Climate Dynamics
https://doi.org/10.1007/s00382-022-06289-y

Assessing the large-scale drivers of precipitation in the northeastern United States via linear orthogonal decomposition

Raymond Sukhdeo[1] · **Paul A. Ullrich**[1] · **Richard Grotjahn**[1]

Received: 30 July 2021 / Accepted: 2 April 2022

Abstract
This study examines the linear orthogonal modes associated with monthly precipitation in the northeastern United States, from CESM1 LENS (35 ensemble members, 1979–2005) and two reanalysis datasets (ERA5, 1950–2018 and NOAA-CIRES-DOE 20CRv3, 1950–2015). Calendar months are aggregated together, and any linear trends in data are removed. Using region-averaged precipitation anomaly time series and monthly anomalies for several global 2D atmospheric fields, a linear orthogonal decomposition method is implemented to iteratively extract time series (based on field and geographic location) of absolute maximum correlation. Linear modes associated with this method are then projected onto the full set of 2D fields to provide physical insight into the mechanisms involved in generating precipitation. In this region, the first mode is associated with vapor transport from the Atlantic seaboard, the second mode is characterized by westward vapor transport associated with extratropical cyclones, and the third mode captures vapor transport from the Gulf of Mexico during the fall and winter. However, the third mode is less robust in the spring and summer. Results are generally consistent across the datasets, and applying multiple linear regression with the linear modes to predict the precipitation anomalies produces R-squared values of around 0.54–0.65 for CESM1 LENS, and around 0.58–0.88 for reanalysis, with the lowest values generally in the spring and late summer. The influence of low-frequency climate variability on the modes is considered for CESM1 LENS, and the modes in late winter can be predicted with some success via a combination of several, prominent large-scale teleconnection patterns.

Keywords Northeastern U.S. precipitation · Linear orthogonal decomposition · Linear modes · Teleconnections · Large-scale circulation

1 Introduction

The northeastern United States (hereafter the Northeast) is home to a dense human population and encompasses a variety of agricultural and economic interests that are reliant on the available water resources and the replenishment of those resources via precipitation. However, water availability in this region is expected to be altered under climate change (Melillo et al. 2014). These trends are particularly important at the ends of the precipitation spectrum (i.e., extreme precipitation and droughts), as these events risk billions of dollars' worth of damage to infrastructure, property, and agriculture (Lott and Ross 2006). Observations have shown an upward trend in the frequency and intensity of extreme precipitation in the Northeast in the last several decades (Kunkel et al. 1999; DeGaetano 2009; Pryor et al. 2009; Kunkel et al. 2013; Howarth et al. 2019). This region has also experienced several major droughts in the twentieth century (Seager et al. 2012), and model projections suggest more short-term droughts in the region due to warming and increased evaporative demand (Hayhoe et al. 2007). The most famous of these droughts, which persisted over the 1962–1967 period (Namias 1966), had major implications for agricultural and water management practices, and still serves as the standard for future water resource planning. Given the many problems that can be associated with an increased frequency in both extreme precipitation and drought conditions, it has become increasingly important to gain a better understanding of the dominant, large-scale atmospheric drivers of precipitation in the Northeast, particularly those that accompany both wet and dry extremes.

✉ Raymond Sukhdeo
rsukhdeo@ucdavis.edu

[1] Department of Land, Air and Water Resources, University of California, One Shields Ave., Davis, CA 95616, USA

Published online: 28 April 2022

Such an understanding provides meaningful insight for stakeholders with interests pertaining to future resource allocations and water management practices in the region.

The Northeast exhibits seasonal cycles for several precipitation characteristics, such as number of precipitating days, precipitation intensity, and precipitation total. These characteristics are discussed in detail in Agel et al. (2015) (hereafter A015), and are summarized in part here. Precipitation intensity tends to have a large annual variation, with a peak in late summer and dual peaks in the spring and fall. These differences in seasonal precipitation characteristics are due in part to the large-scale features and storm tracks associated with daily precipitation in the region. Extratropical storms account for 80–85% of the total precipitation from December to May (Pfahl and Wernli 2012; A015), and the dual peak in precipitation intensity in the spring and fall is related to a weakening of storm-related intensity in May (Pfahl and Wernli 2012; A015). Storm track also plays a key role in this seasonal cycle, and in each season, there are two broad regions with the greatest storm track density: one over the Great Lakes region and another over the Atlantic seaboard (Kocin and Uccellini 2004; Pfahl and Wernli 2012; A015). Track density is highest in the winter and spring, and during the summer and fall, track density in these areas is weaker and shifted northward into the lower portion of eastern Canada. In the summertime, while extratropical cyclone activity is reduced, localized convection, mesoscale convective activity (Feng et al. 2019), and tropical cyclones (Barlow 2011) can be important drivers of precipitation in the region.

Winter precipitation variability in the Northeast has been shown to be more significantly related to the large-scale circulation than in other seasons (Ning et al. 2012a). Previous studies (e.g., Hartley and Keables 1998; Kunkel and Angel 1999; Bradbury et al. 2003) have demonstrated that prominent, large-scale teleconnection patterns, such as the North Atlantic Oscillation (NAO) (Wallace and Gutzler 1981; Barnston and Livezey 1987), the Pacific-North American pattern (PNA) (Wallace and Gutzler 1981; Leathers et al. 1991), and the El Niño-Southern Oscillation (ENSO) (Trenberth 1997), can have important influences on winter precipitation in the Northeast.

However, the relationships identified between winter precipitation and the large-scale teleconnection patterns are generally not robust (Ropelewski and Halpert 1986; Hurrell 1995; Bradbury et al. 2002a, 2002b; Archambault et al. 2008). This lack of statistical significance is hypothesized to be because regional precipitation has inherent, complex variability, and the large-scale modes of climate variability previously discussed also have differing influences across the Northeast. As such, it becomes difficult to identify their influence when examining the domain holistically. Indeed, as determined by Ning and Bradley (2014), linear correlation and composite analysis show that different teleconnection patterns have significant influences on different precipitation patterns, which may explain why previous studies (e.g., Bradbury et al. 2002a; Archambault et al. 2008) did not find robust relationships between total winter precipitation over the whole domain and individual teleconnection patterns.

Our work here seeks to build upon the existing literature related to understanding the large-scale processes that are important in producing conditions favorable to precipitation in the Northeast and assessing the role that large-scale teleconnection patterns may have in those processes. Since this study is primarily concerned with water availability, we focus on monthly accumulated precipitation. Recent studies, such as Agel et al. (2015, 2017, 2019), have addressed the climatology of extreme precipitation in the context of overall precipitation in the region (A015), the large-scale meteorological patterns (LSMPs) associated with that precipitation (Agel et al. 2017; hereafter A017), and the dynamical structures and key ingredients for extreme precipitation, such as moisture and instability, that are linked to those LSMPs (Agel et al. 2019). Their analysis of the climatology of overall precipitation in the Northeast provides an excellent reference point for understanding the seasonal trends in the regional precipitation, and their assessment of the LSMPs and their associated dynamical pathways informs us about which processes are crucial in generating precipitation. However, these analyses have focused on daily, extreme precipitation and have primarily examined the LSMPs associated with that precipitation via the use of clustering methods.

Studies such as these are nonetheless informative to understanding monthly-averaged precipitation, because it is reasonable to assume that most months in which there are more precipitating days will have greater monthly-averaged precipitation amounts (See Fig. S13 in the Online Resource). This also indicates that smaller-scale features, such as localized convection, fronts and extratropical cyclones, which can facilitate heavier precipitation, directly impact monthly precipitation totals.

In this study, we seek to improve the current understanding of environmental conditions favorable to Northeast precipitation and quantify the relative contributions to precipitation amount from upstream drivers. This is achieved by: (I) examining a large number of meteorological fields to understand which are likely associated with precipitation in the region, (II) using a novel linear orthogonal decomposition-based technique, instead of clustering analysis, in order to identify a set of independent large-scale modes that are nearly optimal for prediction of precipitation totals, (III) using multiple linear regression (MLR) to determine the amount of precipitation variability explained by the linear combination of those atmospheric modes, and (IV) determining the influence of several large-scale teleconnection

patterns on precipitation predictability via an analogous MLR framework.

The preceding paragraphs frame our study in the context of previous studies that have examined Northeast precipitation and its associated large-scale drivers, and outline some of contributions our study can make to the existing literature. Specifically, in this present work, we seek to address the following questions:

1) What are the dominant, large-scale atmospheric drivers of precipitation in the northeastern United States, as identified by a linear orthogonal decomposition?
2) How well can multiple linear regression models predict monthly-averaged precipitation in the Northeast?
3) How well does CESM1 LENS compare to reanalysis products with regards to the linear orthogonal modes identified in this study?
4) What is the dependence of the linear orthogonal modes on low-frequency modes of climate variability?

To address these questions, monthly-averaged precipitation in the Northeast is examined in connection with several "basic" atmospheric fields. The linear orthogonal decomposition is used to sub-sample the time series of all predictor fields at all grid points in order to maximize the linear predictability, in the sense of multiple linear regression (MLR). Such an approach is complementary to other methods for decomposing the meteorological fields that drive precipitation, such as k-means clustering analysis; however, one advantage of the linear decomposition is that the linear modes are easily combined within a single linear model, while total precipitation amounts are more difficult to extract from clustered fields. The approach pursued in this study can provide improved insight into the monthly trends and variability associated with Northeast precipitation, as well as inform the level of linear predictability for precipitation in the region via an understanding of key atmospheric features.

The remainder of the paper is organized as follows. Datasets are described in Sect. 2. Methodology is discussed in Sect. 3. Results and analysis of selected linear orthogonal modes are provided in Sect. 4. The paper concludes with a summary and discussion in Sect. 5.

2 Data

Our study makes use of both reanalysis and large ensemble model simulations. For reanalysis, monthly data is drawn from the ERA5 dataset (Hersbach et al. 2020) at $0.25^{\circ} \times 0.25^{\circ}$ latitude/longitude grid spacing, and from the NOAA-CIRES-DOE 20CRv3 dataset (Slivinski et al. 2019), with latitude/longitude grid spacing of $1.0^{\circ} \times 1.0^{\circ}$. ERA5 covers the period from 1950 to 2018 (69 years), while NOAA-CIRES-DOE 20CRv3 covers 1950–2015 (66 years). Both datasets include monthly-averaged data on single levels, as well as monthly-averaged data on several pressure levels (i.e., 500 hPa and 850 hPa). We also employ the CESM1 LENS dataset (Kay et al. 2015), which uses a $0.94240838^{\circ} \times 1.25^{\circ}$ latitude/longitude grid. For CESM1 LENS, we draw on data from a subset of the historical period, from 1979 to 2005, and from the first 35 ensemble members of the dataset (giving 35 ensemble members × 27 years = 945 samples for each month), which provides the advantage of producing many more samples for analysis in each month. We seek to understand the large-scale drivers of Northeast precipitation for each calendar month (i.e., January, February, …, November, December) separately. In the case of CESM1 LENS, this aggregation is also performed over all ensemble members. As such, the monthly data for each calendar month are aggregated together and analyzed separately from the other calendar months. For example, after aggregation, we would have 69, 66, and 945 concatenated Januarys for ERA5, NOAA-CIRES-DOE 20CRv3, and CESM1 LENS, respectively, for use in our analysis. This aggregation is similarly done for the other calendar months. Of important note is the difference in the variety of atmospheric fields between the reanalysis products and the CESM1 LENS dataset. ERA5 and NOAA-CIRES-DOE 20CRv3 offer many more fields for analysis as compared to CESM1 LENS, in part due to the relatively large space requirements when CESM1 LENS was generated. As a result, for the purposes of congruency between our assessment of the reanalysis products and model simulations, we have chosen to examine only atmospheric fields that are common to all three datasets. These fields are referred to now as "basic" atmospheric fields and include: geopotential height at 500-hPa ("Z500"), specific humidity at 850-hPa ("Q850"), zonal wind at 850-hPa ("U850"), meridional wind at 850-hPa ("V850"), sea-level pressure ("PSL"), two-meter temperature ("T2M"), eastward integrated vapor transport ("IVTe"), northward integrated vapor transport ("IVTn"), sea-surface temperature ("SST") and 700-hPa omega ("ω700ε). Total precipitation ("PRECT") is also drawn from the three datasets, and a regional, area-averaged PRECT time series is calculated, using a geographic domain of 40–45 °N and 70–75 °W for the Northeast, for use in our methodology (See Sect. 3). Sensitivity testing for this domain was conducted to verify the robustness of our methodology and results. The results of this testing are given in Sect. 4 of the Online Resource. IVTe and IVTn are not readily given from NOAA-CIRES-DOE 20CRv3 and CESM1 LENS, and so are approximated from the Q850, U850, and V850 fields.

