Synthefy

The World's First Multi-Modal Foundation Model for Time Series Data
Synthesize and Forecast - All from a Simple Text Prompt

Why a *Multi-Modal* Foundation Model for Time Series?

Synthefy is building the world's first foundation model for multi-modal time series data. Our GenAI toolbox allows customers to forecast and create privacy-preserving synthetic time series data – all from a simple text prompt. Notably, we are building the world's first multi-modal forecasting engine that takes into account **rich contextual data**, such as weather, news articles, market sentiment, etc. to forecast and synthesize better. Our team has extensive experience with time series from our time at Uber, OpenAI, Nvidia, and Stanford.
Example use cases of our products (**synthesis and forecasting**) are:

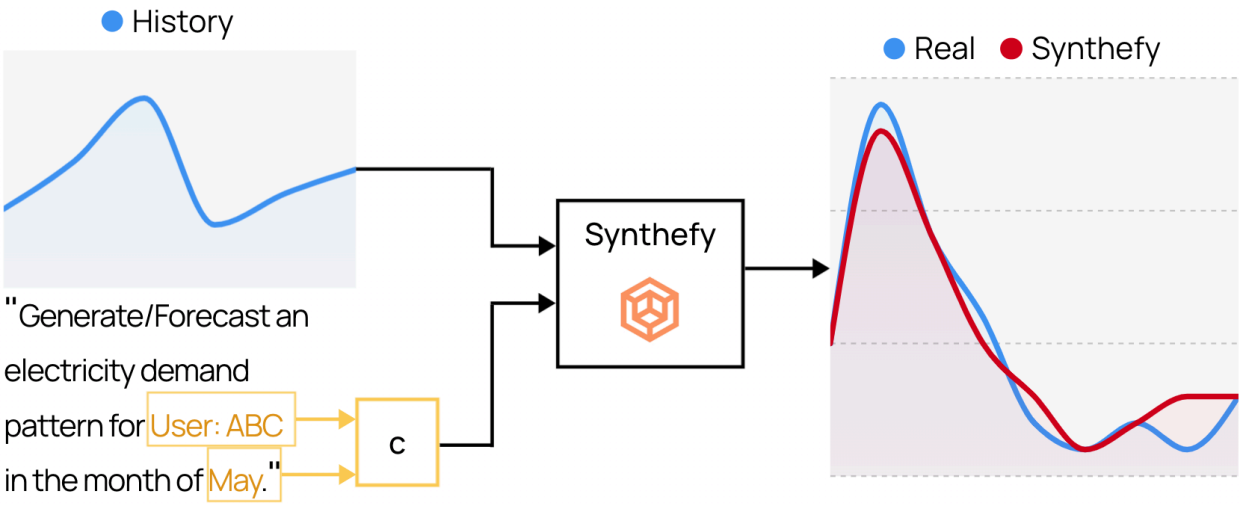

Figure 1: Synthesize and Forecast, all with a *single GenAI foundation model*. Synthefy allows you to improve results with rich contextual data for your domain.

Privacy-Preserving Synthetic Data:
A network operator can ask: "Generate a cell demand pattern that looks like an urban cell but with 10% more congestion on weekends." Use cases include stress testing production engineering systems and anonymizing customer data.

Multi-Modal Forecasting:
A video streaming platform can ask: “Forecast customer demand for streaming movies on a Friday in the USA when a new Bollywood movie is released”. Use cases include demand forecasting, capacity planning, and anomaly detection.

Customer Pain Points Today

Customers in **energy, finance, networking, e-commerce, and medicine** are struggling to gain insights from terabytes of streaming time series and system log data. Today’s time series analysis tools are archaic – they simply forecast the future using past data. However, they can’t handle, and thus simply ignore, rich contextual metadata. In networking, this metadata could be 5G cell location, system logs, and application type, while in e-commerce inventory management, this could be advertisement spend, website traffic, Instagram/TikTok views, and previous sales.

