A decadal climate service for insurance: Skilful multi-year predictions of North Atlantic hurricane activity and US hurricane damage

Julia F. Lockwood, Nick Dunstone, Leon Hermanson, Geoffrey R. Saville, Adam A. Scaife, Doug Smith, Hazel E. Thornton

\textsuperscript{a} Met Office Hadley Centre, Exeter, UK.

\textsuperscript{b} WTW Research Network, WTW, London, UK.

\textsuperscript{c} Canopius, London, UK.

\textsuperscript{d} University of Exeter, Exeter, UK.

Corresponding author: Julia F. Lockwood, julia.lockwood@metoffice.gov.uk

Early Online Release: This preliminary version has been accepted for publication in Journal of Applied Meteorology and Climatology, may be fully cited, and has been assigned DOI 10.1175/JAMC-D-22-0147.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2023 American Meteorological Society
ABSTRACT

North Atlantic hurricane activity exhibits significant variation on multi-annual timescales. Advance knowledge of periods of high activity would be beneficial to the insurance industry, as well as society in general. Previous studies have shown that climate models initialized with current oceanic and atmospheric conditions, known as decadal prediction systems, are skilful at predicting North Atlantic hurricane activity averaged over periods of 2-10 years. We show that this skill also translates into skilful predictions of real-world US hurricane damages. Using such systems, we have developed a prototype climate service for the insurance industry giving probabilistic forecasts of 5-year-mean North Atlantic hurricane activity, measured by the total accumulated cyclone energy (ACE index), and 5-year-total US hurricane damages (given in US dollars). Rather than tracking hurricanes in the decadal systems directly, the forecasts use a relative temperature index known to be strongly linked to hurricane activity. Statistical relationships based on past forecasts of the index and observed hurricane activity and US damages are then used to produce probabilistic forecasts. The predictions of hurricane activity and US damages for the coming period 2020-2024 are high, with ~95% probabilities of being above average. We note that skill in predicting the temperature index on which the forecasts are based has declined in recent years. More research is therefore needed to understand under which conditions the forecasts are most skilful.

SIGNIFICANCE STATEMENT

The purpose of this article is to explain the science and methods behind a recently developed prototype climate service which uses initialized climate models to give probabilistic forecasts of 5-year-mean North Atlantic hurricane activity, as well as 5-year-total associated US hurricane damages. Although skill in predicting North Atlantic hurricane activity on this timescale has been known for some time, a key result in this article is showing that this also leads to predictability in real-world damages. These forecasts could be of benefit to the insurance industry and society in general.

1. Introduction

When North Atlantic hurricanes reach land, the impacts can be devastating, including loss of life and widespread destruction of homes and infrastructure. They are also one of the leading causes of global insured losses: for example, in 2017, Hurricanes Harvey, Irma and
Maria contributed more than 60% of total global insured losses (including all man-made and natural disasters), and Hurricane Katrina in 2005 has the highest insured loss on record for a single event (82bn USD indexed to 2017; Swiss Re Institute 2018).

North Atlantic hurricane activity is known to exhibit multi-annual timescale variation (e.g. Molinari and Mestas-Nuñez 2003; Chylek and Lesins 2008; Klotzbach and Gray 2008), for example there was notably low hurricane activity in the 1970s-1980s, with associated low US hurricane damages followed by a very active and damaging period from the mid-1990s to mid-2000s (Pielke and Landsea 1998; Nyberg et al. 2007; Weinkle et al. 2018). Prediction of hurricanes on multi-annual timescales is therefore of potential use to insurance companies, allowing them to better prepare for periods of high activity. The 5-year timescale is chosen in particular as it smooths out annual peaks in activity, and since legislation prevents large swings in pricing on interannual timescales it can therefore be used to inform longer term pricing strategies. It also captures the decadal variability, informing insurers how the current climate may differ from their historical records. In fact, following the costly 2004 and 2005 hurricane seasons, three major global catastrophe model vendors developed near-term (5-year) hurricane predictions for use by the insurance industry (Karen Clark & Company 2010; Pielke 2009; AIR Worldwide 2015), demonstrating the clear market interest in predictions on this timescale. These predictions used expert elicitation, statistical models and sea-surface temperature trends (e.g. Jewson et al. 2009; Lloyd’s 2014; AIR Worldwide 2015), but when the high activity predicted for 2006-2010 failed to materialize, two of the companies abandoned their efforts in this area (AIR Worldwide 2015).

