Using machine learning to monitor financial markets

Keynote speech by Gaston Gelos, Deputy Head, Monetary and Economic Department,
Bank for International Settlements[1]

At the 6th Biennial Conference on Financial Stability, Bank of Mexico
Mexico City, 18 November 2025

It is a real honour for me to speak at the financial stability conference on the Bank of Mexico's 100th birthday. I have a somewhat special connection to the Bank of Mexico which traces back to an internship here in 1997.

Today, I want to focus on recent research we have been conducting at the Bank for International Settlements (BIS). As you know, at the BIS, our mandate is to support central banks.

One aspect is exploring how advances in artificial intelligence (AI) and machine learning can be applied to the daily work of central banks and supervisors.

A specific question that I want to focus on today is: can we use these techniques to predict episodes of market stress or market dysfunction before they actually materialise?

This is an important policy challenge because episodes of market dysfunction can have significant economic consequences. In some cases, they can even result in a full-blown crisis. In fact, in recent years, central banks have often felt compelled to intervene to address market dysfunction in key markets.

If we could get an early warning of such events, supervisors could prepare for them and potentially even take pre-emptive action to mitigate risks.

I don't need to tell this audience why this is very a difficult task, and research so far has not been terribly successful in this regard. The ability of traditional econometric models to predict market anomalies out of sample is very limited.

This stems in part from the nature of financial markets. They are complex, deeply interconnected – both among themselves and with the real economy – and they often exhibit highly non-linear dynamics. On top of that, true stress episodes are, thankfully, rare events. This means the sample size for forecasting these significant episodes of dysfunction is quite small.

[1] The views expressed here are my own and not necessarily those of the Bank for International Settlements or its member institutions.

For a predictive tool to be truly useful, it would need to have two key characteristics.

First – and this is obvious – it needs to be able to forecast dysfunction sufficiently in advance. But more importantly, it should avoid giving supervisors false alarms or predicting too many such episodes. We've all heard the joke about economists forecasting 10 out of the last five recessions. That's what we want to avoid.

Second, the tool would be even more valuable if it could help supervisors understand the potential source of the dysfunction. Knowing that trouble is coming would be helpful on its own, since it would allow supervisors to prepare for the storm. But if we could pinpoint where the trouble is brewing, supervisors could go a step further – they could intervene to reduce the severity of the issue or even prevent it altogether. Later, I will show you two different approaches to tackle this issue.

Why may using AI and machine learning to forecast market dysfunction be a promising avenue?

These models offer clear advantages over traditional econometric techniques, such as autoregressive models. First, AI models are much better at capturing the non-linear dynamics that often define dysfunction in the financial sector. Unlike traditional econometric methods, which often rely on linear functions, AI models are not constrained in this way. Second, AI models can handle large sets of explanatory variables with ease. They can even identify and select the most important variables for prediction, if needed. Third, they adapt and automatically detect structural breaks.

That said, there's a trade-off. The same ability to discern non-linear patterns that makes AI models so powerful also makes them harder to interpret. They are often seen as "black boxes". It is typically difficult to pinpoint which variables are driving the model's predictions or to provide clear intuition about the economic mechanisms behind them.

The use of AI and machine learning in predicting dysfunction is still small but growing rapidly. So far, most of the focus has been on predicting asset prices and returns, where complex models have proved to be more precise than simple ones across asset classes. There is in particular some evidence that machine learning models outperform econometric models in predicting stock returns. There are also some studies on crisis prediction, which is something different, but these studies suggest that AI has potential.

Let me now highlight two recent pieces of work from the BIS that address these issues.

The first one is authored by my BIS colleagues Iñaki Aldasoro, Peter Hördal, Andreas Schrimpf and Sonja Zhu[2]. It uses machine learning to map out the full distribution of financial conditions in US Treasury, foreign exchange and money markets. For each of these markets, they construct market condition indicators from a variety of components.