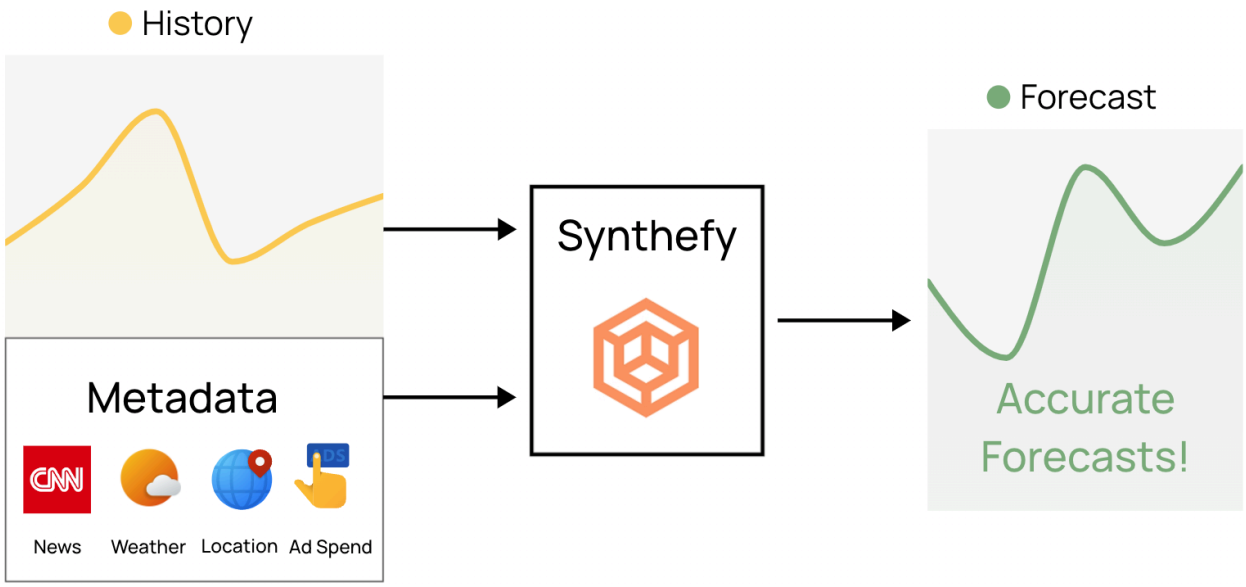

Figure 2: The Power of Multi-Modal Forecasts: Synthefy adds rich metadata for your domain to create better forecasts and synthetic data.

To solve this problem, Synthefy is building a multi-modal foundation model that can leverage rich context in documents, graphs, other timeseries, or even categorical and quantitative variables. In essence, Synthefy solves the following customer problems:

Privacy-Preserving Synthetic Data:
Train models on real data, anonymize them using the latest privacy-preserving ML techniques, and create infinite synthetic variants to test your applications.

Synthesize Rare Training Data, Anomalies, or Stress-Tests:
System failures or anomalies are often the most important for customers, but occur too rarely to train robust ML models. Synthefy's tools can be used to generate new failure cases or anomalies to improve ML model performance or stress-test engineering systems.

Multi-Modal Forecasting: Type in a text prompt, give us a short window of past historical data, provide any contextual metadata, and automatically boost the accuracy of forecasts!

The Synthefy SaaS Platform

Synthefy's platform is a software-as-a-service (SaaS) package. It can be licensed for on-premise use or is available as a cloud subscription. All our platform needs is a pointer to a multi-modal dataset of time series and optional metadata. Synthefy provides trained models for synthetic data, forecasting, and anomaly detection. We never need to see your private data if you opt for on-premise use.