To our knowledge, the insurance industry has not yet made use of initialized decadal predictions using physically based climate models, which in the following years were shown to be highly skillful in predicting North Atlantic tropical storm and hurricane activity (Smith et al. 2010; Caron et al. 2014; Caron et al. 2015; Caron et al. 2018). The skill comes from both external forcing and initialization with the current state of the ocean and atmosphere. For external forcing, the high anthropogenic aerosol levels in 1970s-1980s are believed to have caused the hurricane ‘drought’ in that period (Booth et al. 2012; Dunstone et al. 2013; Rousseau-Rizzi and Emanuel 2022). Initialization, on the other hand, particularly enhances sea surface temperature skill in the North Atlantic sub-polar gyre, a key region known to be related to North Atlantic tropical cyclone activity (Goldenberg et al. 2001; Dunstone et al. 2011; Hermanson et al. 2014).
In this paper we describe how we have used a multi-model decadal prediction system ensemble to make probabilistic 5-year forecasts of hurricane activity over the North Atlantic basin (measured by the accumulated cyclone energy, ACE\(^1\)) and the associated US damages (measured in US dollars, adjusted to 2020). Rather than tracking tropical cyclones/hurricanes in the models directly, the forecasts use a relative temperature index known to be strongly linked to hurricane activity (e.g. Vecchi and Soden 2007; Swanson 2008; Vecchi et al. 2008; Ramsay and Sobel 2011; Dunstone et al. 2013). An important aspect of this work is that the index is predicted from the fundamental dynamics of the climate model, unlike the aforementioned near-term predictions developed by the insurance industry which used statistical methods to predict sea surface temperatures. Statistical relationships between past forecasts (hindcasts) of the index and observed hurricane activity and US damages are then used to produce probabilistic forecasts of these measures.

The rest of the paper is structured as follows. In Section 2 we describe the model and observational data used for this study, and in Section 3 we outline the statistical verification methods used. Section 4 describes the methods used to generate the ACE and damage forecasts and gives an assessment of their skill, and describes development of the final forecast product. In Section 5 we discuss the limitations of the predictions and present the conclusions.

2. Data

a. Model (forecast) and re-analysis data

A 40-member multi-model ensemble consisting of 4 dynamical decadal prediction systems (10 members each) was used to make the hurricane predictions. The systems used are based on fully-coupled atmosphere-ocean dynamical models: HadGEM3-GC3.1

\(^1\) ACE is defined as \(10^{-4} \sum_{i,t} v_{i,t}^2\), where \(v_{i,t}\) is the estimated maximum 1-minute wind speed of cyclone \(i\) at 6-hourly time interval \(t\), when the cyclone is at tropical storm intensity \((v_{i,t} > 33 \text{ kn})\) or greater.
(Williams et al. 2017), CMCC-CM2-SR5 (Cherchi et al. 2019), EC-Earth3 (Bilbao et al. 2021), and MPI-ESM-HR (Müller et al. 2018).

Each ensemble member is initialized with the current state of the atmosphere, ocean and land surface every November from 1960-2018 to form a set of retrospective forecasts (hindcasts), and run for 10 years. The forecast for 2020–2024 is calculated from the runs initialized in November 2019. All models use the full-field initialization method for the atmosphere and ocean, with the exception of MPI-ESM-HR which uses anomaly initialization in the ocean. All models are forced with the CMIP6 historical greenhouse gas, aerosol and natural forcings from 1960-2014, and the SSP2-4.5 scenario thereafter (O’Neill et al. 2016). A full description of the models is given in the C3S Sectoral Applications of Decadal Predictions Technical Appendix at https://climate.copernicus.eu/sites/default/files/2021-09/Technical_appendix_2020.pdf (see Müller et al. 2018 for details of the MPI-ESM-HR model).

To calculate the relative temperature index described in Section 3, we extract monthly mean near surface temperature (tas) for years 1 to 5 of the forecast for the North Atlantic hurricane season (June – November; referred to as JJASON). Ideally sea surface temperature would have been used but this data was not readily available at the time the analysis was undertaken. This has minimal impact on the results, since the sea surface temperatures and tas are strongly coupled on 5-year timescales, especially in the tropics (their correlations in DePreSys4 members over the MDR and TROP regions (defined in Section 4), where both variables were available are >0.99).

The ERA5 re-analysis dataset (Hersbach et al. 2020) is used as the ‘observations’ to assess the skill in predicting tas and the relative temperature index. To assess the observed relationship between the relative temperature index with ACE and hurricane damages over a longer time period we use sea surface temperature data from the HadISST data set (Rayner et al., 2003), for 1900–2020.

b. Hurricane and loss data

North Atlantic tropical cyclone and hurricane numbers and total accumulated cyclone energy (ACE) are from the Atlantic HURDAT2 database (National Hurricane Center HURDAT2, accessed 2020; Landsea and Franklin 2013), for cyclones formed in the North Atlantic basin over the hurricane season (JJASON), over the period 1900–2020. We have
attempted to correct for missing tropical cyclones in the pre-satellite era (pre-1965) using the estimated missing tropical cyclone numbers from Vecchi and Knutson (2008), and assuming the missing cyclones have average ACE.

US hurricane damages, which include both insured and un-insured losses, are compiled from the NOAA NCEI U.S. Billion-Dollar Weather and Climate Disasters database (Smith and Katz 2013) for the period 1980-2020. Damages from 1961 to 1979 are compiled from the Annual Summaries of North Atlantic Storms (NOAA Miami Regional Library archive, 2021).