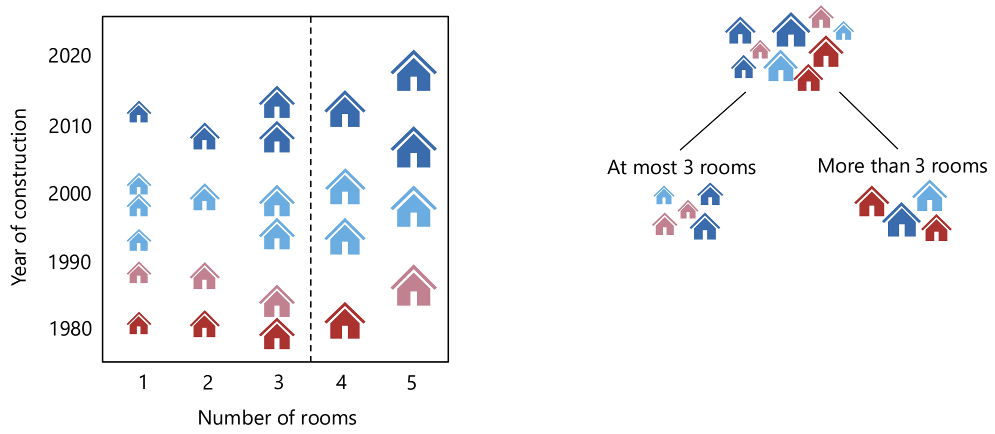
[2] I Aldasoro, P Hördahl, A Schrimpf and S Zhu, "Predicting financial market stress with machine learning," BIS Working Papers, no 1250, March 2025

The second one is by my BIS colleagues Matteo Aquilina, Doug Araujo, Taejin Park, Fernando Perez-Cruz and me.[3] It relies on recurrent neural networks to forecast market dysfunction in foreign exchange (FX) markets and shed light on the drivers of these forecasts.

Let me start by broadly sketching out the methods used in both papers. The paper by Aldasoro and others uses a machine learning algorithm called random forest. A random forest is an ensemble learning method that uses a large number of decision trees to make predictions for classifications and regression tasks.

You may be familiar with decision trees, a popular method for prediction and classification to predict the value of a target variable based on input features. You start with grouping individual data points by sequentially splitting data into finer categories. For instance, if you wanted to forecast the value of a house, you could split houses into those with at most three rooms and those with more than three rooms (see Graph 1). Then you could split them further into those built before or after a specific year and keep adding variables to improve your prediction.

Decision trees group individual data points by sequentially partitioning data into finer categories — Graph 1



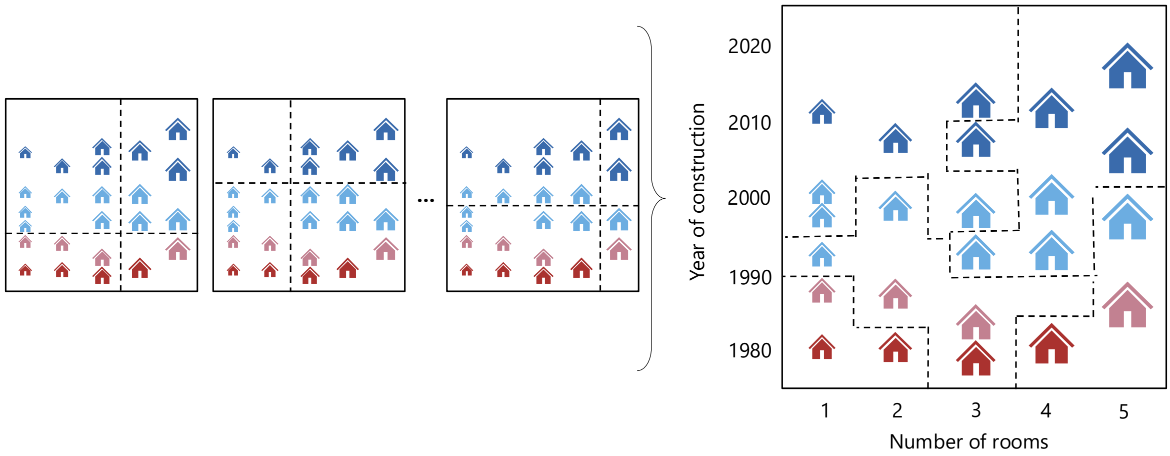
The problem with decision trees is that they have a tendency to overfit, meaning they can learn very irregular patterns in the data you have but then do not generalise well once you want to predict out of sample.

The random forest algorithm addresses this by creating multiple independent decision trees and then averaging their predictions. The decision trees are typically trained on different slices of the same data to improve predicting power (Graph 2). By averaging across many trees, the method reduces the risk of overfitting and produces more robust predictions.