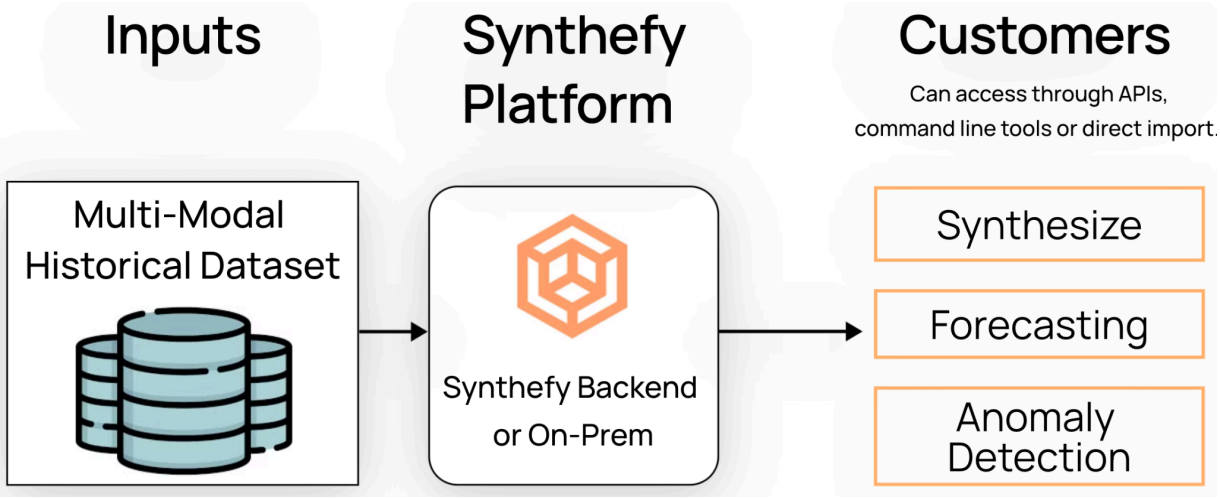

Figure 3: Our SaaS platform can be deployed in the cloud or locally for privacy-preserving training and inference. The ***same model*** serves synthesis, forecasting, and anomaly detection.

Our Novel Generative AI Models

Synthefy has pioneered a novel joint architecture for time series synthesis, forecasting, and anomaly detection using GenAI models. GenAI models, such as DALL-E, Midjourney, and Stable GenAI, have achieved tremendous success at synthesizing photo-realistic images from a text prompt. However, synthesizing time series is fundamentally more challenging, as we discuss next.

Why is Synthesizing Time Series Fundamentally Harder than Images?

Traditional GenAI models for images use a static text prompt, such as “generate me an image of a dog in Disneyland”. However, the metadata conditions for time series are much more nuanced. They could be a time-variant weather and precipitation pattern, which is *itself a time series*, or a mixture of quantitative and categorical variables, such as the age, gender, weight, and presence of a pacemaker for a patient’s electrocardiogram (ECG).

As humans, we can visually inspect a synthetic image and assess its quality or use a host of standard metrics, such as the Frechet Inception Distance (FID) score. However, by just glancing at a time series, we can’t tell if it retains key statistical moments, Fourier coefficients, etc. As described later, Synthefy provides a comprehensive evaluation toolbox to rigorously test the fidelity of synthetic data.

Compute Efficiency: Train on GPU, *Run Inference on a CPU*

Given we are operating with time series, not high-dimensional inputs like video, our models train efficiently on a GPU. For example, we can train a very good model on terabytes of data on a lower-end Nvidia GPU (A5000) in just 6-8 hours. Once a model is trained, we are ready for real-time deployment of a model (i.e., *inference*). Our models are efficient to run on CPU, which is a huge cost and efficiency benefit for customers.

Case Studies: Medicine, Energy, Networking, and Transportation

We now showcase the broad applicability of the Synthefy toolbox on public datasets in medicine, transportation, energy, etc. Remarkably, our uniform architecture works for all these datasets and more!