3 Methodology

3.1 Linear orthogonal decomposition (LOD)

Here we provide an overview and summary of the LOD method. The mathematical details of the method can be found in Appendix A.

As stated in Sect. 1, the goal of the LOD method is to sub-sample the time series of all predictor fields at all grid points in order to maximize linear predictability in the context of MLR. To begin, for a particular calendar month of aggregated data (ex. January), a reference state (in our analysis this is the climatological monthly mean) is first subtracted away from the precipitation time series and the 2D atmospheric fields, to generate anomalies of those data. If an obvious and undesired trend is present in the time series, it can be removed by using a time-dependent linear field instead. The iterative procedure begins with the following: (1) the precipitation anomaly time series is correlated with each 2D anomaly atmospheric field, at each grid point, (2) the 2D anomaly atmospheric field and grid point of absolute maximum correlation is then identified, and used to create a linear model of the original precipitation anomaly time series, (3) this linear model precipitation (in this case, LOD Mode 1) is then subtracted away from the original precipitation anomaly time series. This will give us the remainder of the original precipitation anomaly time series that has yet to be predicted. Steps 1–3 are then repeated iteratively, with the original precipitation anomaly time series replaced with the unpredicted remainder. In Steps 2 and 3, we impose orthogonality constraints so that the time series of each identified linear mode is linearly independent from all others. In reality, the processes associated with each linear mode are not independent from each other. However, in our study, applying this artificial orthogonality criteria allows us to better isolate the individual contribution of particular large-scale drivers to the regional precipitation.

To our knowledge, such an approach has not been employed for assessing the linear predictability of precipitation in a region from relevant atmospheric fields. Instead, most studies employ principal component analysis (PCA) as a means of determining key modes. The linear orthogonal decomposition approach is somewhat analogous, with each extracted linear mode being orthogonal to the others. However, it brings an added nuance in that modes are tied directly to the variable being predicted, and allows us to iteratively extract the features across all 2D fields that are most strongly correlated with the monthly, region-averaged precipitation. The iterative method may be terminated at any point—here, we terminate our CESM1 LENS analysis at iteration N when the maximum absolute correlation at iteration $N+1$ is < 0.10. For reanalysis products, global correlations tend to be higher, and so the process is terminated with a higher absolute correlation threshold of 0.20. Note that, while the iterative strategy described above does not guarantee the modes selected are the optimal set of modes to maximize predictability, they are certainly a close approximation to this optimum. Indeed, we expect finding such an optimal combination of field vectors would require an exhaustive search of all possible sets, which would make the problem computationally intractable.

Through the employ of multiple climate data products, we can ascertain whether or not a particular mode is likely to have physical meaning. Namely, we posit that if a particular linear orthogonal mode from our analysis appears consistent across such a set of high-quality products, we can say that the patterns exposed by the linear orthogonal decomposition are robust. On the other hand, if a particular mode is inconsistent across products, it is likely to be a manifestation of structural uncertainty or weather variability (i.e., the component of the precipitation that is inherently unpredictable).

3.2 Multiple linear regression (MLR) using linear modes to predict PRECT

For the analysis of each particular calendar month of aggregated data, a multiple linear regression (MLR) model is generated, using the time series of the linear modes as predictors and the original precipitation anomaly time series as predictand. Indeed, it can be shown that the root mean square error of this MLR model is related to the "unexplained precipitation" at each step of the process via the expression $RMSE^n = \sqrt{\frac{1}{N-n-1}(\hat{\mathbf{p}}^n \cdot \hat{\mathbf{p}}^n)}$, where $\hat{\mathbf{p}}^n$ is the unpredicted precipitation at iteration step n. The MLR model enables us to validate our above scheme by providing additional confirmation of the explained variance afforded by the "basic" atmospheric fields that was calculated in Part A. Additionally, MLR also provides the 95% confidence range for the coefficients associated with each linear mode time series. These coefficients, in conjunction with the linear mode time series, allow us to determine how well our extracted modes combine linearly in the predictability of the monthly, region-averaged precipitation, and allow us to test whether or not these time series are in fact statistically significant predictors.

3.3 Multiple linear regression (MLR) using climate indices to predict linear modes

In a similar process to that discussed in Sect. 3.2, for the analysis of a particular month, an MLR model is built with several climate indices being used as predictors, and each

linear mode time series being the predictand. The climate indices under consideration are: AMO, DMI, IPO, NAO, Nino1 + 2, Nino3, Nino3.4, Nino4, PDO, and PNA. All climate indices are initially included in the linear model regression for each linear mode time series. After the initial regression output, a recursive elimination approach is used to select only the climate indices that are significant at the 99% confidence level (p-value < 0.01). This feature selection based on recursive elimination is done for each linear mode time series for each calendar month of aggregated data.

4 Results

4.1 CESM1 LENS: discussion of LOD modes

As discussed previously, the linear orthogonal decomposition method allows us to determine the large-scale modes and processes that are most closely associated with monthly-averaged precipitation in the Northeast. These processes that are important in the generation of precipitation will often be embedded within large-scale meteorological patterns (LSMPs; Grotjahn et al. 2016), and these LSMPs in turn describe the atmospheric circulation that can be attributed to specific phenomena, in this case, precipitation in the Northeast. However, often the LSMPs themselves may not be the primary drivers and/or direct influencers of the generation of precipitation. Rather, these LSMPs provide a favorable environment for other processes to generate precipitation (A017). Our linear orthogonal modes are analogous to these LSMPs, and while our LSMP-like linear orthogonal modes lack a local consistency metric, contour shading in our plots indicates statistically significant (p-value < 0.05) regions in our modes based on Student's t-testing. As such, the modes generated by the linear orthogonal decomposition are able to capture several important large-scale drivers of precipitation in the region. Given the similarity in results between the reanalysis products and CESM1 LENS, the larger sample size afforded by CESM1 LENS that allows us to avoid overestimation of the predictability of the LOD modes, and because this work is focused on understanding these large-scale drivers of precipitation as a means for informing future resource allocations and water management practices in the region, we choose to highlight the CESM1 LENS results here in Sect. 4 Results for the reanalysis products are similar to these, and are shown in the Online Resource.

i. Winter and Early Spring

Our analysis of January for CESM1 LENS data (Figs. 1, 2, and 3) is used here to capture the large-scale conditions

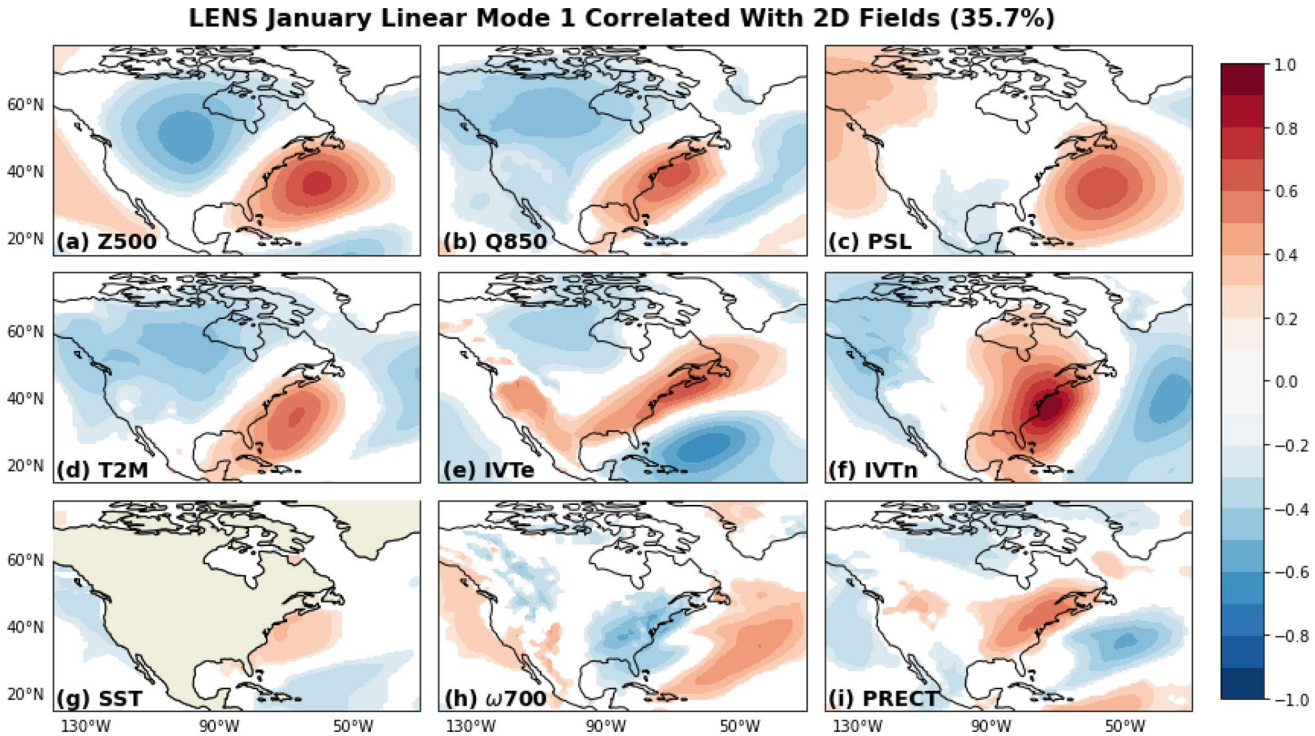

Fig. 1 CESM1 LENS January Linear Mode 1 for overall precipitation correlated with anomaly fields for **a** 500-hPa geopotential height, **b** 850-hPa specific humidity, **c** sea-level pressure, **d** two-meter temperature, **e** eastward integrated vapor transport, **f** northward integrated vapor transport, **g** sea-surface temperature, **h** 700-hPa omega, and **i** total precipitation. Contour shading indicates significance at the 95% confidence level. Percentage is fraction of precipitation variance explained by the first linear mode in January