[3] M Aquilina, D Araujo, G Gelos, T Park and F Perez-Cruz, "Harnessing artificial intelligence for monitoring financial markets," *BIS Working Papers*, no 1291, September 2025.

Random forests combine several trees, trained on different slices of the same data, to improve prediction out of sample

Graph 2



To interpret the drivers of the prediction, the authors use so-called Shapley values to analyse different predictors. Shapley values quantify the marginal contribution of each predictor to a specific forecast, conditional on all possible combinations of other variables. ural network (RNN). The architecture of these models is complex and more difficult to explain succinctly compared with random forests.

Broadly speaking, RNNs are a type of machine learning algorithm that, in some ways, are similar in spirit to Kalman filters. Kalman filters estimate and predict the state of a system in the presence of uncertainty, such as measurement error or unknown factors. The key difference is that Kalman filters assume a linear relationship between inputs and outputs, while RNNs do not rely on such a simplifying assumption. This makes RNNs better suited for capturing the non-linear dynamics often present in financial markets.

One of the major contributions of the paper is to add a module to the RNN that: (i) starts from the full set of predictors; (ii) learns recursively the importance of each predictor for the prediction; and (iii) updates weights given to the predictors in a recursive way.

In each period the different variables compete for importance in the prediction, with the model "learning" how to optimise forecasting performance by allocating the prediction weights of each variable based on current and past observations of all variables. By inspecting the weights from the second RNN, we can understand where the model is finding a signal at each point in time.

This is a crucial feature for supervisors and regulators, since it gives them information on what is changing in the data and driving the prediction.

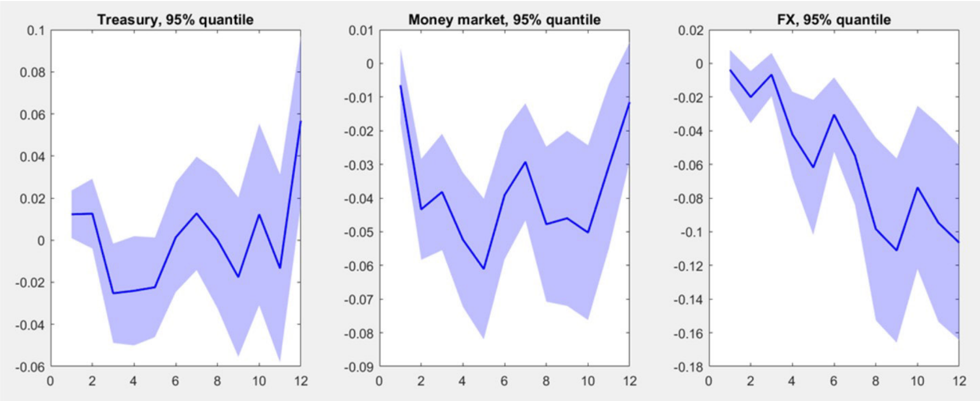
Now, let's take a closer look at what these approaches deliver.

As a reminder, Aldasoro and others construct market condition indicators from a variety of components for three markets: the US Treasury, foreign exchange and money markets.

The paper first finds that simple AR (1) quantile regressions perform better than multivariate regression analysis at longer horizons.

But the random forest model compares well with the AR model. Graph 3 focuses on the 95th percentile of the market condition indicators – the extreme values that are most likely related to stress. The prediction horizon in the graph goes up to 12 months. What the graph shows is the difference in "quantile loss" between the random forest model and an autoregressive (AR) model, across different forecast horizons, which are measured in months on the x-axis. Negative numbers indicate that the random forest model performs better than the AR model.

Random forests in Aldasoro et al (2025) can predict market conditions in money and FX markets well, especially at relatively long forecast horizons — Graph 3



Negative numbers imply RF is better than AR at predicting the 'tail' of the distribution

Now, the AR model is simple but also a tough benchmark to beat. And the random forest model outperforms the AR benchmark at all horizons for the money market and FX indicators. However, it struggles more when predicting Treasury market conditions, particularly at longer forecast horizons.

Let me now turn to the interpretation of signals. I mentioned earlier that while knowing stress is coming is helpful, identifying the likely root cause of that stress is even more valuable, since it allows supervisors to take action in advance.

