated by experts.

In this analysis, these annotations were used for temporal localization: start of the transient phase, transition into the special status, and end of the event. By contrast, the SF2 fingerprints and the overlap curves are to be interpreted as sensor-fusion signals. They are not based on domain interpretation of individual sensor channels, but on the semantic fusion of the measurement-value fingerprints. The example fingerprints reconstructed in Figure 3 explicitly use only operational sensor channels; status and TRANSIENT_ON are excluded.

3. Description of the Derived State Vector

For each timestamp t, a synchronous sensor record

$$x(t) = [x_1(t), x_2(t), \dots, x_m(t)]$$

is formed. The raw file contains 13 operational sensor/state channels: ABER_CKP, ESTADO_PXO, ESTADO_W1, ESTADO_XO, P_ANULAR, P_JUS_CKP, P_MON_CKP, P_PDG, P_TPT, QGL, T_MON_CKP, T_PDG, T_TPT.

The SF2 project well25_raw.sf2project.json loaded for the documented overlap curves uses 12 active feature channels from these. ESTADO_XO is inactive in petrobras_norm_binning.json via Delete; status and TRANSIENT_ON are status/evaluation channels and are not included in the state fingerprints reconstructed in Figure 3.

The derivation of the SF2 state vector proceeds in four steps:

- **Normalization:** numerical values are mapped into a comparable value range according to the same scheme, in this setup via min-max clipping.
- **Binning:** normalized values are converted into discrete categories. This turns a continuous measurement value into a symbolic sensor state.
- **Measurement-value fingerprint lookup:** each discrete sensor symbol addresses a sparse histogram stored in the Retina.
- **Fingerprint aggregation:** the measurement-value histograms are accumulated in binary form and then reduced to 3% sparsity. With a 64 x 64 Retina, this corresponds to floor(4096 * 0.03) = 122 active bits.

The preprocessing configuration used is structured uniformly for all sensor and feature columns: numerical values are normalized and then mapped into discrete bins. All columns not treated as required or evaluation fields are interpreted by SF2 Suite RT as sensor channels; the later semantic fusion does not distinguish methodologically between raw channels and derived feature columns.

3.1 Definition of the Derived Features

In addition to the raw-core setup, a feature-extended variant was created. It contains 129 sensor channels and serves as a methodological extension when not only instantaneous values but also local process dynamics should enter the semantic space. The derived features are computed causally, i.e., a feature at timestamp t uses only values up to and including t.

Pressure topology

$$\text{DP_DOWNHOLE_TREE} = \text{P_PDG} - \text{P_TPT}$$
$$\text{DP_DOWNHOLE_CHOKE} = \text{P_PDG} - \text{P_MON_CKP}$$
$$\text{DP_CHOKE_CKP} = \text{P_MON_CKP} - \text{P_JUS_CKP}$$
$$\text{DP_CHOKE_TREE} = \text{P_MON_CKP} - \text{P_TPT}$$
$$\text{DP_JUS_TREE} = \text{P_JUS_CKP} - \text{P_TPT}$$
$$\text{DP_ANNULUS_TREE} = \text{P_ANULAR} - \text{P_TPT}$$

These features describe pressure drops and pressure coupling along the production path.

Ratio and actuator proxies

$$\text{RATIO_TREE_DOWNHOLE} = \text{P_TPT}/\text{P_PDG}$$
$$\text{RATIO_MON_DOWNHOLE} = \text{P_MON_CKP}/\text{P_PDG}$$
$$\text{RATIO_JUS_MON} = \text{P_JUS_CKP}/\text{P_MON_CKP}$$
$$\text{GL_PER_CKP_OPEN} = \text{QGL}/\text{ABER_CKP}$$
$$\text{P_MON_PER_CKP_OPEN} = \text{P_MON_CKP}/\text{ABER_CKP}$$
$$\text{DP_CKP_PER_OPEN} = \text{DP_CHOKE_CKP}/\text{ABER_CKP}$$
$$\text{QGL_ACTIVE} = 1(\text{QGL} > 1\text{e-}6)$$

This group captures relative operating states, choke action, and gas-lift activity.