Dataset and Use Cases

We selected the following real-world datasets since they feature a diversity of challenging practical attributes. Specifically, they feature a diverse mix of seasonalities, discrete and categorical conditions, a wide range of horizons, and multivariate correlated channels.

Air Quality:
The Air Quality dataset has six multivariate channels that contain various air quality parameters recorded every hour at 12 different stations from 2013 to 2017. The air quality metrics are paired with the corresponding weather metrics. An example use case is: “Generate an air quality pattern for the next 4 days given a time series of the weather forecast and the station.”

Medical Electrocardiogram (ECG):

The PTB-XL Electrocardiogram (ECG) dataset tests Synthefy's technology on highly non-linear and non-stationary biomedical signals. This data is challenging since it has 12 correlated channels corresponding to ECG leads on the human body. Moreover, the dataset has 71 rich conditions, ranging from the patient's heart disease to age, gender, weight, and presence of a pacemaker etc.

Electricity Demand:
The UCI Electricity Load Diagrams dataset records power consumption usage, recorded every 15 minutes, from 370 households between 2011 and 2014. An example use case is: "Generate an energy demand pattern for User ABC for a day in February."

Highway Traffic:
The traffic dataset contains a univariate time series of traffic volume recorded for every hour on a particular highway from 2012 to 2018. The dataset is particularly challenging since the metadata conditions *are themselves a time series*, such as hourly rainfall, snowfall, temperature, and holiday seasonality etc.

Synthefy Beats Today's State-of-the-Art GANs by 5x

Figure 5 shows the quality and realism of Synthefy's synthetic data. Each column is a challenging public dataset. The first row is Synthefy's model, and the second row is the previous state-of-the-art of Generative Adversarial Networks (GANs). Each time series is for a test prompt/condition whose data was never seen before during training. The real time series is in blue, while Synthefy's synthetic time series is in red.

Clearly, Synthefy's model (row 1) closely matches the ground-truth data and significantly outperforms GANs. The ECG test conditions were for specific patient conditions *that don't even exist,* and hence, we don't have a ground truth blue curve. Still, the pattern resembles a standard ECG wave from medical literature. As discussed next, we show that training a disease classifier on synthetic data, but testing it on real patients yields extremely high accuracy.

The **key take-away** is that Synthefy's model closely matches ground-truth, challenging test time series and beats GANs by 5x.