The damages are adjusted to 2020 equivalent US dollars, taking into account inflation and changes in wealth and population, using the method described in Pielke and Landsea (1998) and Weinkle et al. (2018; hereafter referred to as W18). Assuming that the populations of the counties affected by hurricanes follow the same proportional changes as the population of the US as a whole and using gross domestic product the measure of wealth, the equation in Pielke and Landsea (1998) simplifies to:

\[ NL_{2020} = L_y \times \frac{GDP_{2020}}{GDP_y} \]

where \( NL_{2020} \) is the annual damage normalised to 2020, \( L_y \) is the loss recorded in year \( y \), and \( GDP_{2020} \) and \( GDP_y \) are the nominal gross domestic products (i.e. the values recorded at the time) in 2020 and year \( y \) respectively. This simplification arises because the nominal GDP ratio implicitly includes the population and inflation increase. Gross domestic product data is from the Bureau of Economic Analysis (2021).

\[ \]

2 Use of US population rather than that of the affected counties, and GDP as a measure of wealth instead of ‘fixed reproducible tangible wealth’ as in Pielke and Landsea (1998), has little effect on the correction factor: the correlation coefficient between the 5-year-total losses in this work and the updated losses presented in W18 (which use the original Pielke and Landsea 1998 equation) is 0.96 over the common time period.
We also use the normalized loss data from W18, covering the period 1900–2017, to assess the relationship between damages and the relative temperature index in observations over this longer time period. The two damage data sets are very similar for the over-lapping years (see Section 4c). The largest discrepancy is for the damages of 2017, since the W18 data set only includes continental U.S. and thus excludes the damages from Hurricane Maria, which caused major damage to Puerto Rico.

3. Verification Methods

Following the C3S Sectoral Application of Decadal Forecasts “Recommendations on forecast quality assessment for decadal predictions”, we include assessment of both the deterministic and probabilistic forecast skill (see also Delgado-Torres et al. 2022). We use the Spearman rank correlation coefficient, $\rho$, to assess the deterministic skill of the forecasts. This is to take into account the non-linear relationship between US damages and forecast median US damage (see Section 4.3), although the standard Pearson correlation coefficient gives similar results.

We use the Brier skill score (BSS) for measuring the probabilistic skill of predicting above or below average ACE and US damages, defined as:

$$BSS = 1 - \frac{BS_{fc}}{BS_{ref}},$$

where $BS_{fc}$ and $BS_{ref}$ are the Brier scores of the forecast and a reference forecast. The Brier score is defined as

$$BS = \frac{1}{n} \sum_{k=1}^{n} (f_k - o_k)^2,$$

where $n$ is the number of the observed–forecasted pairs, $f_k$ is the forecast probability (range 0–1) of above average ACE/damage for the 5-year period $k$, and $o_k$ the observed occurrence ($o_k = 1$ if the observed ACE/damage was above average, 0 otherwise). A BSS of 1 indicates a perfect forecast, and BSS > 0 indicates that the forecast performs better than the reference. Here the reference used is 5-year persistence, where the forecast probability of above average ACE/damage is equal to 1 if the previous 5 years had above average ACE/damage.

Contingency tables for forecasting above or below average ACE/US damage are also shown. A forecast is considered to have predicted above average ACE/loss if the forecast
probability is greater than 0.5. The hit rate is the ratio of correct forecasts to the number of times this event occurred (hits/(hits+misses)) and the false alarm rate is the ratio of false alarms to number of non-occurrences of the event (false alarms/(false alarms+correct rejections)). When the hit rate is high and the false alarm rate is low, the forecast is said to exhibit discrimination, i.e. the ability to distinguish situations leading to the occurrence of an event from those leading to the non-occurrence of the event (Wilks 2019b). Forecasts with good discrimination can be useful for decision making, although required hit rate/false alarm thresholds depend on the application of the forecasts and the cost of action or in-action.

The correlations and Brier Skill Scores are tested for statistical significance using the nonparametric block bootstrapping approach as described in Smith et al. (2020), using a block length of 5 years to take into account auto-correlation of the data.

As the forecast probability distributions are based on a statistical relationship between past forecasts of the temperature index and observations, we use cross-validation to assess out-of-sample skill. The ACE/damages for each 5-year period are estimated from the statistical relationship between observed ACE/damages and forecast temperature index for data points with no common years with the period (see Section 4b).

4. ACE and US damage forecasts

To make the ACE and damage predictions, we first calculate a relative temperature index from the ensemble mean of the decadal prediction systems as an indicator of predicted hurricane activity. A description of the index and its predictability is described in Section 4a. The statistical models described in Sections 4b and 4c then transform the index into probabilistic 5-year-mean ACE index and 5-year-total total US damage predictions. In Section 4d we describe the final forecast product.

a. Relative temperature index

The relative temperature index used to indicate North Atlantic hurricane activity, MDR-TROP, is defined as:

$$\text{MDR-TROP} = T_{\text{MDR}} - T_{\text{TROP}}$$

where $T_{\text{MDR}}$ is the 5-year mean (JJASON only) temperature anomaly in the North Atlantic hurricane main development region (MDR; 10°N to 25°N, 80°W to 20°W), and $T_{\text{TROP}}$ is the 5-
year mean (JJASON only) temperature anomaly over the tropical oceans (averaged over 30°S to 30°N).