Temperature and thermodynamic proxies

$$\text{DT_DOWNHOLE_TREE} = \text{T_PDG} - \text{T_TPT}$$
$$\text{DT_TREE_CHOKE} = \text{T_TPT} - \text{T_MON_CKP}$$
$$\text{THERMAL_PRESSURE_GRADIENT} = \text{DT_DOWNHOLE_TREE}/\text{DP_DOWNHOLE_TREE}$$

These features represent temperature gradients and coupled pressure/temperature changes.

First order derivatives

For central raw and differential signals X, rates of change are computed on three time scales:

Source note: The original report does not supply formulas or time intervals for the three first-order derivative time scales.

They describe instantaneous and short-scale state changes.

5 minute rolling mean

$$\text{MEAN300_X}(t) = \text{mean}(\text{X}[t - 299\text{s} \ldots t])$$

The rolling mean describes the smoothed local operating state.

5 minute rolling variance

$$\text{VAR300_X}(t) = \text{variance}(\text{X}[t - 299\text{s} \ldots t])$$

The rolling variance describes local instability, oscillation, and control motion.

15 minute rolling trend

For X[t - 899s ... t], a linear regression

$$\text{X} = a + b \times \text{seconds}$$

is computed. The feature is the slope:

$$\text{TREND900_X}(t) = b$$

The rolling trend describes robust, longer-term drift rather than simple point differences.

15 minute boundary tracking

$$\text{MIN900_X}(t) = \text{min}(\text{X}[t - 899\text{s} \ldots t])$$
$$\text{MAX900_X}(t) = \text{max}(\text{X}[t - 899\text{s} \ldots t])$$
$$\text{RANGE900_X}(t) = \text{MAX900_X}(t) - \text{MIN900_X}(t)$$
$$\text{BPOS900_X}(t) = (\text{X}(t) - \text{MIN900_X}(t))/\text{RANGE900_X}(t)$$

These features describe the current position of the state near the lower or upper boundary of its local 15-minute behavior.

In the present report analysis, the raw core takes precedence because it shows the argument of low technical prior assumptions most clearly. The feature variant is interesting for follow-up analysis, because SF2 can also fuse dozens to hundreds of channels without changing the methodological principle.



Figure 3 Representative state fingerprints. SF2 state fingerprints from the project Retina petrobras_64_64_100; 64 x 64, 3% sparsity.

The fingerprint figure visualizes three real timestamps from WELL-25: normal operation, high warning phase, and event start. The depiction is reconstructed from the stored SF2 Retina: for each active measurement-value histogram, the binary presence of the Retina positions is accumulated; afterwards, the top positions are selected deterministically by descending count and ascending index. Sensor records are therefore not interpreted individually, but made comparable as overall state patterns.

4. Training of the Retina

Retina training uses a completeness-filtered training set of 200,000 records (well25_raw_sensor_train_200k.csv). Compared with the full stream, the selection is small, but deliberately sufficient for the use case: SF2 is not intended to train a globally generalizing model across many wells, but to create a semantic operating space for one concrete physical system.

A practical core point is the low setup and training time. The complete setup and Retina training were implemented in less than one hour; the actual retina training run takes only a few minutes. As a result, cross-well generalization is not mandatory. Each well can receive its own asset-specific model. For industrial applications, this is advantageous because the retina adapts to the individual sensor set, operating mode, and process characteristics of the respective physical system.

The most important training principles were:

- all data come from the same well system;
- the channels are time-synchronized at 1 Hz;
- the training set prefers complete records;
- status 1 is fully represented, while status 0 is supplemented in a balanced and completeness-oriented manner;
- the retina is understood as a local semantic space for WELL-25.

The SF2 retina operates as a 64 x 64 semantic grid. Sensor states are mapped onto this retina, and timestamp-specific fingerprints arise through aggregation of the active semantic positions.

5. Overlap Measurements

After retina training, running sensor states are converted into fingerprints and compared with reference patterns. The present overlap time series contain two curves:

- **Baseline:** overlap of the current fingerprint with an aggregation of the five preceding snapshots. A high value means that the current state closely matches the immediate past locally.
- **Hydrate Production Line:** overlap of the current fingerprint with a stored fault-example fingerprint from the status library. A high value means semantic proximity to the referenced Hydrate Production Line state.