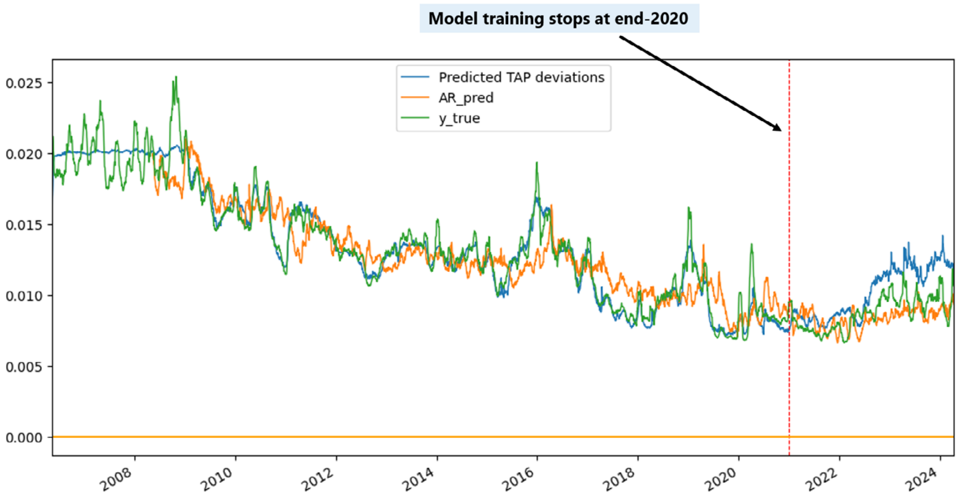
The authors find that, for FX markets, two sets of variables matter most: (i) risk preferences – measured by implied volatility in FX markets; and (ii) sentiment – measured by fund flows. These are key predictors for the extreme right tail of the distribution, which is most relevant for stress scenarios.

In the paper by Aquilina and others we pursue a different approach. The study focuses on a single measure of market dysfunction – the standard deviation of triangular arbitrage parity conditions in the euro-yen exchange rate. It is a useful measure to focus on because deviations in triangular arbitrage parity are strongly correlated with market-wide dislocations.

In FX markets, the euro-yen exchange rate should equal the euro-US dollar exchange rate multiplied by the US dollar-yen exchange rate. In other words, exchanging euros for yen should be equivalent to exchanging euros for dollars and then dollars for yen. Any discrepancy in these exchange rates should be quickly arbitraged away, as these markets are large and typically highly liquid. If such discrepancies persist, it suggests that markets are not functioning properly.

The study takes the minute-by-minute differences between the direct euro-yen exchange rate and the dollar-intermediated exchange rate. These differences should in theory be zero. The study then calculates the daily standard deviations of these differences. It turns out that, historically, this measure has been strongly correlated with episodes of significant overall market stress.

Predicting variance of triangular parity deviations in the euro-yen exchange rate — Graph 4



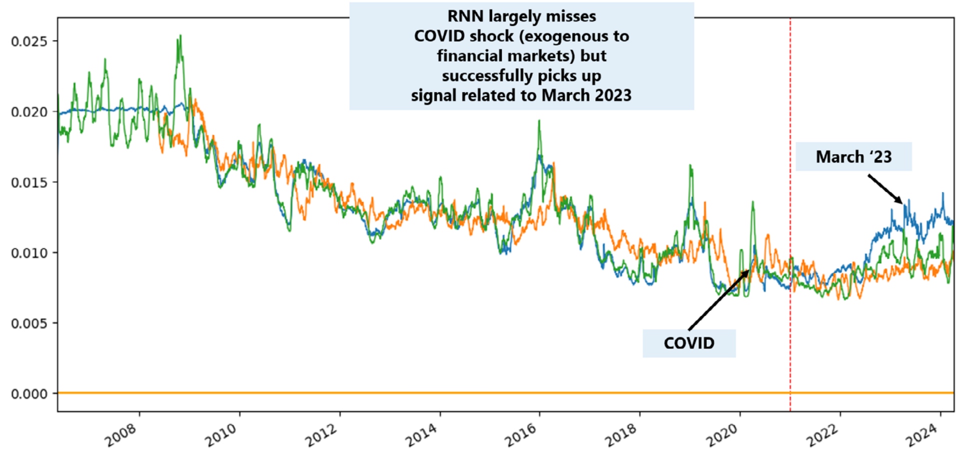
Now, let’s move to the results.