This index, and similar variants, have often been used to estimate annual and multiannual North Atlantic hurricane activity at this timescale (e.g. Vecchi and Soden 2007; Swanson et al. 2008; Vecchi et al. 2008; Vecchi et al. 2011; Dunstone et al. 2013; Caron et al. 2018). Its use is justified by considering the conditions necessary for hurricane formation: warm local sea surface temperatures and moist, unstable air, low vertical windshear, and high vorticity (Gray 1998; Emmanuel and Nolan 2004; Emmanuel 2010). All of these conditions can be satisfied in the MDR, which lies on the northward flank of the intertropical convergence zone (ITCZ). Although high surface temperatures increase the probability of hurricane formation locally, high temperatures elsewhere in the tropics, for example the Pacific or Indian Ocean, lead to increased atmospheric stability and increased wind shear in the tropical Atlantic due to changes in the Walker circulation, impeding hurricane development (Latif et al. 2007). In addition, Dunstone et al. (2013) showed that positive values of MDR-TROP are associated with a northward shift in the ITCZ and Hadley cell (and hence the subtropical jet), again reducing vertical wind shear in the MDR. The increase in vorticity with a more northerly ITCZ (due to the increased planetary contribution) also favors hurricane formation (Merlis et al. 2013, Sobel et al. 2021).

We note that there has been some debate in the literature on whether decadal hurricane predictions based on this relative SST index would have been able to predict the shift from low to high hurricane activity in the mid-1990s (Vecchi et al. 2013). However, the analysis present in Smith et al. (2014) shows that such systems would have provided clear evidence for an impending reversal to a period of above average hurricane frequency had they been available in the mid-1990s, before the observed increase occurred.

The correlation between 5-year-mean observed tas and ACE over the hindcast period is shown in Fig 1(a), showing a strong positive relationship between ACE and tas in the North Atlantic. The correlation in this region is still present after linearly detrending the data (Fig 1(c)), and is particularly strong in the MDR region (marked by the black box) and the subpolar gyre region in the far North Atlantic (around Iceland and the southern tip of Greenland). Figures 1(b) and (d) show the correlation skill in predicting 5-year-mean tas. Correlations are generally very high across the globe and part of this skill can be attributed to the warming trend, but Fig 1(d) shows the much of the predictive skill in the North Atlantic.
remains after linearly detrending the data. The detrended skill is particularly strong in the North Atlantic MDR and subpolar gyre regions, which are the regions most important for predicting hurricane activity as shown in Figs 1(a) and (c).

Figure 1: Relative temperature index (MDR-TROP) as predictor of hurricane activity. (a) Correlation of observed 5-year-running-mean tas and observed 5-year-running-mean North Atlantic hurricane activity (ACE) over the model hindcast period; (b) Correlation of predicted 5-year-running-mean tas (from multi-model ensemble mean) and observed 5-year-running-mean ACE; (c) and (d) as in (a) and (b), but with all time series linearly detrended before calculating correlations. In panels (a) to (d), the Pearson correlation coefficient is shown, and ERA5 is used for the observations; (e) Time series of observed observed 5-year-mean ACE (black, showing the raw HURDAT data with the dotted line, and the corrected HURDAT data with the solid line), and 5-year running mean MDR-TROP calculated from ERA5 tas (thick blue line), and HadISST SST (thin blue line); (f) Time series of observed (ERA5) and predicted 5-year-running-mean MDR-TROP index (ensemble mean). The rank correlation coefficient between time series is shown in legends in panels (e) and (f).

Fig 1(e) shows the time series of observed 5-year-mean MDR-TROP and North Atlantic ACE, showing an overall correlation of 0.70 (using ERA5 data, $p<0.05$) over the hindcast
period. Both measures show positive trends over the hindcast period, although the longer-term time series from 1900-2018 (using HadISST SST data to calculate MDR-TROP), show that the long-term trend is small. The correlation holds over this longer period ($\rho=0.61$), with both time series showing troughs for approximately 1900–1920 and 1965–1995, and peaks for approximately 1920–1965 and after 1995. This adds further evidence that there is a relationship between MDR-TROP and hurricanes, rather than the correlation being caused by positive trends in both measures.

The correlation of the multi-model ensemble mean in predicting MDR-TROP is 0.67 ($p<0.05$) (Fig 1(f)), demonstrating significant skill in predictions of this index. Given the high correlations shown in Figs 1(e) and (f), we expect the predicted MDR-TROP by the models to be a skilful predictor of observed ACE, which is indeed the case: The correlation between the predicted MDR-TROP and observed ACE is 0.77 ($p<0.01$); see Section 3b).

We note that Fig 1(f) shows the MDR-TROP is poorly predicted at the end of the time series (central years 2014—2018). During this period the observed MDR-TROP and observed ACE time series also diverge. This is discussed further in Section 5.

We also investigated the skill of several other North Atlantic proxy hurricane indices which have previously been mentioned in the literature, including the lambda index (which is also based on the MDR-TROP relative temperature; Vecchi et al. 2011; Caron et al. 2014; Caron et al. 2015; Caron et al. 2018), the Atlantic Multidecadal Oscillation (AMO) index developed by Klotzbach and Gray (2008), and a relative temperature index based on the temperatures in the subpolar gyre region (-50°E to -10°E; 50°N to 60°N; Klotzbach and Gray 2008; Caron et al. 2018) and the tropics (SPG-TROP, with TROP defined as above). We find little difference in the skill between the above indices in predicting North Atlantic ACE, with all model predicted indices having a correlation with observed ACE of approximately 0.7–0.8. In addition, all indices show the same discrepancy between model and observed values for 2014–2018 as found for MDR-TROP.

b. ACE forecasts

The relationship between observed 5-year-mean ACE index and 5-year-mean forecast MDR-TROP is quantified using ordinary least squares linear regression on past data. Rather than using the ensemble spread to estimate the forecast ACE prediction intervals, the prediction intervals are estimated from the residuals of the linear regression fit (Wilks 2019a, Accepted for publication in Journal of Applied Meteorology and Climatology. DOI 10.1175/JAMC-D-22-0147.1).
This method has successfully been used in seasonal forecasts (e.g. Bett et al. 2018; Bett et al. 2022), and has the advantage that it implicitly recalibrates the forecast probability distribution to the observed variability.