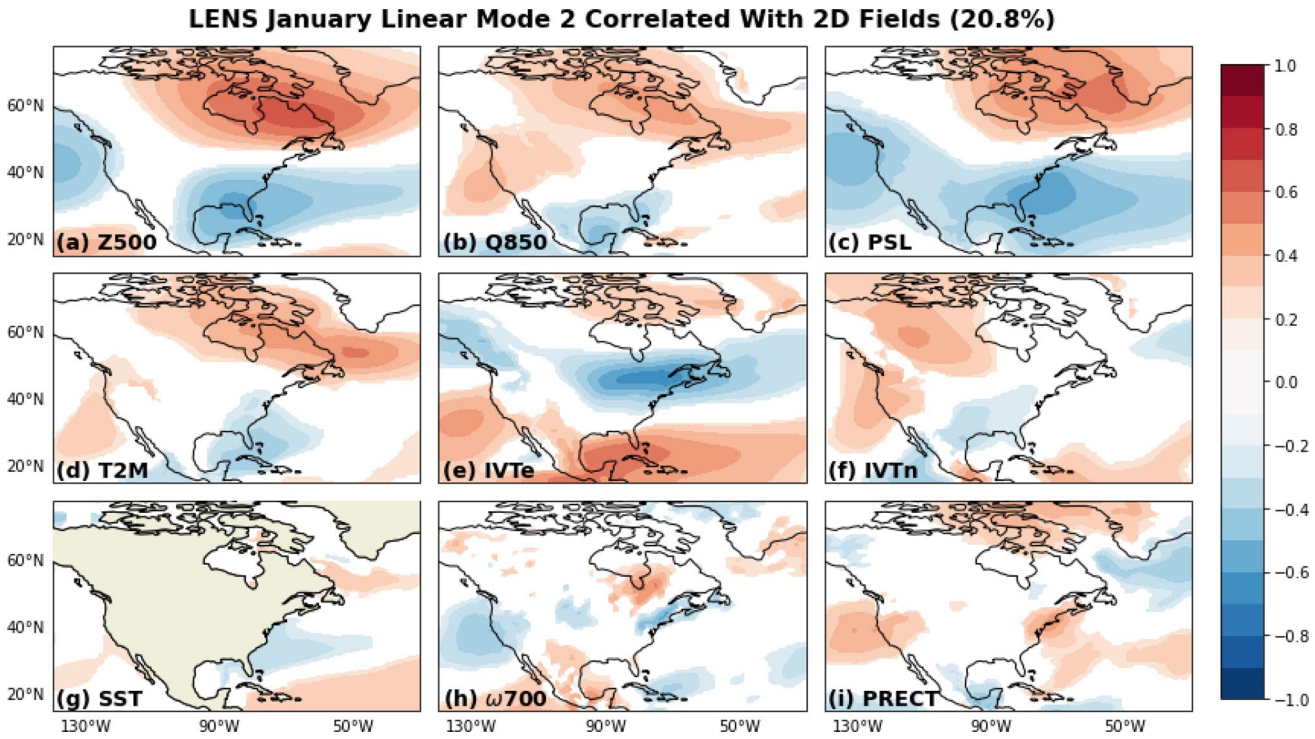

Fig. 2 Same as Fig. 1, but for Linear Mode 2

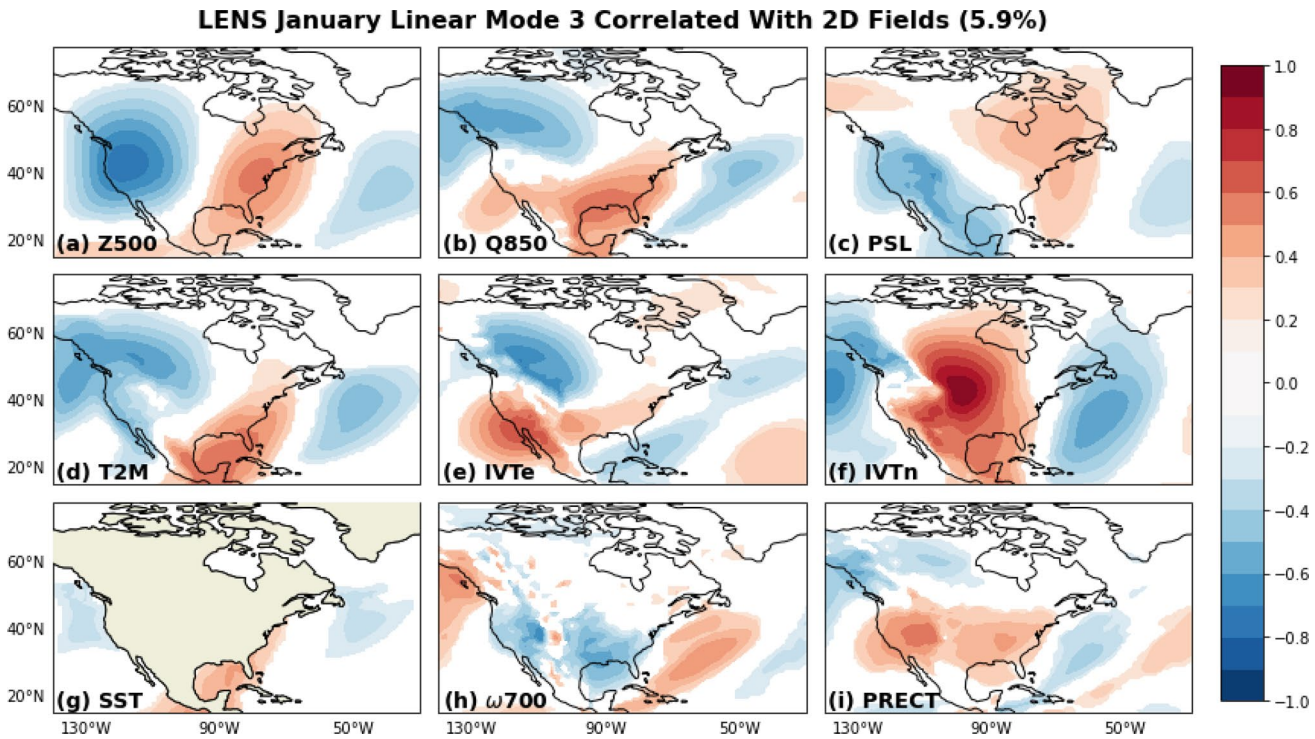

Fig. 3 Same as Fig. 1, but for Linear Mode 3

that are conducive to wintertime precipitation in the Northeast. Results for December, February, and March are similar, and can be found in the Online Resource. The results of our linear orthogonal decomposition method for

Table 1 Linear orthogonal decomposition results for CESM1 LENS data

LOD Results for CESM1 LENS					
	$d\hat{p}^1$	$d\hat{p}^2$	$d\hat{p}^3$	$d\hat{p}^4$	R-Squared Value
January	IVTn(72 °W,37 °N) 0.5979 [0.56,0.64] 35.7%	U850(71 °W,45 °N) 0.4556 [0.42,0.49] 20.8%	IVTn(98 °W,42 °N) 0.2436 [0.21,0.28] 5.9%	Z500(92 °W,18 °N) 0.1208 [0.08,0.16] 1.5%	0.639
February	IVTn(71 °W,37 °N) 0.6317 [0.59,0.67] 39.9%	U850(70 °W,46 °N) 0.3884 [0.35,0.43] 15.1%	Z500(110 °W,49 °N) 0.2784 [0.24,0.32] 7.8%	PSL(90 °W,47 °N) 0.1506 [0.11,0.19] 2.3%	0.650
March	IVTn(70 °W,38 °N) 0.6085 [0.57,0.65] 37.0%	U850(71 °W,46 °N) 0.4216 [0.38,0.46] 17.8%	Z500(110 °W,48 °N) 0.2664 [0.23,0.31] 7.1%	IVTe(75 °W,37 °N) 0.1164 [0.08,0.16] 1.4%	0.632
April	IVTn(70 °W,39 °N) 0.5752 [0.54,0.62] 33.1%	U850(71 °W,46 °N) 0.4504 [0.41,0.49] 20.3%	Z500(105 °W,50 °N) 0.2136 [0.17,0.25] 4.6%	IVTe(76 °W,38 °N) 0.1486 [0.11,0.19] 2.2%	0.601
May	IVTn(70 °W,40 °N) 0.5310 [0.49,0.57] 28.2%	U850(71 °W,46 °N) 0.5108 [0.47,0.55] 26.1%	IVTe(77 °W,38 °N) 0.2398 [0.20,0.28] 5.7%	T2M(52 °W,41 °N) 0.1821 [0.14,0.22] 3.3%	0.634
June	IVTn(68 °W,40 °N) 0.6450 [0.61,0.68] 41.6%	IVTe(70 °W,46 °N) 0.4190 [0.38,0.46] 17.8%	IVTe(75 °W,39 °N) 0.2225 [0.19,0.26] 4.9%	IVTn(89 °E,20 °N) 0.1038 [0.07,0.14] 1.1%	0.652
July	IVTn(70 °W,40 °N) 0.6964 [0.66,0.74] 49.5%	IVTe(71 °W,46 °N) 0.3632 [0.33,0.40] 13.2%	Z500(92 °W,36 °N) 0.1390 [0.10,0.18] 1.9%		0.636
August	IVTn(68 °W,40 °N) 0.5355 [0.49,0.58] 28.7%	IVTe(72 °W,46 °N) 0.4661 [0.42,0.51] 21.7%	IVTe(76 °W,39 °N) 0.2466 [0.20,0.29] 6.1%		0.565
September	IVTn(68 °W,39 °N 0.4827 [0.44,0.53] 28.8%	U850(71 °W,46 °N) 0.4411 [0.40,0.48] 19.1%	IVTe(80 °W,39 °N) 0.2888 [0.25,0.33] 4.8%	T2M(56 °W,42 °N) 0.1779 [0.14,0.22] 2.4%	0.543
October	IVTn(70 °W,39 °N) 0.5918 [0.55,0.63] 35.0%	U850(70 °W,47 °N) 0.4646 [0.43,0.50] 21.6%	Z500(80 °W,27 °N) 0.1920 [0.15,0.23] 3.7%	Z500(57 °W,24 °N) 0.1168 [0.08,0.16] 1.4%	0.617
November	IVTn(71 °W,36 °N) 0.5719 [0.53,0.61] 32.7%	U850(70 °W,46 °N) 0.4549 [0.41,0.50] 20.7%	V850(62 °W,26 °N) 0.1958 [0.16,0.24] 3.8%	Q850(73 °W,40 °N) 0.1531 [0.11,0.19] 2.3%	0.596
December	IVTn(72 °W,37 °N 0.5550 [0.52,0.60] 30.8%	U850(70 °W,45 °N) 0.4783 [0.44,0.52] 22.8%	Z500(111 °W,42 °N) 0.2516 [0.21,0.29] 6.3%	Q850(106 °W,12 °N) 0.1074 [0.07,0.15] 1.2%	0.612

First four columns represent the first four linear orthogonal modes. Rows for the first four columns indicate, respectively, (i) selected atmospheric field and geographic grid point, (ii) linear regression coefficient value, (iii) 95% confidence range for the coefficient value, (iv) fraction of explained variance of the precipitation data. Last column indicates R-squared value produced from multiple linear regression using the linear orthogonal modes to predict precipitation

CESM1 LENS are given in Table 1, and are discussed in part here and in the other subsections of Sect. 4.1.