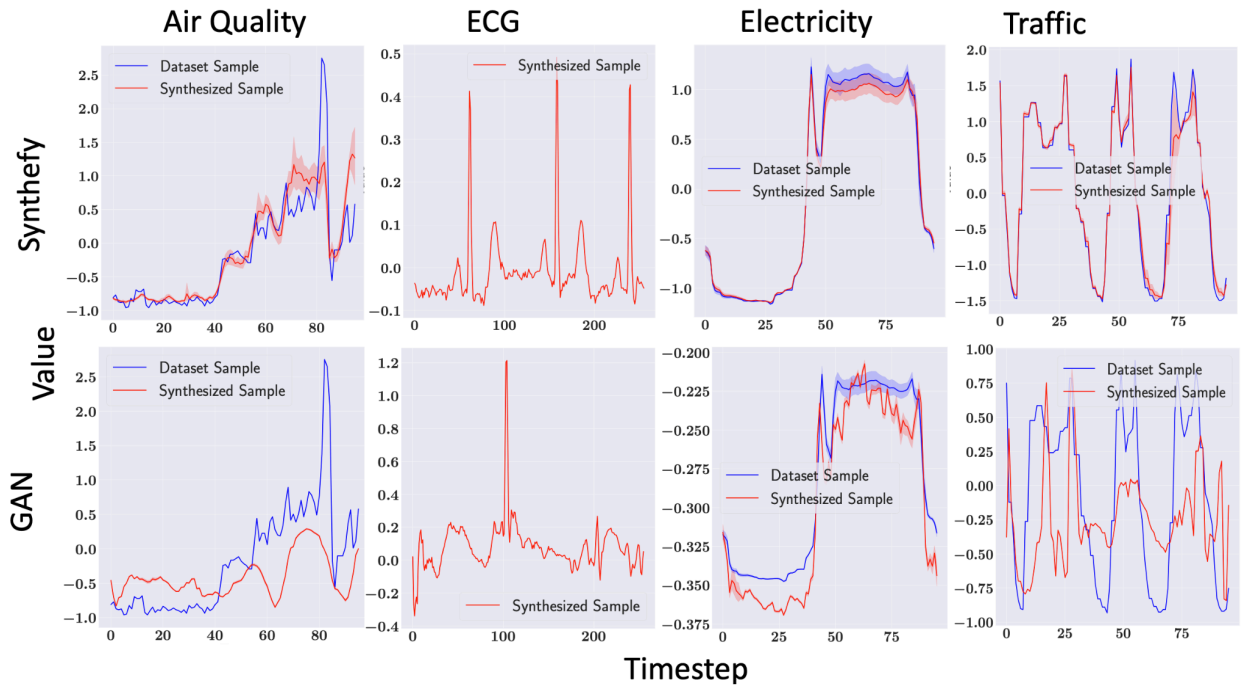

Figure 5: Synthefy significantly outperforms GANs. Synthefy (top row) creates realistic synthetic data (red) compared to today's competitors, which rely on GANs (bottom row).

How Realistic is Synthefy's Synthetic Data?

The Synthefy platform provides rigorous out-of-the-box evaluation metrics to test the fidelity and utility of synthetic data.

Time Domain Evaluation Metrics

Figure 6 shows that Synthefy accurately models the real distribution of time series data. Again, each column is a challenging dataset. We plot the distribution of time series values across a dataset for real, unseen test conditions (blue) and synthetic data (red). The first row is Synthefy's model, while the second row is for our closest competitors of GAN methods.

Clearly, Synthefy's models accurately represent the time domain distribution much more accurately than GANs. In particular, GANs are notorious for only being able to create uni-modal data distributions and suffer from the well-known *mode collapse* problem. This is clearly seen for the Traffic dataset (bottom right), where the real traffic data distribution is bi-modal, while the GAN only learns an inaccurate "average" mode.