**Figure 2: ACE forecast.** (a) Scatter plot showing observed 5-year-mean hurricane activity (ACE) against the predicted 5-year-mean MDR-TROP from the ensemble mean of the decadal prediction systems. The rank correlation and Brier skill scores are shown in the top left. The solid line shows the linear regression. (b) Time series of observed and forecast 5-year-mean ACE. On each panel the shading shows the 75% and 95% prediction intervals. The forecast for 2020-2024 is shown with the red dot, with the box and whiskers showing the 75% and 95% prediction intervals.

The linear regression fit is performed on non-overlapping 5-year-mean data, to avoid over-confidence in the prediction intervals. There are 5 sets of non-overlapping 5-year-means (for example the first set is the means of 1961-1965, 1966-1970, 1971-1975, etc; the second set is the means of 1962-1966, 1967-1971, 1972-1976, etc). Each set has 11–12 data points (10–11 when performing cross-validation). The linear-regression is performed on each set separately, and we use the mean fit coefficients for the line of best fit and prediction interval estimation. The linear regression and 75% and 95% prediction intervals are shown in Fig 2(a), and the resulting predicted and observed ACE time series is shown in Fig 2(b).

The correlation between forecast and observed ACE is \( \rho=0.77 \) (\( p<0.01 \)), and 0.72 (\( p=0.01 \)) using cross-validation. The correlation remains significant after linearly detrending both variables (\( \rho=0.68 \)). The skill of persistence (using the last 5 years to predict the next 5 years) is \( \rho=0.42 \), significantly lower than the model forecasts (the difference has \( p \)-values of 0.02 and 0.12 with and without cross-validation, respectively). The Brier skill score of
predicting above or below average ACE, using 5-year persistence as the reference forecast, is 0.62 ($p=0.02$), and 0.56 ($p=0.06$) using cross-validation. The contingency table for predicting above average ACE (Table 1) demonstrates the forecasts are able to accurately discriminate between above and below average ACE index, with a high hit rate (80%) and low false alarm rate (10%).

<table>
<thead>
<tr>
<th>Above-average ACE index</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Predicted</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>23 Hits</td>
</tr>
<tr>
<td>No</td>
<td>5 Misses</td>
</tr>
</tbody>
</table>

Hit rate 80%

False alarm rate 10%

Table 1: Contingency table for predicting above average ACE index. The average is defined as the mean observed ACE over the validation period (1960-2018), equal to 98.4 $10^4$km$^2$.

The forecast for 2020–2024 (initialization date November 2019) is shown with the box-and-whisker on Figs 2(a) and (b). The forecast is for a very active period, with a 5-year-mean prediction of 145 $10^4$km$^2$ (75% prediction intervals 110–180 $10^4$km$^2$; 95% prediction intervals 85–205 $10^4$km$^2$) and a 95% chance of above average ACE. Note that there are some caveats to this forecast, discussed in Section 5.

c. US damage forecasts

As for the ACE forecast, we use a statistical relationship between observed damage and the forecast MDR-TROP index based on past data to make the damage prediction. The time series of 5-year-total US hurricane damage over the hindcast period is shown in Fig 3(a), along with the time series of observed and forecast MDR-TROP, and the scatter plot of 5-year US hurricane damage (on a logarithmic scale) against forecast MDR-TROP is shown in Fig 3(b). Although the rank correlation coefficient with forecast MDR-TROP is high (0.70, $p<0.01$), the relationship is non-linear, as can be seen in the scatter plot of 5-year-total loss.
Two prominent jumps are seen in the loss time series, centred on the years 1992 and 2005. These peaks are caused by Hurricanes Andrew (1992) and Katrina (2005), and they also appear as outliers on the scatter plot in Fig 3(b), where points containing these years have been highlighted. It could be argued that data containing the 2017 hurricane season also have anomalously high damage, but in this case it was not a single event causing the large losses.

The observed 5-year US damages and observed MDR-TROP are plotted over the long-term period 1902–2015 (central years) in Fig 3(c) (using W18 US damage data and HadISST MDR-TROP). This panel confirms that the relationship holds over a longer time period and with alternative data sets, and that the high correlation is not simply due to positive trends over the hindcast period. In the pre-hindcast period, a further three jumps are seen in the damage time series, caused by the cyclones Galveston (1900), Galveston (1915), the Great Miami storm (1926). The correlation between observed 5-year US damages and MDR-TROP is 0.34 for longer period 1902–2015, but this increases to 0.66 after removing the damages from the five outlier storms.
Figure 3: US damage forecast. (a) Timeseries of observed 5-year-total US damages (adjusted to 2020) and observed (ERAS) and predicted MDR-TROP; (b) Scatter plot of 5-year-total US damages (log-scale) against forecast 5-year-mean MDR-TROP, highlighting 5 year periods which contain the damages from Hurricanes Andrew (points with black circles) and Katrina (red circles). (c) Long-term time series of observed MDR-TROP (calculated from HadISST) and 5-year-total US damages (adjusted to 2018; W18 data). The solid black line (‘NoExt’) shows the data after damages from the extreme storms Galveston (1900), Galveston (1915), Great Miami (1926), Andrew (1992) and Katrina (2005) have been removed; (d) Scatter plot of hindcast and 2020-2024 forecast 5-year-total US damage (the data is the same as in (b) but on a linear scale); (e) Time series of hindcast and 2020-2024 forecast 5-year-total US damage. In (d) and (e) the solid line shows the median damage prediction, the shading shows the 75% and 95% prediction intervals, and the horizontal dashed line marks the long term median damage. The forecast for 2020-2024 is shown with the red dot, with the box and whiskers showing the 75% and 95% prediction intervals.