The graph here shows: (i) the target variable, which is in green; (ii) predictions based on a simple autoregressive model in orange; and (iii) the predictions based on the RNN in blue.

All predictions are for 60 business days ahead – roughly three months. The model was trained on data up to the end of 2020. Everything you see after that is purely out of sample.

Predicting variance of triangular parity deviations in the euro-yen exchange rate

Graph 5



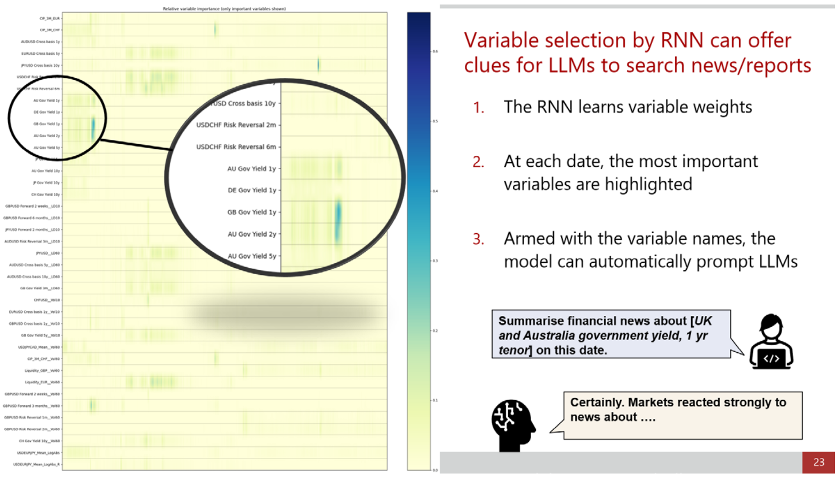
The RNN model's predictions have several interesting characteristics. First, the prediction captures the overall shape of the target variable, but it is less volatile. It is especially less volatile than the predictions from the autoregressive model. This lower volatility is particularly useful for supervisors. It means that changes in the model's forecast are more likely to represent true signals of something significant, rather than just noise.

The model also performs well out of sample: (i) it clearly signals the banking turmoil of March 2023; (ii) the RNN does not forecast the Covid-related stress even in sample, because there were no signals from markets; and (iii) in terms of levels, it is not very close to the actual realisation, but in contrast to the AR(1) model, it correctly signals the increase in probability of stress.

What about the interpretation of the signals?

Using variable selection to guide news search by LLMs

Graph 6



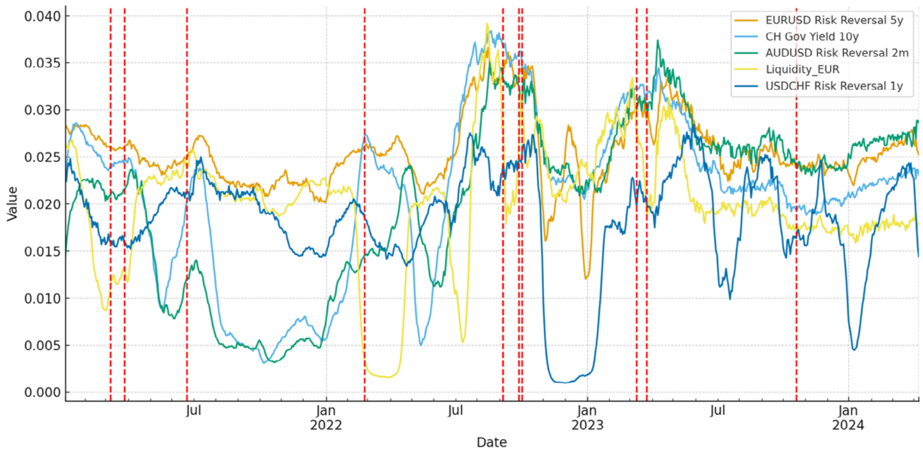
Note: LLM=Large Language Model

We use the novel way to set up the RNN that I described earlier. The RNN creates time-varying weights for each variable, which we then inspect. At any given point in time, the variables with the highest weights are the most important for the model's forecast. In Graph 6, high weights appear in blue.

Then, if the model predicts future stress, we can use the names of the key variables for that particular date to prompt large language models (LLMs) to search through news and reports for relevant information.