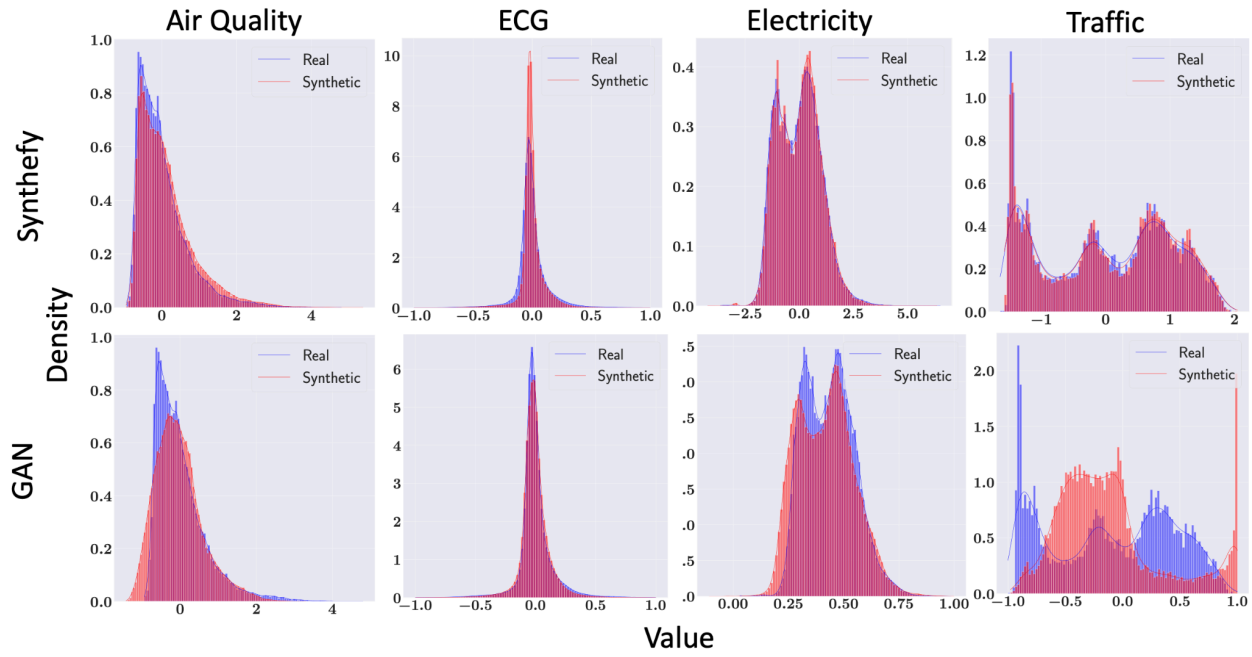

Figure 6: Synthefy accurately models the distribution of time series data.

Synthefy Accurately Models Frequency Spectra

Synthefy's platform also automatically evaluates that synthesized time series have a similar Fourier spectra as the real time series. This is crucial to retain high frequency content and seasonality patterns in real-world signals, such as radio frequency signals.

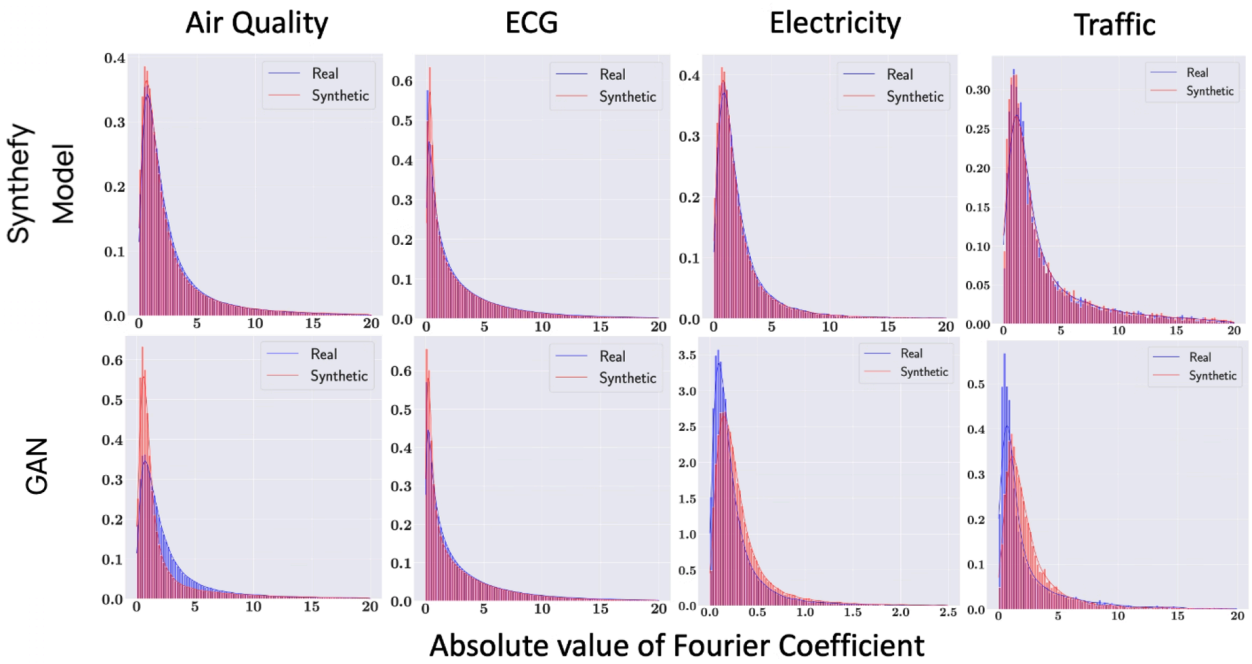

Figure 7: **Synthefy accurately models the Fourier spectrum** (frequency distribution) of time series data.