Given the apparent strong relationship between log(damage) and predicted MDR-TROP index, a simple linear regression between these measures was attempted. However, the
resulting prediction intervals appeared unrealistically large for high values of forecast MDR-TROP: For example, for a standardized forecast MDR-TROP value of 2.25 (the highest in the time series), the 95% prediction upper limit was $2.2tn, equivalent to 11 Hurricane Katrinas happening in a single 5-year period. A generalized linear model assuming a gamma distribution was also fit to the data using the Python Statsmodels package (Seabold and Perktold 2010), but again a satisfactory fit and prediction intervals could not be obtained. This is possibly because of the constraint of a constant shape parameter in the gamma distribution with MDR-TROP.

<table>
<thead>
<tr>
<th>Above-median US damages</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Predicted</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Hits</td>
</tr>
<tr>
<td>No</td>
<td>7</td>
</tr>
<tr>
<td>Hit rate</td>
<td>75%</td>
</tr>
<tr>
<td>False alarm rate</td>
<td>30%</td>
</tr>
</tbody>
</table>

Table 2: Contingency table for predicting above median 5-year-total US damages (adjusted to 2020). The median observed 5-year-total US damages (adjusted to 2020) is 57.9 bn USD, calculated over the validation period (1960-2018).

The difficulty of fitting loss probability distributions for extreme hazards is well known (Lloyd’s 2014), and is one of the reasons why insurers resort to using catastrophe modelling, where thousands of extreme events and their resulting losses are simulated. One issue with hurricanes is that hurricanes with similar ACE can inflict vastly differing amounts of damage, for example small changes in paths or size could lead to large differences in exposure (Czajkowski and Done 2014), or if one manages to breach flood defenses as in the case of Hurricane Katrina (Anderson et al. 2007). Such detailed catastrophe modelling is beyond the scope of this project. Instead, we have devised a method based on linear regression and random sampling to produce damage probability distributions for a given forecast MDR-TROP which takes into account the unpredictable nature of extreme single events such as those which cause the jumps in the damage time series, whilst retaining a dependency on MDR-TROP. The method is described below.
First we remove the damages attributed to Andrew and Katrina from the data and perform linear regression of log(US damage) onto forecast MDR-TROP. We then assume that such extreme single events occur randomly at the climatological frequency (2 events in 60 years of data). To incorporate this into the predicted probability distributions, for a given MDR-TROP we estimate the ‘base’ damage probability distribution from the linear regression, then randomly sample a ‘base’ total damage for the 5-year period. We then loop through each of the 5 years in the period, randomly sampling whether an extreme event occurs. If an extreme is selected, then its loss is added to the base damage. This gives the possibility of having more than one extreme damage hurricane in a single 5-year period. The value of damage for the extreme is sampled from a uniform probability distribution between 90bn\(^3\) and 210 bn USD (which are the approximate damages of Andrew and Katrina respectively, after adjusting to 2020). This is repeated 5000 times for each MDR-TROP to build up a probability distribution as a function of MDR-TROP.

The resulting damage hindcast and forecast prediction intervals are shown in the scatter plot in Fig 3(d) and time series in Fig 3(e). Although the method has involved estimations of unknown parameters such as the frequency and distribution of the extremes, it qualitatively reproduces the distributions we expect, i.e. a low expected damage for low values of MDR-TROP, but with heavy tails indicating an ever-present risk of a random extreme, and high expected damages for high values of MDR-TROP, but without the upper limit of the distribution giving unrealistically high estimations. Note that the resulting predicted median loss and probability of above average damages are insensitive to the choice extreme distribution: they remain relatively unchanged even after doubling the extreme upper limit to $400bn and doubling the frequency of extreme storms to 4 in 60 years. However, the predicted upper damage limits for low MDR-TROP years are sensitive to the choice of

\[3\] For low forecast values of MDR-TROP, where the loss probability distribution from the linear regression is very narrow and centered on losses \(\ll $90bn, the lower limit of the ‘extreme’ is decreased to the 95% upper limit of damage from the linear regression. This avoids having a gap in the final probability distribution.

File generated with AMS Word template 2.0
extreme distribution. This uncertainty is stated in the product and more research is needed to understand the probabilities of such extreme events.

The rank correlation coefficient of the median loss prediction is 0.70 ($p<0.01$), and with the cross-validation it is 0.66 ($p<0.01$). The Brier skill score for predicting above average loss is 0.53 ($p<0.01$), and 0.50 with cross-validation ($p=0.01$). Note that the cross-validated skill scores have used knowledge of Hurricanes Andrew and Katrina for estimated the probability and distribution of extreme outlier events, but since the median and probability of above average damages are insensitive to the extremes, this should have little effect on the cross-validated scores. Table 2 shows the contingency table for predicting above average US damages, demonstrating good discrimination with a high hit rate and low false alarm rate.