The overlap values lie on a scale from 0 to 100. Mathematically, this is a normalized comparison of active fingerprint bits; the concrete normalization follows the SF2 Suite RT implementation.

The expert annotation status is not used here to explain individual sensors. It marks only the point in time at which the special status begins in the ground truth. Likewise, TRANSIENT_ON marks only the lead-in phase. The scientific observation is therefore: sensor fusion already moves toward the stored Hydrate Production Line fingerprint before the status switch occurs.

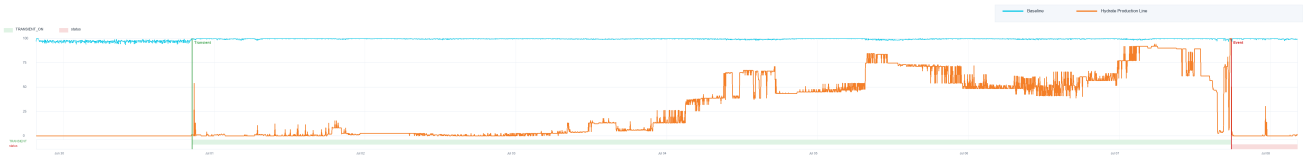


Figure 4 Overlap time series as a single-line timeline. Full WELL-25 time span on one continuous linear x-axis.

6. Analysis of the Overlap Curves

The analysis combines PetroBras_status_timeseries.csv and PetroBras_overlap_timeseries.csv. Both files are synchronized with well25_raw_sensor_in.csv and contain 722,192 timestamps.

6.1 Event timestamps

Table 3 Event timestamps

Event	Row	Timestamp	Distance to event
Transient start	89,356	2020-06-30 20:30:57	6d 21:16:07
Event start	684,323	2020-07-07 17:47:04	0
Event end in file	722,191	2020-07-08 04:18:12	10:31:09 after event start

6.2 Phase statistics

Table 4 Overlap statistics by phase

Phase	Duration	Baseline mean	HPL mean	HPL median	HPL p95	HPL max
Pre-transient	1d 00:49:16	96.88	0.005	0.0	0.0	1.6
Transient, status 0	6d 21:16:07	99.51	28.61	31.97	77.05	100.0
Event, status 1	10:31:09	99.52	0.197	0.0	0.0	100.0

The pre-transient phase shows practically no hydrate overlap. In the transient phase, HPL overlap rises strongly. At the same time, the baseline remains almost continuously high. This leads to a central interpretation: the system is not drifting chaotically, but is moving continuously and locally stably toward a semantically similar fault state.

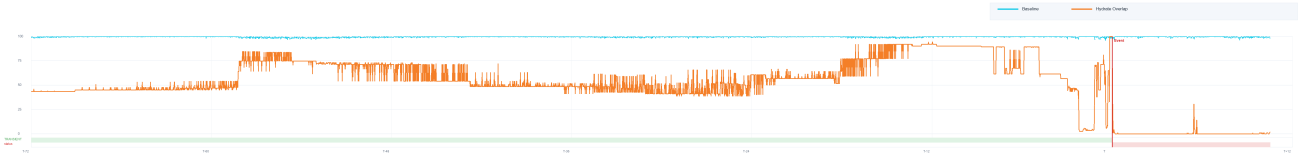


Figure 5 Overlap zoom as a single-line timeline. Critical region from T-72h to T+12h on a linear axis; T marks the annotated event start.

6.3 Early warning thresholds

Table 5 Early warning threshold exceedances

Threshold	First exceedance	Stable 60 s exceedance	Stable 300 s exceedance
≥ 1	2020-06-30 20:27:11	6d 21:18:51 before event	5d 23:44:04 before event
≥ 70	2020-07-04 17:06:50	2d 09:04:29 before event	12:51:18 before event
≥ 80	2020-07-05 07:52:05	13:40:34 before event	12:21:17 before event
≥ 90	2020-07-07 00:42:25	13:40:34 before event	12:21:17 before event

The thresholds are not to be interpreted as final alarm rules. Rather, they show that the overlap curve contains leading decision information at several persistence levels. For a productive system, a staged alarm strategy would be sensible: observation, diagnosis, and operational intervention.