Our Data Significantly Boosts Accuracy on Downstream ML Tasks

The ultimate test of synthetic data is its utility for downstream tasks, such as data augmentation or training robust ML models. As such, we train an ML inference model (classification or regression) on purely synthetic data, but test it on real data. In the ECG dataset, a disease classifier trained on real patient data achieves 95% accuracy on real, unseen test patients. Remarkably, a classifier trained on Synthefy's *purely synthetic data* achieves a staggering 93% accuracy when tested on the same real, unseen test patients. Our academic papers highlight a plethora of similar results.

Synthefy's Model Learns Causal Relationships Between Relevant Metadata and Time Series

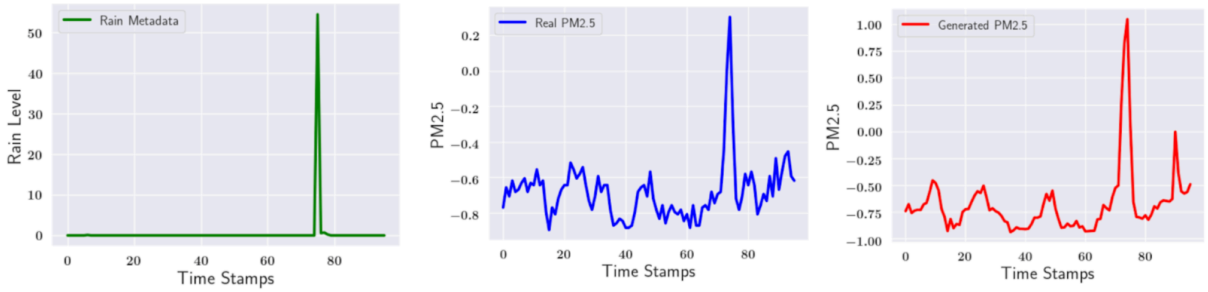

Figure 8: **Synthefy accurately learns the relationship between causal changes in metadata with resulting changes in synthesized time series.** Here, a spike in precipitation (left, green) yields a spike in synthesized air quality values (right - blue and red).

A key innovation of Synthefy's model is to explain "why" our models synthesize or forecast a specific time series trend based on metadata. For example, in e-commerce, we want to say that "sales increased because I saw a spike in website traffic 2 weeks after I began an ad campaign". In essence, we want to find causal relationships between changes in relevant metadata with the resulting time series.

The above figure shows our model learns such causal relationships. In this example, we synthesize air quality (particulate matter or PM) time series based on a time series of rainfall, which serves as part of the contextual metadata. Clearly, on the left, we see a spike in the rainfall (green) just before time step 80. Thus, our model learns to synthesize a peak in PM count (air quality) in red.