The prediction for 2020–2024 is for total US damages of 230bn USD (approximately 46bn USD per year), with 75% and 95% prediction intervals of 100–475bn USD and 55–770bn USD respectively, and a greater than 95% chance of being above average.

d. Forecast product

This work was part of the European Union funded Climate Copernicus Climate Change Service (C3S) project “Sectoral applications of decadal predictions”, which was set up to develop climate services on decadal timescales (Dunstone et al. 2022) for users in agriculture (Solaraju-Murali et al. 2021; 2022), water management (Paxian et al. 2022), energy (Tsartsali et al. 2023) and insurance (this work). Up until this point there were very few services on this timescale, despite the skill of decadal predictions demonstrated in scientific literature and clear user needs (Kushnir et al. 2019).

This service was co-developed with the insurance broker Willis Towers Watson. User needs were gathered in an initial meeting, and research results and drafts of the forecast product were shared throughout the project. Feedback on earlier versions of the forecast product included the need to make the forecast more accessible by reducing the use of meteorological jargon. The final product is a 2-page pdf document (available at https://climate.copernicus.eu/decadal-predictions-insurance), which follows the common template for all case studies in this project. Page 1 of the document gives the headline results, namely that there will be 95% chance of above-average North Atlantic hurricane activity and >95% chance of above average US damages for the period 2020–2024. This is followed by a more detailed forecast (giving the forecast ACE and US damage median and
75% and 95% prediction intervals, and the time series shown in Figs 2(b) and 3(c) to put the forecast into context). The second page briefly describes the method to generate the forecasts and presents forecast verification measures (rank correlation, Brier skill score and contingency tables). Note that the forecast statistics on the website differ slightly to those presented here, since more ensemble members became available during the writing of this paper.

The product was well received by the users, and they showed an interest in continuing the service and exploring ways to apply the information to business practices. In terms of usability, their view was that there would have to be a strong forecast signal, outweighing other market influences, to make a difference to business overall. Nevertheless, forecast products such as this can provide some guidance for how insurers may think about hurricane risk in the next few years, and can become part of the underwriting conversation in selecting which lines of business to grow or shrink. If predictions were available at higher resolutions, such that locations could be assessed across a portfolio of risks, then further utility could be derived, however, this is currently beyond the scope of this product. More information on probabilities of extreme periods was also requested, although for the damage model in particular more research and verification is needed to be confident in the predicted upper percentiles, as these are highly dependent on extremes like Hurricanes Andrew and Katrina. Another suggestion for future versions of the forecast was to show predictions of the number of landfalling hurricanes or number of damaging events (defined as the number of events for which there is a recorded loss, which is similar, but not exactly equal to, the number of landfalling hurricanes). There is significant skill in predicting both of these measures ($\rho=0.47, p <0.05$ and $\rho =0.50, p <0.01$ respectively).

5. Discussion and conclusions

We have used decadal predictions systems to make probabilistic forecasts of North Atlantic hurricane activity and US damages, and have demonstrated the predictions are skilful and able to discriminate between periods of below and above average activity. Figures 1(e) and (f), however, show two issues: Firstly, the time series of observed 5-year-mean MDR-TROP and ACE diverge at the end of the time series, particularly for periods with central years 2014–2018 (Fig 3(a) shows the damage time series also diverges from observed MDR-TROP during this period); and secondly, the model predictions of MDR-TROP are poor during the same period. The model error in predicting MDR-TROP compensates for the divergence.
between observed MDR-TROP and ACE/damages, giving apparently skilful predictions during this period (Figs 2(b) and 3(c)), which may indicate that the correlations between predicted MDR-TROP and observed ACE/US damages over-estimate the true skill.

Assuming a two-stage framework where the observed ACE is linearly related to the observed MDR-TROP, which is linearly related to the predicted MDR-TROP, the final correlation between observed ACE and predicted MDR-TROP should be the product of the correlations of the individual stages (ie. \( \rho(\text{forecast MDR-TROP, obs ACE}) = \rho(\text{forecast MDR-TROP, obs MDR-TROP}) \times \rho(\text{obs MDR-TROP, obs ACE}) \); Kretschmer et al. 2021). This would give estimates of correlation skill of 0.47 for ACE and 0.31 for damages. Both of these skill estimations are lower than the values obtained from modelling ACE/US damages on predicted MDR-TROP directly, but note that there is large uncertainty in these values, owing to the large uncertainty in the correlations for each stage.

![Figure 4: Performance of MDR-TROP index.](image)

(a) Time series of single season observed ACE and ACE derived from observed MDR-TROP

(b) Obs tas anom 2014-2018 (JJASON)

(c) Model tas anom y1-5 2014-2018 (JJASON)
observed MDR-TROP (dashed line). The vertical lines mark the 2017 and 2018 seasons, and the rank correlation score is shown in the legend. (b) and (c) Observed and forecast 5-year-mean tas anomaly for 2014–2018 (JJASON) respectively. Anomalies are with respect to 1981–2010. The boxes show the MDR and TROP regions as in Fig 1. The subpolar gyre (SPG) region (-50°E to -10°E; 50°N to 60°N) is also marked.