After applying the methodology discussed in Sect. 3, it emerges that the first linear mode is associated with northward integrated vapor transport (IVTn) (Fig. 1f), which is maximal along the Atlantic seaboard near coastal Virginia, the second linear mode is associated with weakened westward integrated vapor transport (IVTe) over the Northeast (Fig. 2e), and the third linear mode is associated with northward integrated vapor transport (IVTn) originating from the Gulf of Mexico region (Fig. 3f). Recall that in each iteration step, the selected time series of data (based on field and geographic location) is determined by absolute maximum correlation with the region-averaged precipitation anomaly time series. Projecting these selected time series (our linear modes) onto the 2D meteorological fields (i.e., calculating their correlations with each point in the field, as in Step 1 of Sect. 3.1) provides further insight into the physical mechanisms that can lead to precipitation. Looking at Fig. 1a, the east–west dipole of Z500 over the central United States and the North Atlantic provides a favorable mechanism for integrated vapor transport from the west and south. Increased moisture transport into the region, coupled with enhanced vertical motion (as seen in the ω 700 field, Fig. 1h), creates a conducive environment for precipitation to occur. This finding is reasonable, as ocean-enhanced transport from the south or southeast can provide a rich source of water vapor into the Northeast. A017 demonstrate that IVTn is stronger for extreme precipitation as compared to interquartile (25–75th percentile) precipitation. This suggests that precipitation totals in January (and winter and early spring as a whole) are strongly connected to the amount of southerly moisture transport into the region. Correlation plots based on the second linear mode for PSL, IVTe, and ω 700 (Fig. 2) suggest that low surface pressure to the south of the Northeast drives a weakening of the westerly flow and leads to moisture convergence. This is attributed, at least in part, to enhanced extratropical cyclone activity near this region. Enhanced surface lows, originating from the midwestern or southeastern United States, can move northward into the region, and the associated cyclonic rotation of these systems can stimulate a weakening of the westerly flow that facilitates increased westward vapor transport and moisture convergence, while frontal activity can provide an additional lifting mechanism for precipitation to occur. Examining the Z500 and IVTn correlation plots for the third linear mode (Fig. 3) suggests that the location of high- and low-pressure systems can facilitate integrated vapor transport from the Gulf of Mexico into the central and east-central United States. In a similar (albeit weaker) manner to the first linear mode, increased southwesterly moisture convergence into the region can help to generate precipitation.

ii. Late Spring

Results for April (as shown in the Online Resource) demonstrates good similarity with the results for the winter period and March, with our respective modes of Atlantic seaboard vapor transport (first linear mode), westward vapor transport associated with extratropical cyclone activity (second linear mode), and vapor transport from the Gulf of Mexico (third linear mode). However, as we move later into the season, there is an eastward shift of the east–west dipole of Z500 over the central United States and the North Atlantic. This configuration continues to provide a favorable mechanism for integrated vapor transport from the west and south, but indicates a transition period in the large-scale pattern over the conterminous United States. In terms of generating precipitation, this transition leads to the configuration for May (Figs. 4, 5, and 6), in which there is more noticeable shallow troughing in the Ohio River Valley. This pattern accompanies vapor transport from the south and southwest (Fig. 4), but reduced vapor transport from the Gulf of Mexico (Fig. 6). As was the case in preceding months, the location of the trough acts to modulate processes such as vapor transport, low pressure development, and whether available moisture from the Gulf of Mexico can reach the Northeast. In the case of May, it appears that Gulf of Mexico vapor transport is not a relatively important process because it is does not make its way as far northward as in the winter months.

iii. Summer and Early Fall

Our results for July (Figs. 7 and 8) exhibit several notable differences from those discussed already for the winter and late spring months (similar results for June, August, and September; shown in Online Resource). While vapor transport from the Atlantic seaboard (first linear mode) is still a prominent process for the generation of precipitation, both the influence of extratropical cyclone activity (see subsection *v* in Sect. 4.1) and vapor transport from the Gulf of Mexico are diminished during this seasonal period, and seem to be replaced in part by other processes. Regarding the second linear mode, inspection of the Z500 correlation field (Fig. 8a) indicates a pressure ridge in southeastern Canada. The anticyclonic flow associated with a ridge pattern in the north could facilitate moisture advection into the region. The warmer summer environment also leads to a substantial increase in moisture capacity, and warm, moist unstable air suggests an association with convective activity (Fig. 8h), such as localized thunderstorms. Additionally, while localized convection can be important for generating precipitation, shortwave vorticity maxima propagating through longwave ridges can create a mechanism for convection to occur within these ridge patterns themselves (Milrad

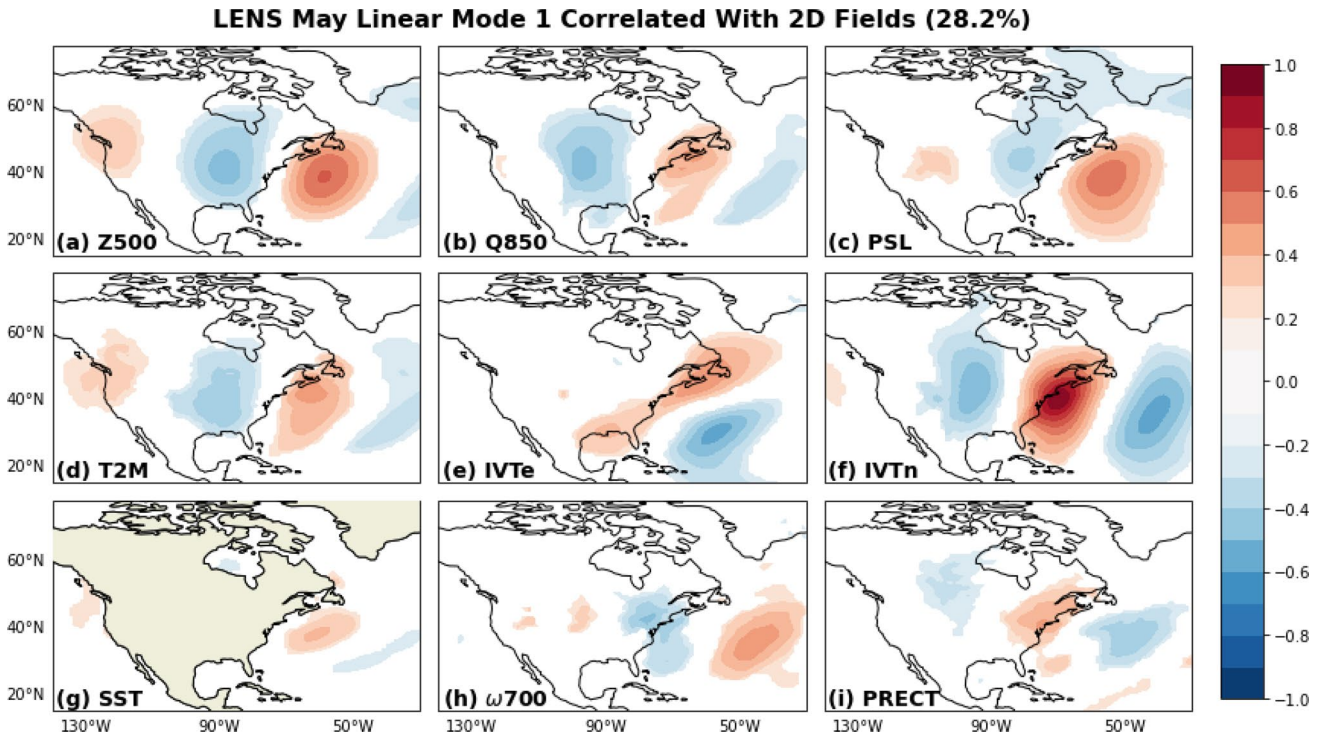

Fig. 4 CESM1 LENS May Linear Mode 1 for overall precipitation correlated with anomaly fields for **a** 500-hPa geopotential height, **b** 850-hPa specific humidity, **c** sea-level pressure, **d** two-meter temperature, **e** eastward integrated vapor transport, **f** northward integrated vapor transport, **g** sea-surface temperature, **h** 700-hPa omega, and **i** total precipitation. Contour shading indicates significance at the 95% confidence level. Percentage is fraction of precipitation variance explained by the first linear mode in May