Synthefy's Model Ensures Hard Constraints are Satisfied

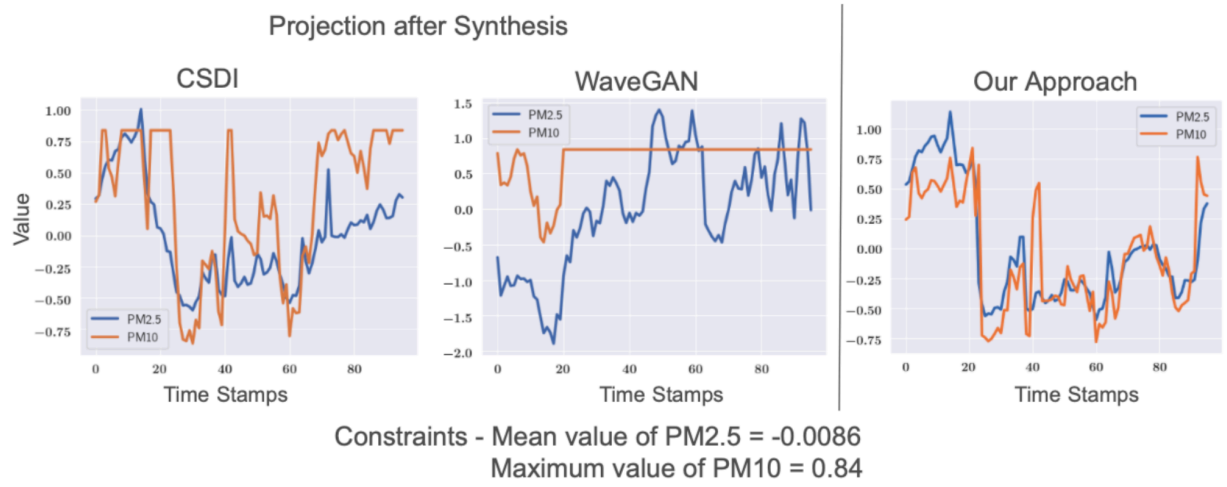

Figure 9: **Synthefy can strictly adhere to a diverse set of hard constraints on the synthesized time series.** Here, the orange and blue time series should be correlated and satisfy the constraints on the mean and maximum values shown below the figure. Our approach (right) yields smooth time series that are realistic and obey constraints.

In many engineering and finance applications, we want to ensure that synthesized and forecasted time series satisfy hard constraints from domain knowledge or physical limits. For example, if the maximum inventory in an e-commerce warehouse is 100 boxes, we should never forecast/synthesize an inventory of 105 boxes. Likewise, in finance, if we want to stress-test a trading strategy on an S&P 500 window with 10% extra price volatility, we need to ensure that the synthesized time series indeed satisfies this constraint on volatility. This will ensure we accurately assess risk when tuning parameters of a trading strategy.

A key innovation of Synthefy's model is that we **can enforce a wide range of real-world hard constraints** on time series *with zero extra training*. Figure 9 shows an example of our constraint generation approach for the same air quality dataset. We have two highly correlated time series in orange and blue that should track each other. Moreover, we have a different constraint per channel – a maximum limit on the blue signal and a desired mean value for the orange. Prior methods on the left, such as CSDI and WaveGAN, perform poorly and the synthetic data does not match the ground truth. However, our approach on the right satisfies the constraints and generates highly realistic, correlated time series.

Case Study: Multi-Modal Forecasting

We now show how Synthefy significantly outperforms competitors like Meta's Prophet, Amazon Chronos, and so-called time series foundation models. Our key differentiation is that we have better models and we effectively incorporate multi-modal contextual metadata (e.g., text, categorical variables) to improve the quality of forecasting models.

The below table shows the mean squared error (MSE) for forecasting using Synthefy's models vs. Amazon Chronos on a real network operator's base station traffic dataset, and public air quality, electricity, and road traffic datasets. Clearly, Synthefy's models (orange) significantly reduce forecasting error.

Adding **Multi-Modal Data** Significantly Improves Forecasting

MSE x Models	Base Station Traffic (5 Channels)	Air Quality (6 Channels)	Electricity	Road Traffic
Amazon Chronos Foundation Model	0.64	0.86	0.104	4.845
Models **without** Multi-Modal Data	0.54	0.770	0.068	3.527
Synthefy with Multi-Modal Data	0.38 (1.42x Better)	0.574 (1.49x Better)	0.036 (2.8x Better)	2.922 (1.65x Better)

Synthefy

Work with Us

The Synthefy platform is already solving real business problems for Forbes Fortune 500 Global companies. Our platform broadly applies to use cases in finance, energy, e-commerce, networking, and more. We are happy to partner with customers for proof-of-concepts and enterprise deployments. Please contact sales@synthefy.com for more!