To investigate the divergence between observed 5-year-mean MDR-TROP and observed 5-year-mean ACE, in Fig 4(a) we plot the seasonal values of observed ACE alongside ACE predicted from the observed MDR-TROP (using linear regression). ACE is still well predicted by observed MDR-TROP on this timescale ($\rho=0.56$), although a persistent negative bias appears in the last six seasons, with notably large errors in 2017 and 2018. This bias is also present when using the other hurricane proxy indices to predict ACE described in Section 4b. The reasons for the bias are currently unclear, but for such a short time period it is insufficient evidence to definitively conclude non-stationarity in the relationship between ACE and MDR-TROP. However, hurricane and tropical cyclone genesis is not fully understood (Sobel et al., 2021; Studholme et al 2021), and the link between observed hurricane activity and proxy indices such as the simple relative temperature index used here, or even more complex genesis potential indices may change in a warming climate (e.g. Wang et al., 2020). It will be necessary to monitor the performance of the index in the coming years.

With regards to the poor predictions of MDR-TROP late in the time series, Figs 4(b) and (c) show the observed and predicted temperature anomalies for the 5-year period with the largest error, 2014–2018. The observations show notable cool anomalies in the subpolar gyre region, which were driven by the years 2015 (Duchez et al. 2016) and 2018 (Dunstone et al. 2019). Temperatures in the subpolar gyre are strongly linked to the MDR (Smith et al. 2010), and it is indeed overprediction of the MDR temperatures that are causing the errors in the MDR-TROP index (not shown). Despite decadal prediction systems generally having high skill in predicting SPG temperatures (e.g. Smith et al. 2019), this recent period of cooling has clearly not been well predicted. Cool anomalies in the SPG can arise through various mechanisms (Robson et al. 2018) and therefore some SPG temperature transitions may be more predictable than others. The reasons for the cooling in this recent period and why it was not captured by the models are under investigation.

A further caveat to these predictions is that recent studies have shown climate change may be exacerbating the damage caused by individual hurricanes, through increased precipitation, slower moving systems, and higher sea-levels leading to higher probabilities of
coastal flooding (e.g. Risser and Wehner 2017; Trenberth et al. 2018; Zhang et al. 2020; Guzman and Jiang 2021; Strauss et al. 2021). This is difficult to account for in the damage time series, and may lead to errors in estimating the true skill in predicting hurricane damage. On the other hand, several studies have shown no trend in hurricane damages (after removing the effects of growth in population and wealth) when considering longer time periods (e.g. Weinkle et al. 2018; Klotzbach et al. 2018), which can be seen in Figure 3(c). Furthermore, a positive correlation between forecast MDR-TROP and observed damages remains after linearly detrending both variables ($\rho=0.38$, where the damage detrending has been performed on the logarithm of the damages). Although this correlation is no longer statistically significant, detrending will have removed any increase over the hindcast period due to decadal variability and thus could be removing genuine skill. It will be necessary to re-evaluate this skill as time progresses and more data becomes available.

To conclude, we have demonstrated skilful 5-year predictions of North Atlantic hurricane activity (as measured by the total ACE index) using ensembles of initialized climate predictions. We have shown that this translates into predictability of real-world US damages. The predictions are made using a relative temperature index from the model ensemble mean, and then using linear regression of past observations onto the index to calculate probabilistic forecasts. The regression method is modified for the damage predictions to take into account the random nature of very extreme loss events such as Hurricanes Andrew and Katrina. A prototype climate service for the insurance industry has been co-developed with users based on this work (Dunstone et al. 2022), which is publicly available at https://climate.copernicus.eu/sectoral-applications-decadal-predictions. The predictions of ACE and US damages for the period 2020–2024 are high, both with ~95% probabilities of being above average. We note that skill in predicting the temperature index on which the forecasts are based has declined in recent years. More research is therefore needed to understand under which conditions the forecasts are most skilful.

Acknowledgments.

The authors would like to thank two anonymous referees whose comments led to considerable improvements to this manuscript. This study received support from the

File generated with AMS Word template 2.0
C3S_34c contract (number: ECMWF/COPERNICUS/2019/C3S_34c_DWD) of the Copernicus Climate Change Service (C3S) operated by ECMWF.

Data Availability Statement.

The datasets used in this study are available from the following sources:


• Decadal prediction systems: Hindcast data is available from ESGF, https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/DCPP/index.html. Forecast data can be accessed by request to the corresponding author.
REFERENCES


Guzman, O., and Jiang, H., 2021: Global increase in tropical cyclone rain rate. *Nat. Commun.*, 12, 5344. [https://doi.org/10.1038/s41467-021-25685-2](https://doi.org/10.1038/s41467-021-25685-2).


File generated with AMS Word template 2.0


File generated with AMS Word template 2.0


Solaraju-Murali, B., Bojovic, D., Gonzalez-Reviriego, N., Nicodemou, A., Terrado, M., Caron, L.-P., and Doblas-Reyes, F. J., 2022: How decadal predictions entered the climate

File generated with AMS Word template 2.0

Accepted for publication in Journal of Applied Meteorology and Climatology. DOI 10.1175/JAMC-D-22-0147.1