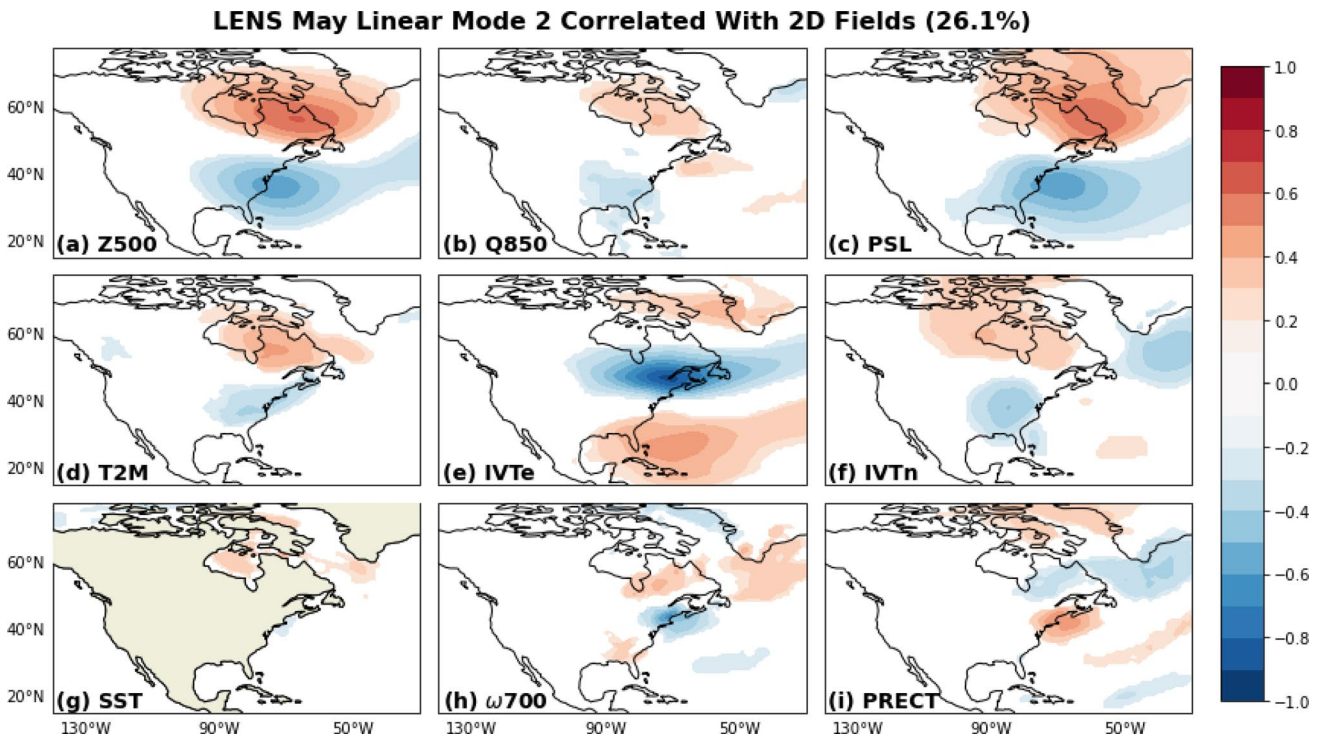

Fig. 5 Same as Fig. 4, but for Linear Mode 2

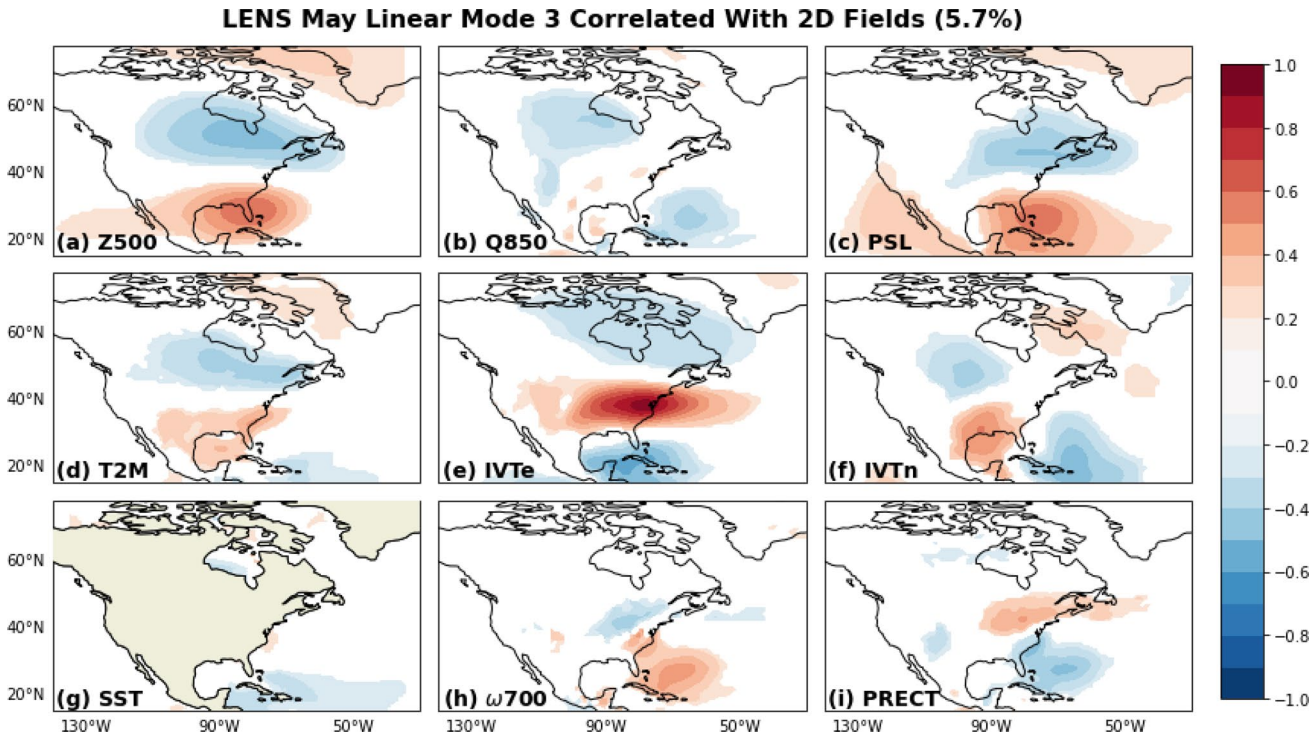

Fig. 6 Same as Fig. 4, but for Linear Mode 3

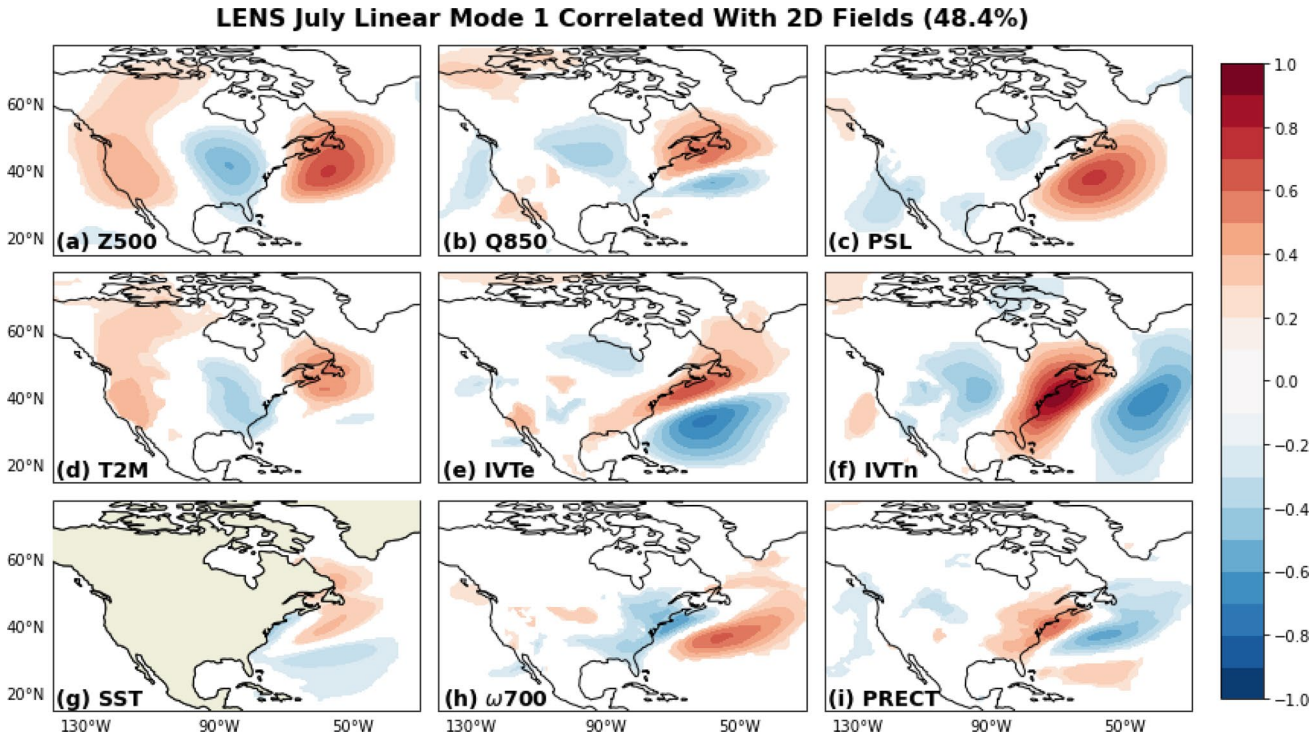

Fig. 7 CESM1 LENS July Linear Mode 1 for overall precipitation correlated with anomaly fields for **a** 500-hPa geopotential height, **b** 850-hPa specific humidity, **c** sea-level pressure, **d** two-meter temperature, **e** eastward integrated vapor transport, **f** northward integrated vapor transport, **g** sea-surface temperature, **h** 700-hPa omega, and **i** total precipitation. Contour shading indicates significance at the 95% confidence level. Percentage is fraction of precipitation variance explained by the first linear mode in July

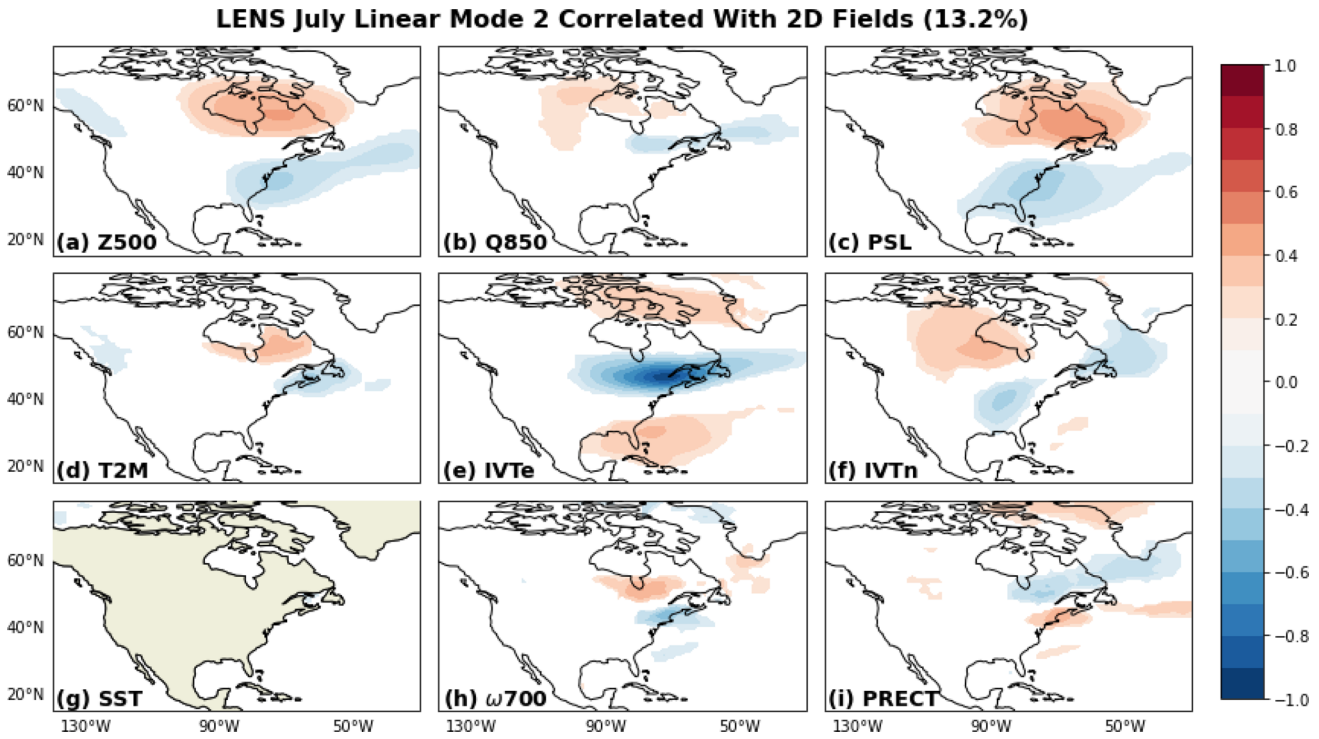

Fig. 8 Same as Fig. 7, but for Linear Mode 2

et al. 2014). Given that we are working with monthly aggregated data, it would be difficult to fully resolve the influence of localized convective activity, and the mechanisms proposed in the previous few sentences are only hypotheses about what large-scale drivers may be associated with the second linear mode. It may just be the case that extratropical cyclone activity is still the relevant driver of precipitation in the mode, but is reduced in importance, relative to the winter and spring periods. The third linear mode for July (as shown in the Online Resource) is difficult to physically interpret and provides only a small fraction of explained precipitation variance (1.9%), and as such is not discussed in great detail here. We hypothesize that this third linear mode may correspond to enhancement of the environment conducive to mesoscale convective system (MCS) activity. It must be noted, however, that MCSs are more of a prominent feature in the central U.S., and occur with much less frequency in the Northeast, even though these systems account for roughly 10–20% of total precipitation in the region during the summer period (Feng et al. 2019). So, while MCSs can be heavy-precipitation events when they do occur, they are infrequent and would not be resolved in our monthly aggregated data, and we would refer readers to Feng et al. (2019) for further discussion of these systems. If this third mode is in fact related to MCSs, this may explain why our third linear mode in July explains only 1.9% of the precipitation variance.

iv. Fall to Late Fall

Our analysis for October and November (as shown in the Online Resource) demonstrates a return to a large-scale configuration with drivers that are similar to those found during the winter and early spring, namely Atlantic seaboard vapor transport (first linear mode), westward vapor transport associated with extratropical cyclone activity (second linear mode), and vapor transport from the Gulf of Mexico (third linear mode). More specifically, the east–west Z500 dipole found during October and November is similar to the configuration found during April, in which the dipole is now phase-shifted westward. This is likely due to the late fall being a transition period before the winter, in which this Z500 dipole is shifted further westward and stays in place during much of the winter and into the early spring.