6.4 Sensor level shifts

Table 6 Sensor level shifts

Channel	Absolute change	Relative change
ABER_CKP	0.0977	7.3%
P_ANULAR	5.86 million	31.5%
P_JUS_CKP	38.3k	2.1%
P_MON_CKP	13.52 million	233.3%
P_PDG	15.20 million	44.8%
P_TPT	12.87 million	84.6%
QGL	-0.662	-100.0%
T_MON_CKP	-6.76	-19.8%
T_PDG	-4.04	-6.6%
T_TPT	-35.5	-89.6%

These changes are physically interpretable, but in the SF2 approach they are not presupposed as hand-modeled features. The semantic representation uses the joint signature of the system state.

7. Advantages of the SF2 Method Compared with Traditional Methods

Table 7 Comparison with traditional methods

Criterion	Traditional thresholds / models / ML	SF2 Suite RT
Prerequisite	Sensor types, limits, feature engineering, or physical model	Synchronized sensor records from the same system
Model objective	Threshold violation, classification, or simulation	Semantic comparability of states
Data demand	Often many labeled fault cases or expert models	Local operating space with small per-well training
Generalization	Cross-well generalization is difficult	Individual model per well is possible
Interpretability	Single values, model parameters, or black-box scores	Overlap to known state fingerprints
Deployment	Often engineering-, cloud-, or model-maintenance-intensive	Can be operated locally and edge-ready

The central methodological advantage is not that SF2 replaces domain knowledge. Rather, SF2 moves domain knowledge to a different place: experts are used for evaluating and naming relevant states, but not necessarily for the initial construction of the sensory state space.

8. Further Procedure

1 **Prospective replay test:** temporally causal replay in which only historical information is available at each point in time.

2 **Ablation Raw vs. Derived:** comparison between raw-channel retina and feature retina with rolling means, variances, slopes, derivatives, and boundary tracking.

3 **Per-well replication:** training individual retinas for additional 3W wells with Class 8, without having to force a shared cross-well model.

4 **Alarm design:** definition of persistence, hysteresis, and multi-stage rules for overlap thresholds.

5 **Benchmarking:** comparison against SCADA thresholds, classical ML models, autoencoder/LSTM-based anomaly detection, and flow-assurance heuristics.

6 **Status-library design:** investigation of whether multiple fault fingerprints per event phase provide better persistence after event onset.

7 **Industrial ROI validation:** linking early-warning lead time with real costs for production loss, hydrate inhibitors, remediation, and deferred production.

9. Conclusion

The WELL-25 analysis shows that SF2 Suite RT can make a Hydrate Production Line event visible as a semantic state shift. Particularly relevant is the combination of high local baseline stability and increasing proximity to the hydrate fingerprint: the process state remains continuous, but moves over days toward a known fault state.

For the oil and gas industry, this property is practically significant. It enables per-well early-warning models without initial dependence on sensor-type knowledge, threshold design, or physical hydrate modeling. Because retina training takes only a few minutes, each asset can receive its own semantic model. The next scientific step is prospective validation across multiple wells and event types.

References

1 Petrobras, 3W Project / 3W Dataset, GitHub Repository: https://github.com/petrobras/3W

2 Vargas et al., A realistic and public dataset with rare undesirable real events in oil wells, Journal of Petroleum Science and Engineering, 2019: https://doi.org/10.1016/j.petrol.2019.106223

3 Vaz Vargas et al., 3W Dataset 2.0.0: a realistic and public dataset with rare undesirable real events in oil wells, Scientific Data, 2026: https://www.nature.com/articles/s41597-026-07225-z

4 Offshore Magazine, Flow assurance: Chemical inhibition of gas hydrates in deepwater production systems, 2010: Offshore Magazine article, 2010

5 Offshore Magazine, Flow assurance: Maintaining plug-free flow and remediating plugged pipelines, 2000: Offshore Magazine article, 2000