So far, this is done manually, but of course one could think about automating this step. This can help identify any brewing issues related to these variables.

Time-varying weights Graph 7



The graph shows how these weights change over time. As explained, they can be used in conjunction with an LLM to sift through news. As a proof of concept, we test this approach using information available in July 2023.

For this case, we use a sophisticated language model, namely Google's Gemini 2.5 pro. We choose this model because, among the key LLMs available, this one was trained with data that went only up to early 2023, which affords us the possibility to check its use in an out-of-sample way for one financial market stress in our study period, namely the "Treasury Tantrum" of October 2023. Other LLMs were trained with more recent data, and we would therefore not rule out the possibility that the model used post-event information.

We collect a large number of financial news from the first half of July 2023 to mimic the information set available at the time of forecast. Then, we prompt the model to read the news and identify and summarise developments that financial supervisors should monitor. We point it to the small set of variables that the RNN had assigned high weights to and ask the model to point us towards which developments a supervisor should look at.

Gemini 2.5 (trained until early 2023)

- **Prompt:** *Attached is a collection of financial-market news in the first half of July 2023. We search for predictive signals across various daily financial market data, and our prediction model flagged unusual signals in USD/CHF 1-year risk reversal, AUD/USD 3-month forward points, and triangular-arbitrage parity (TAP) deviations in USD–EUR–AUD and USD–EUR–MXN. Informed by that, identify and summarize from your corpus the top three developments a financial supervisor should monitor for potential market dysfunction in the coming months. Output: exactly three bullets, ordered by importance/relevance. For each bullet, 1–2 sentences on why it matters. Use only the attached articles. Keep it concise.*

Gemini 2.5 says:

Based on the provided news articles from July 2023, here are the top three developments a financial supervisor should monitor for potential market dysfunction:

- **Diverging views on the U.S. Federal Reserve's monetary policy.** Markets are increasingly pricing in an end to the Fed's rate-hike cycle due to softening inflation data. However, Fed officials have indicated they may favor more rate increases this year, creating a significant gap between market expectations and central bank guidance that could trigger volatility.
- **Broad and rapid depreciation of the U.S. dollar.** The dollar has fallen sharply, hitting a 15-month low against a basket of currencies and an eight-year low against the Swiss franc. This trend is fueled by the view that the Fed's tightening cycle is nearly over, and a continued, rapid repricing of the world's primary reserve currency could lead to broad market instability .
- **Sharp appreciation of emerging market currencies, particularly the Mexican peso.** The Mexican peso has strengthened to its highest level against the U.S. dollar since 2015. This signals potentially crowded carry trades into higher-yielding currencies, which are susceptible to abrupt unwinds if risk sentiment shifts.

Sources

Based on the context from the financial news and the knowledge about which financial market variables provide more signal, the model lists three candidates for monitoring. Interestingly, the first such item relates to the Treasury Tantrum episode.

To summarise the exercise: suppose in July 2023, the analyst sees that the RNN forecasts higher values for triangular arbitrage parity deviations for October 2023, as shown in the earlier graph. Then, the analyst can use an LLM to sift through current information (financial market news as in this example, or other relevant documents) while knowing which financial market variables contributed more to the forecast. The result is a more targeted search process that increases the odds that the LLM provides meaningful responses to help the process of monitoring financial markets.

Before concluding let me summarise what we think is the key contribution of our papers.

We do not claim that we built the ultimate stress forecasting model. But we have shown that AI can be used by supervisors to enhance their monitoring capabilities. Both papers I presented focus on reasonably long-term forecasting windows to allow supervisors to intervene. The focus is to look at problems that may be brewing under the surface.

We also attempt to "open the black box" of AI models. In one case, this is done through the use of Shapley values, and in the other by adjusting the neural network architecture to identify which variables are more relevant. Then LLMs can be used to retrieve the most relevant news and context.

To wrap up:

First, machine learning models show great promise in predicting episodes of market dysfunction and stress.

Second, BIS research demonstrates that these models can help forecast dysfunction across different markets, including government bond and foreign exchange markets.

Third, while the black box nature of machine learning models is often seen as a limitation, it can be addressed.

Fourth and finally, integrating large language models into this process can further help supervisors and regulators understand drivers of potential stress.

Thank you.