*v. Extratropical Cyclone (*ETC*) Activity*

We have proposed that the second linear mode during the fall, winter, and spring is most closely associated with extratropical cyclone (ETC) activity to the south of our region of interest. This assertion, however, has not yet been demonstrated. It can be difficult to directly link synoptic-scale (~ 1–3 day) systems, such as ETCs, with monthly precipitation totals, as averaging can potentially dilute the contribution of individual strong systems. However, in our study here, we are more interested in how our LOD modes modulate the large-scale environment in which these ETCs

are present and not necessarily with their actual precipitation amounts. To examine this relationship, we track the total number of time steps in 6-hourly daily data when ETC activity is present for a specific geographic domain relevant to the Northeast. In this way, we are able to link the synoptic-scale influence of ETCs to their relevance in influencing monthly precipitation totals. For tracking ETCs, we use the TempestExtremes software package (Ullrich and Zarzycki 2017), which allows for the tracking of pointwise features within climate datasets. Six-hourly data for surface pressure (PS) and geopotential at 300 hPa and 500 hPa (Z300 and Z500, respectively) are taken from CESM1 LENS for the period 1990–2005 in each of the first 35 ensemble members and used to drive TempestExtremes, which then outputs the relevant ETC tracking information, such as time, the latitude/longitude of the cyclone centers and local sea-level pressure. Output from TempestExtremes is then filtered for geographic relevance to the Northeast, with a latitude/longitude domain of 25–55 °N and 60–100 °W, respectively (Fig. 9, left). This choice of domain accounts for both the high-density North Atlantic storm track as well as the extratropical storms traveling across the Great Lakes region (Hoskins and Hodges 2002). After the filtering process, the total number of time steps with ETC activity are tabulated to generate separate "ETC Activity" time series for each calendar month of aggregated data (January–December). These time series are generated with the same aggregation method as discussed in Sects. 2 and 3. The time series can then be correlated with the appropriate subsection of our linear mode time series (Table 2).

Figure 9 (right) shows the monthly distribution of ETC activity. There are far more instances of ETC activity during the late fall and winter periods, and a decline during the late spring and summer. These results agree well with those of Hawcroft et al. (2012) and A015, which find that extratropical storms account for nearly 80–85% of total precipitation in late fall / winter periods, and indicates that extratropical cyclones are not a prominent feature for Northeast precipitation during the late spring and summer periods. We can further validate the role of ETC activity in certain calendar months by examining the correlation values between our linear modes and the ETC Activity time series. From Table 2, we see that ETC activity is not well correlated with either the first or third linear modes. For the second linear mode, the correlation values are much higher and scale well with the total number of time steps with ETC activity in each month. This provides reasonable confidence that the second linear mode is related to ETCs traveling northward into the Northeast. As discussed previously, extratropical storms are an important driver of precipitation in the Northeast, and can be a primary driver of extreme precipitation conditions, as a large percentage of extreme precipitation along the

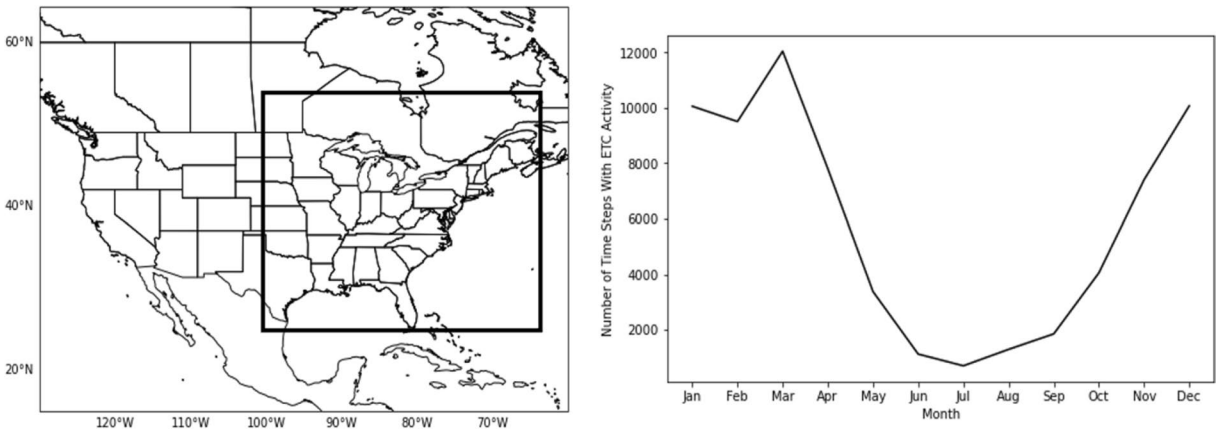

Fig. 9 Geographic domain for ETC tracking (left) and total monthly instances of ETC activity (right) for CESM1 LENS. ETC tracking domain accounts for storms that move northward along the Atlantic seaboard and those from the Great Lakes region

Table 2 Correlation coefficient values between first three linear orthogonal modes and ETC Activity time series for CESM1 LENS

Correlation Values Between Linear Modes and ETC Activity Time Series												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Linear Mode 1	**0.163**	0.027	0.072	0.059	0.022	0.060	0.046	0.097	0.048	0.025	0.049	**0.244**
Linear Mode 2	**0.411**	**0.485**	**0.435**	**0.304**	**0.232**	**0.279**	0.113	**0.161**	**0.215**	**0.193**	**0.436**	**0.390**
Linear Mode 3	0.079	0.003	0.022	0.044	0.093	0.047	0.017	0.025	0.137	0.115	0.011	0.002

Bold font indicates correlation values that are significant at the 99% confidence level

Atlantic seaboard occurs near extratropical cyclone centers (Kunkel et al. 2012; Pfahl and Wernli 2012). Additionally, Pfahl and Sprenger (2016) showed that cyclone intensity and moisture availability scale well to precipitation amount, meaning that stronger cyclones and greater moisture content likely lead to increased precipitation totals. Looking back at our plots for the second linear mode, particularly during the late fall / winter periods, we see that this relationship is validated. It may seem surprising that the ETC Activity time series are generally not well correlated with the first linear mode, given the overall importance of ETCs in contributing to monthly precipitation totals. However, it is worth keeping in mind that LOD modes represent orthogonal contributions to monthly-averaged precipitation and as such can either reinforce or counteract one another. Effectively what we are seeing is that precipitation is a nonlinear combination of both moisture availability and dynamic uplift. Our first LOD mode appears to be primarily related to moisture availability in the Northeast, which is modulated primarily by southerly vapor transport. Our second LOD mode appears to be more strongly related to dynamic uplift, which is primarily influenced by frontal activity. Strong positive values in both modes would be indicative of both a moistening of the Northeast and strong dynamic uplift, likely contributing to extremely wet conditions in such a month. With that said, opposite dynamic patterns are apparent in the PSL field, and seem to limit the co-occurrence of these conditions: the first mode suggests that optimal moisture transport occurs when PSL is enhanced off the Atlantic seaboard, while the second mode suggests that ETCs primarily occur in conjunction with negative PSL anomaly in this region.

Additionally, when testing alternative ETC tracking domains (see Sect. 6 of the Online Resource), we see that the strength of the correlation between the first linear mode and the ETC Activity time series is sensitive to the choice of domain (i.e., tracking ETCs in a particular domain can lead to more or less Atlantic seaboard vapor transport). While this sensitivity does exist regarding the first linear mode, the correlations with the second linear mode are insensitive to domain choice, such that the strongest correlations are between the ETC Activity time series and the second linear mode, regardless of domain choice. This further confirms our finding that the second linear mode is most strongly related to ETCs.

vi. A Link Between the Daily and Monthly Large-Scale Circulations

These three mechanisms discussed thus far for the generation of precipitation, localized vapor transport into the region from the Atlantic seaboard, reduced westerlies associated in part with extratropical cyclone activity, and remote vapor transport from the Gulf of Mexico (for the winter and fall periods), suggest a prominent role for certain LSMPs in creating favorable conditions by which these mechanisms can occur. Several studies (Bradbury et al. 2002b; A017; Suriano et al. 2018) have examined how the large-scale circulation can influence the amount of precipitation received in the Northeast. For this region, a dominant troughing pattern in the eastern United States is often responsible for overall precipitation, with trough strength having a seasonal dependence, and this pattern is accompanied by increased southerly ocean-enhanced vapor transport that extends along the Atlantic seaboard, increased easterly zonal flow, and strong surface lows into the northern portion of the region (A017; Agel et al. 2019). Additionally, the relative zonal position of the trough can act to modulate precipitation amounts, such that a more western (eastern) trough position is associated with greater (lower) monthly-averaged precipitation (Bradbury et al. 2002b). This last point is reasonable, as the most active midlatitude storm tracks occur directly downstream of troughs (Lau 1988; Cai and Van den Dool 1991), while areas just upstream of troughs are accompanied by surface anticyclones and generally have drier conditions (Harman 1991). Based on our above discussion, these processes seem to be well-captured in part by our extracted linear orthogonal modes. Previous studies conducted at different timescales (daily for A017 and monthly for Bradbury et al. (2002b) and Suriano et al. (2018)) indicate a similar role for the large-scale circulation in influencing precipitation totals. The clustering methods used by A017 on daily data (to study extreme precipitation) demonstrate in part that wintertime strength and positioning of troughs can modulate precipitation amounts. A similar conclusion is reached by Bradbury et al. (2002b) when correlating monthly trough position and strength indices to monthly winter precipitation amounts, and Suriano et al. (2018) when examining monthly mid-tropospheric wave train patterns over North America. This likely suggests that monthly analyses such as ours, Bradbury et al. (2002b), and Suriano et al. (2018) are in fact representing more frequent occurrences of particular daily large-scale configurations. If a particular daily, large-scale pattern occurs more frequently in a month, it will have a higher representation in the monthly-averaged large-scale pattern for that month. In the context of studies such as A017 that examine the daily LSMPs associated with extreme precipitation, this indicates that more frequent occurrences of high precipitation-causing LSMPs in a given month will likely lead to increased precipitation totals for that month. There is thus a direct link between daily large-scale patterns and monthly large-scale patterns and the variability associated with monthly precipitation totals. As such, our analysis suggests that, with regards to eastern United States trough strength and positioning, higher precipitation totals are more likely in a large-scale environment with an enhanced ridge in the North Atlantic, well-defined storm tracks with more frequent storms, and increased southerly moisture transport.

Conversely, we would expect drier conditions to be accompanied by a weaker ridge in the North Atlantic, weakly-defined storm tracks with less frequent storms, and reduced southerly moisture transport.

4.2 CESM1 LENS: multiple linear regression using linear modes to predict PRECT

As discussed in the methodology in Sect. 3, the linear modes for each month are used as predictors for the corresponding $\hat{\mathbf{p}}^0$ for each month in a multiple linear regression (MLR) model. The R-squared value generated for each model, which accounts for the variance of the precipitation that can be explained by our predictors, provides us with an estimation of how well our linear mode time series are able to predict $\hat{\mathbf{p}}^0$ for each month. Following this approach, in all months, all of our linear mode time series are significant at the 99% confidence level (p-value < 0.01). R-squared values are listed in Table 1 and example scatterplots comparing observations to MLR predictions for January, May, July, and October are shown in Fig. 10.

Predictability is somewhat similar across all calendar months, as the R-squared values range from 0.54 to around 0.65. Predictability is generally highest during the winter (December, January, February, and March) and early summer (June and July) seasons, and displays a noticeable decline for CESM1 LENS in April, May, August, and September. This raises the question as to what causes this decline in predictability? The decline can be explained by examining the seasonal cycle of precipitating days and precipitation intensity during these months. In April and May, there is a near peak in number of precipitating days, but a noticeable decline in precipitating intensity (A015). Mechanisms and processes such as blocking episodes, localized convection, and backdoor cold fronts are active features of the late spring period, and can contribute to consecutive days of light precipitation (i.e., more precipitating days with less intensity) (Bosart et al. 1973). Additionally, the April–May time period usually sees a shifting of the jet stream to a more northerly position, which can reduce the number of extratropical storms that move northward along the Atlantic seaboard (A015). This, combined with delayed warming due to cold ocean temperatures during the spring season (Trenberth

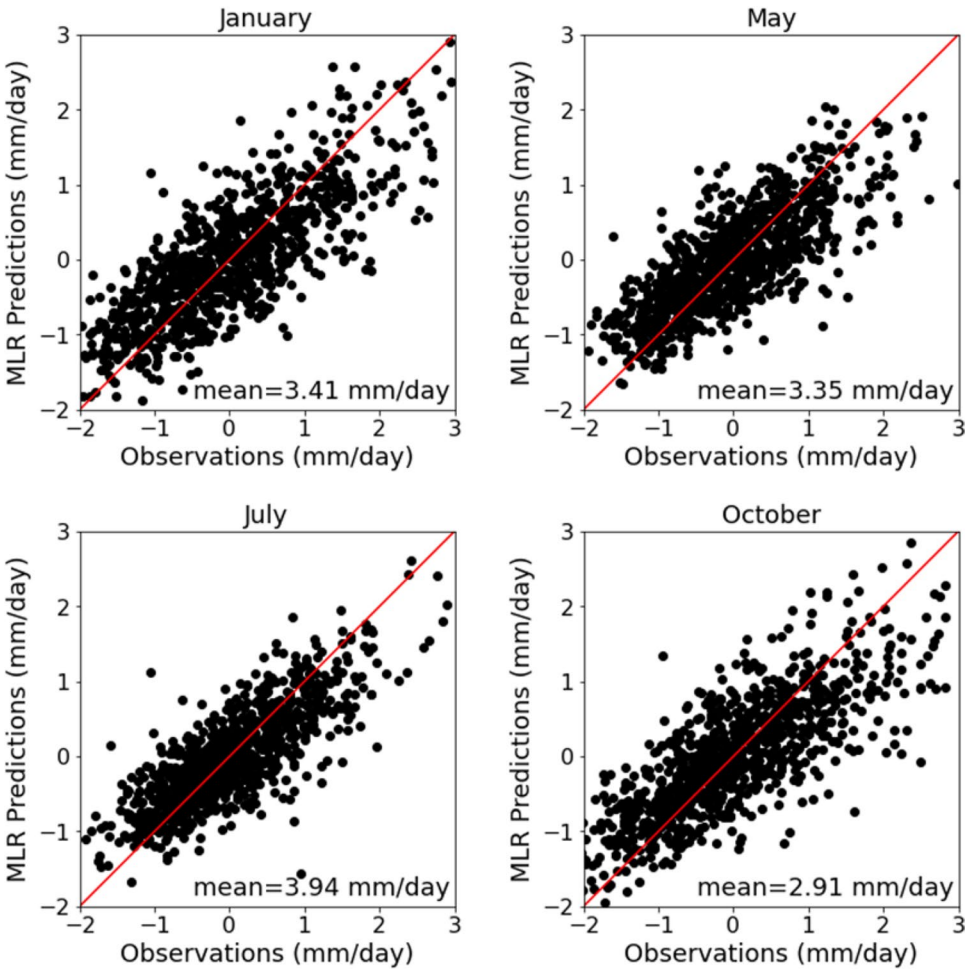

Fig. 10 Scatterplots for CESM1 LENS PRECT anomalies observations vs. MLR predictions for January, May, July and October. Red line in each frame represents the 1:1 line, and value in bottom right corner of each frame indicates the mean precipitation for that month

et al. 2003), can limit precipitation intensity. In August and September, we see an opposite seasonal cycle behavior from that of April and May, in that there is a decline in number of precipitating days, but a near peak in precipitation intensity (A015). This indicates that some precipitation in August–September can tend to come from short-lived, intense systems such as mesoscale convective systems (Feng et al. 2019) or tropical cyclones (Barlow 2011).

Given the inherent interannual variability associated with the frequency and intensity of localized convection, fronts, and extratropical and tropical cyclones, and the infrequent and often spotty nature of MCSs, and the fact that the influence of these systems is not captured well in monthly data, this may explain in part the decline in linear predictability in the spring and summer periods. As discussed in A015, 50% of overall precipitation in the Northeast occurs as single-day events, and most extreme precipitation occurs over several multiday events. Additionally, Hawcroft et al. (2012) estimates that roughly 85% of total precipitation in DJF comes from extratropical storms. Similarly, Catto et al. (2012) find that cold fronts and warm fronts account for 28% and 29%, respectively, of midlatitude precipitation, with frontal activity accounting for 90% of extreme precipitation (Catto and Pfahl 2013). Given our analysis in this study, which is predicated on assessing monthly predictability of precipitation, it can be difficult to fully account for the sub-monthly time-scale processes discussed in these previous studies, as monthly averaging of precipitation totals will tend to partially dilute the impact of some high-precipitating events. We acknowledge this limitation, and would refer readers to the aforementioned studies for further discussion.

4.3 Comparison between CESM1 LENS, ERA5, and NOAA-CIRES-DOE 20CRv3

For the methods presented in Sect. 3, it is important to test whether the identified modes are reasonably consistent between the reanalysis products and CESM1 LENS. To validate this, the LOD method was applied to the data for ERA5, NOAA-CIRES-DOE 20CRv3 and CESM1 LENS. The results for each dataset are shown in Table 1 (CESM1 LENS), Table 3 (ERA5), and Table 4 (NOAA-CIRES-DOE 20CRv3). In terms of predictability, CESM1 LENS has a lower predictability range (determined by the R-squared values) of around 0.54–0.65, as compared to the higher predictability range for ERA5 and NOAA-CIRES-DOE 20CRv3 of around 0.58–0.88. This is perhaps not surprising, as the two reanalysis products have a smaller sample size than CESM1 LENS, and thus a greater level of variability can be captured with fewer modes. More importantly, the lower predictability range for CESM1 LENS likely results from the model's internal variability with regards to representing the precipitation drivers. Overall, the maximum predictability afforded by the linear regression models for $\hat{\mathbf{p}}^0$ across all months and datasets (determined by the R-squared values) is limited to a value of 0.88. We hypothesize that this cap on linear predictability may be because the regional precipitation in the Northeast has inherent, complex variability (Ning and Bradley 2014) that cannot be fully captured by linear prediction methods.

Recall that in each iteration step of the LOD method, the selected time series of data (based on field and geographic location) is determined by absolute maximum correlation with the region-averaged precipitation anomaly time series. For each dataset in January, the first two iterations of the methodology select for IVTn along the Atlantic seaboard and U850 in western New York for CESM1 LENS and ERA5, and IVTn for NOAA-CIRES-DOE 20CRv3, respectively. The third iteration selects for U850 in ERA5, IVTe in NOAA-CIRES-DOE 20CRv3, and IVTn in CESM1 LENS. The fourth iteration selects for U850 in ERA5, IVTn in NOAA-CIRES-DOE 20CRv3, and Z500 in CESM1 LENS. Although different fields and locations are selected for the latter modes, this does not necessarily indicate divergence in the meteorological patterns identified by these modes; it is instead largely indicative of strong correspondence between different fields and locations. By projecting these selected time series onto the 2D fields for each dataset, we can better ascertain similarities in the large-scale features that are selected within each dataset. Figure 11 shows the projections onto the respective IVTn, Z500 and PRECT fields for each dataset for January (other fields and months are shown in the Online Resource). Visually, we see that there is good similarity across the three datasets for the correlation fields between the first two selected time series and the 2D fields. The visual similarity is reduced when examining the correlation fields between the third selected time series and the 2D fields. This similarity is quantitatively corroborated by calculating the spatial correlation in the domain of 25–60 °N and 60–100 °W between the three datasets (tabulated values for the first three linear orthogonal modes are shown in Sect. 2 of the Online Resource). This choice of domain allows us to account for how well the datasets compare with regards to capturing the large-scale features that influence precipitation in the Northeast. Generally, spatial correlation values are strongest during the fall and winter periods and for the first two linear orthogonal modes, with generally lower values during the spring and summer periods and for the third linear orthogonal mode. The fourth mode, however, exhibits very little visual or quantitative (not shown) similarity across the three datasets, which suggests that this mode is perhaps related to the inherent variability within each respective dataset. Interestingly, for certain months, CESM1 LENS compares better with ERA5, and for other months, CESM1 LENS compares better with NOAA-CIRES-DOE

Table 3 Same as Table 1, but for ERA5 data

LOD Results for ERA5 (1950–2018)					
	$d\hat{p}^1$	$d\hat{p}^2$	$d\hat{p}^3$	$d\hat{p}^4$	R-Squared Value
January	IVTn(69 °W,40 °N) 0.7071 [0.60,0.82] 50.0%	U850(70 °W,46 °N) 0.4780 [0.37,0.59] 22.8%	U850(122 °E,25 °N) 0.2293 [0.12,0.34] 5.3%	U850(18 °E,13 °N) 0.1810 [0.07,0.29] 3.3%	0.814
February	IVTn(71 °W,44 °N) 0.6525 [0.54,0.77] 42.6%	U850(103 °W,22 °N) 0.4276 [0.31,0.55] 18.3%	U850(71 °E,41 °N) 0.2937 [0.18,0.41] 8.6%	Z500(21 °W,13 °N) 0.2920 [0.18,0.41] 8.5%	0.780
March	IVTn(66 °W,42 °N) 0.5266 [0.39,0.66] 27.7%	U850(70 °W,45 °N) 0.4798 [0.35,0.61] 23.0%	U850(91 °E,23 °N) 0.3873 [0.25,0.52] 15.0%	Q850(178 °W,15 °N) 0.2395 [0.11,0.37] 5.7%	0.715
April	IVTn(69 °W,41 °N) 0.6635 [0.53,0.79] 44.0%	U850(66 °W,45 °N) 0.4085 [0.28,0.54] 16.7%	IVTn(93 °E,39 °N) 0.2636 [0.13,0.39] 6.9%	IVTe(48 °E,27 °N) 0.2266 [0.10,0.36] 5.1%	0.728
May	IVTn(67 °W,43 °N) 0.7061 [0.60,0.81] 49.9%	IVTe(67 °W,35 °N) 0.4215 [0.32,0.53] 17.8%	PSL(69 °W,55 °N) 0.3042 [0.20,0.41] 9.3%	PSL(12 °W,15 °N) 0.2338 [0.13,0.34] 5.5%	0.823
June	IVTn(65 °W,44 °N) 0.6028 [0.48,0.72] 36.3%	U850(71 °W,46 °N) 0.4446 [0.33,0.56] 20.0%	IVTe(71 °W,39 °N) 0.3621 [0.24,0.48] 13.1%	T2M(16 °E,29 °N) 0.2785 [0.16,0.40] 7.8%	0.770
July	IVTn(101 °W,40 °N) 0.6113 [0.45,0.77] 37.4%	U850(71 °W,47 °N) 0.3823 [0.22,0.54] 14.6%	PSL(127 °E,35 °N) 0.2401 [0.08,0.40] 5.8%		0.578
August	IVTn(71 °W,40 °N) 0.5323 [0.38,0.69] 28.3%	Q850(62 °E,21 °N) 0.4587 [0.31,0.61] 21.0%	T2M(177 °E,16 °N) 0.3569 [0.20,0.51] 12.7%		0.621
September	IVTn(70 °W,39 °N 0.5453 [0.41,0.68] 29.7%	U850(70 °W,46 °N) 0.4816 [0.35,0.61] 23.2%	U850(151 °W,48 °N) 0.3329 [0.20,0.47] 11.1%	IVTe(143 °E,50 °N) 0.2831 [0.15,0.42] 8.0%	0.720
October	IVTn(70 °W,40 °N) 0.6543 [0.54,0.77] 42.8%	U850(70 °W,47 °N) 0.4646 [0.35,0.58] 21.6%	Q850(124 °W,28 °N) 0.2977 [0.18,0.41] 8.9%	IVTn(100 °E,9 °N) 0.2351 [0.12,0.35] 5.5%	0.788
November	IVTn(71 °W,39 °N) 0.6535 [0.54,0.77] 42.7%	U850(70 °W,47 °N) 0.4781 [0.36,0.59] 22.8%	V850(93 °E,51 °N) 0.3045 [0.19,0.42] 9.3%	PSL(132 °E,26 °N) 0.2028 [0.09,0.32] 4.1%	0.789
December	IVTn(69 °W,40 °N) 0.7119 [0.63,0.80] 51.1%	U850(69 °W,46 °N) 0.4741 [0.38,0.56] 22.3%	Z500(114 °W,37 °N) 0.3377 [0.24,0.42] 11.0%	IVTe(19 °E,23 °N) 0.2082 [0.10,0.27] 3.4%	0.877

20CRv3, which again highlights some of the variability that exists across the three datasets. These results demonstrate several key points: (1) our methodology applies reasonably well to both reanalysis products and a large ensemble model dataset, and (2) the lack of consistency after the third iteration of the methodology amongst the three datasets suggests that the first three modes are the only modes that we can be reasonably confident in interpreting as drivers of precipitation in the Northeast.

With regards to explained variance in January, ERA5 produces an R-squared value of 0.81, NOAA-CIRES-DOE 20CRv3 a value of 0.75, and CESM1 LENS a value of 0.64.

Table 4 Same as Table 1, but for NOAA-CIRES-DOE 20CRv3 data

LOD Results for NOAA-CIRES-DOE 20CRv3 (1950–2015)					
	$d\hat{p}^1$	$d\hat{p}^2$	$d\hat{p}^3$	$d\hat{p}^4$	R-Squared Value
January	IVTn(71 °W,39 °N) 0.6280 [0.50,0.76] 39.4%	IVTn(104 °W,52 °N) 0.4260 [0.30,0.56] 18.2%	IVTe(110 °E,39 °N) 0.3259 [0.20,0.46] 10.6%	IVTn(126 °W,2 °N) 0.2556 [0.13,0.38] 6.5%	0.747
February	Q850(71 °W,41 °N) 0.6789 [0.55,0.81] 46.1%	T2M(92 °W,42 °N) 0.3864 [0.26,0.51] 14.9%	Q850(107 °W,72 °N) 0.2661 [0.14,0.39] 7.1%	Q850(89 °W,12 °N) 0.2651 [0.14,0.39] 7.0%	0.751
March	IVTn(70 °W,39 °N) 0.5420 [0.39,0.70] 29.4%	T2M(88 °E,24 °N) 0.3691 [0.21,0.52] 13.6%	IVTn(79 °W,4 °N) 0.3261 [0.17,0.48] 10.6%	IVTn(166 °E,74 °N) 0.3117 [0.16,0.47] 9.7%	0.633
April	IVTn(69 °W,40 °N) 0.7049 [0.58,0.83] 50.0%	U850(67 °W,47 °N) 0.3342 [0.21,0.46] 11.2%	PSL(100 °E,21 °N) 0.2690 [0.14,0.40] 6.4%	Q850(82 °W,19 °N) 0.2533 [0.12,0.38] 5.2%	0.728
May	IVTn(68 °W,45 °N) 0.6832 [0.57,0.80] 46.7%	U850(71 °W,36 °N) 0.4352 [0.32,0.55] 18.9%	T2M(140 °E,61 °N) 0.2966 [0.18,0.41] 8.8%	U850(100 °E,48 °N) 0.2270 [0.11,0.34] 5.2%	0.796
June	IVTn(68 °W,42 °N) 0.6114 [0.48,0.74] 37.4%	IVTe(74 °W,37 °N) 0.4232 [0.29,0.56] 17.9%	V850(137 °E,11 °N) 0.3236 [0.19,0.46] 10.5%	U850(82 °E,44 °N) 0.2733 [0.14,0.41] 7.5%	0.732
July	V850(73 °W,39 °N) 0.6938 [0.56,0.83] 48.1%	IVTe(8 °W,34 °N) 0.3812 [0.24,0.52] 14.5%	IVTn(109 °W,31 °N) 0.2713 [0.13,0.41] 7.4%		0.700
August	IVTn(73 °W,39 °N) 0.6064 [0.46,0.77] 36.8%	IVTn(22 °E,40 °N) 0.4004 [0.25,0.55] 16.0%	V850(67 °E,32 °N) 0.3562 [0.21,0.51] 12.7%		0.655
September	V850(71 °W,40 °N) 0.5827 [0.45,0.72] 34.0%	U850(69 °W,48 °N) 0.4522 [0.32,0.59] 20.5%	Q850(73 °W,40 °N) 0.3115 [0.18,0.45] 9.7%	Q850(141 °E,12 °N) 0.2913 [0.16,0.43] 8.5%	0.726
October	PSL(75 °W,38 °N) 0.6730 [0.54,0.81] 45.3%	IVTe(131 °E,9 °N) 0.3473 [0.21,0.48] 12.1%	PSL(58 °W,39 °N) 0.2903 [0.16,0.43] 8.4%	U850(169 °W,39 °S) 0.2503 [0.12,0.39] 6.3%	0.720
November	IVTn(74 °W,35 °N) 0.5516 [0.41,0.70] 30.4%	IVTe(76 °W,48 °N) 0.3913 [0.25,0.54] 15.3%	T2M(118 °E,46 °N) 0.3471 [0.20,0.49] 12.0%	U850(2 °E,2 °S) 0.3219 [0.18,0.47] 10.4%	0.682
December	V850(69 °W,41 °N) 0.6922 [0.58,0.81] 47.9%	Z500(123 °W,46 °N) 0.4378 [0.32,0.55] 19.2%	IVTn(98 °E,40 °N) 0.2738 [0.16,0.39] 7.5%	IVTe(86 °W,44 °N) 0.2334 [0.12,0.35] 5.4%	0.800

The higher R-squared values associated with the reanalysis products begs the question as to whether there is overfitting of the MLR model, due to the relatively small sample size of the reanalysis products. To determine whether overfitting is an issue here, a train/test procedure was used for the reanalysis results. For a given $\hat{\mathbf{p}}^0$ time series and the selected time series associated with each dataset, the data was partitioned, such that 80% of the data was used to train a linear model, and the remaining 20% was used to test the quality of the model. In order to get a robust result, 1000 random train/test samples were taken. The linear models were trained, and the test sets were used to see how well the models performed, as

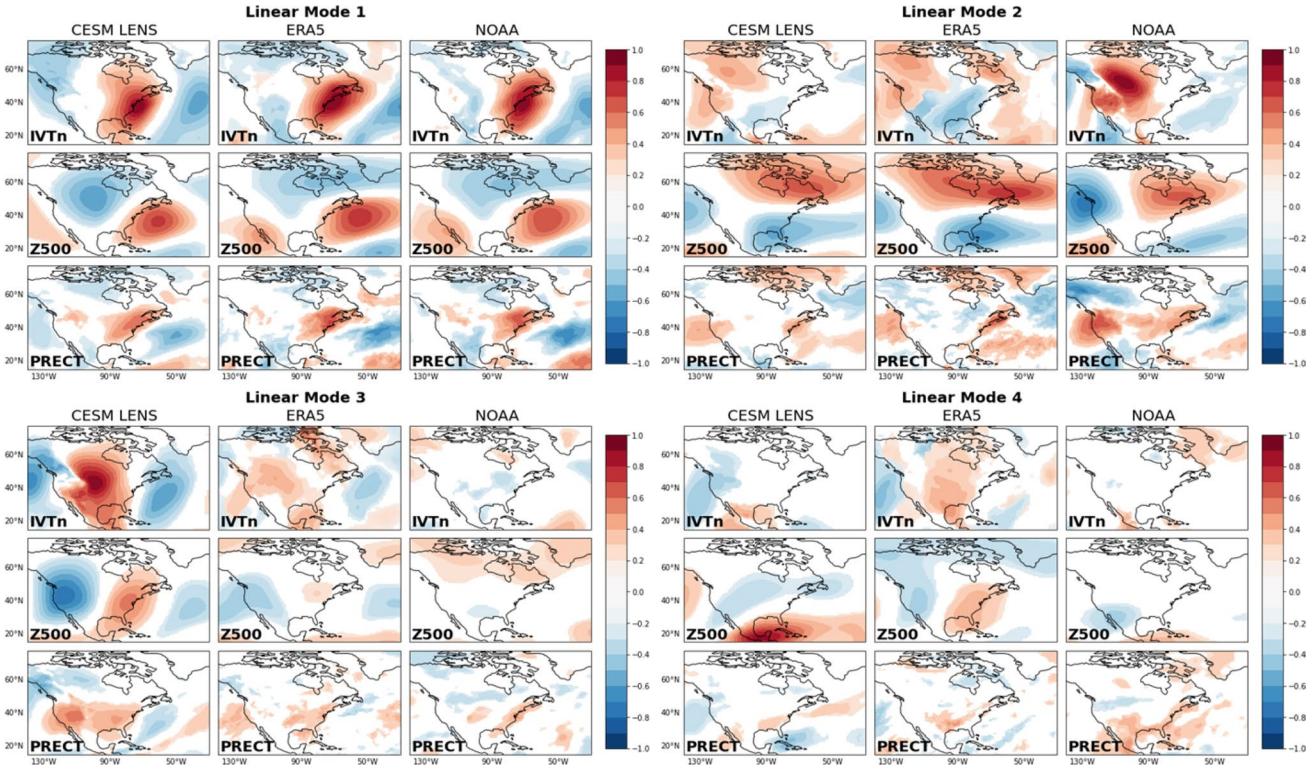

Fig. 11 Comparison for January between CESM1 LENS, ERA5, and NOAA-CIRES-DOE 20CRv3 for the projection of the first four linear orthogonal modes onto the respective IVTn, Z500, and PRECT fields. Contour shading indicates significance at the 95% confidence level

determined by the R-squared values from the 1000 samples. Nearly 40% of samples predict a R-squared value comparable to the R-squared value determined when all of the dataset is used. This seems to indicate that there is not an overfitting issue for the reanalysis datasets, and the increased level of predictability given by reanalysis products as compared to CESM1 LENS may be simply due to the difference in sample size and the overall methods used to generate the respective datasets.

Another question is whether the use of single ensemble members or smaller subsets of ensemble members (than using all 35) produce similar predictability and LOD modes to reanalysis, or are the shortcomings of the model output mitigated in part by using most of the ensemble members? To address this question, we implement the LOD method on single ensemble members and smaller non-overlapping groups of five and eleven ensemble members, respectively, and calculate the spatial correlation of the resulting LOD fields for Z500, IVTn, and PRECT compared to those of ERA5 in the domain of 25–60 °N and 60–100 °W. For individual ensemble members and some groupings of five and eleven ensemble members, R-squared values are comparable to that of ERA5, in the range of 0.65–0.95; however, the spatial correlation values are either small or negative for some of these, indicating that the LOD modes are not being captured well. Example scatterplot results for this approach in January are shown in Fig. 12. For all three fields and for each of the linear modes, there is a broad range of spatial correlation values when considering individual ensemble members. This spread is reduced, and spatial correlation values are improved, as we include more ensemble members, with the best spatial correlation generally obtained when all 35 ensemble members are considered. This result highlights the usefulness and necessity of considering most of the ensemble members when working with CESM1 LENS. If one were to implement LOD on a single ensemble member at random, it could potentially lead to results that are unrepresentative of the larger ensemble. Given that there is good similarity between the results for the reanalysis products and CESM1 LENS, it is thus reasonable to suggest that use of the larger sample size of CESM1 LENS is a compromise between accurately capturing the LOD modes versus reduced predictability in terms of R-squared values relative to the reanalysis products.

Overall, the consistency of the results across the datasets when our methodology is implemented provides confidence in its use. And again, the lack of consistency across the datasets once the third mode is extracted suggests that only three important modes are robust in our regional analysis. These findings demonstrate that the CESM1 LENS dataset

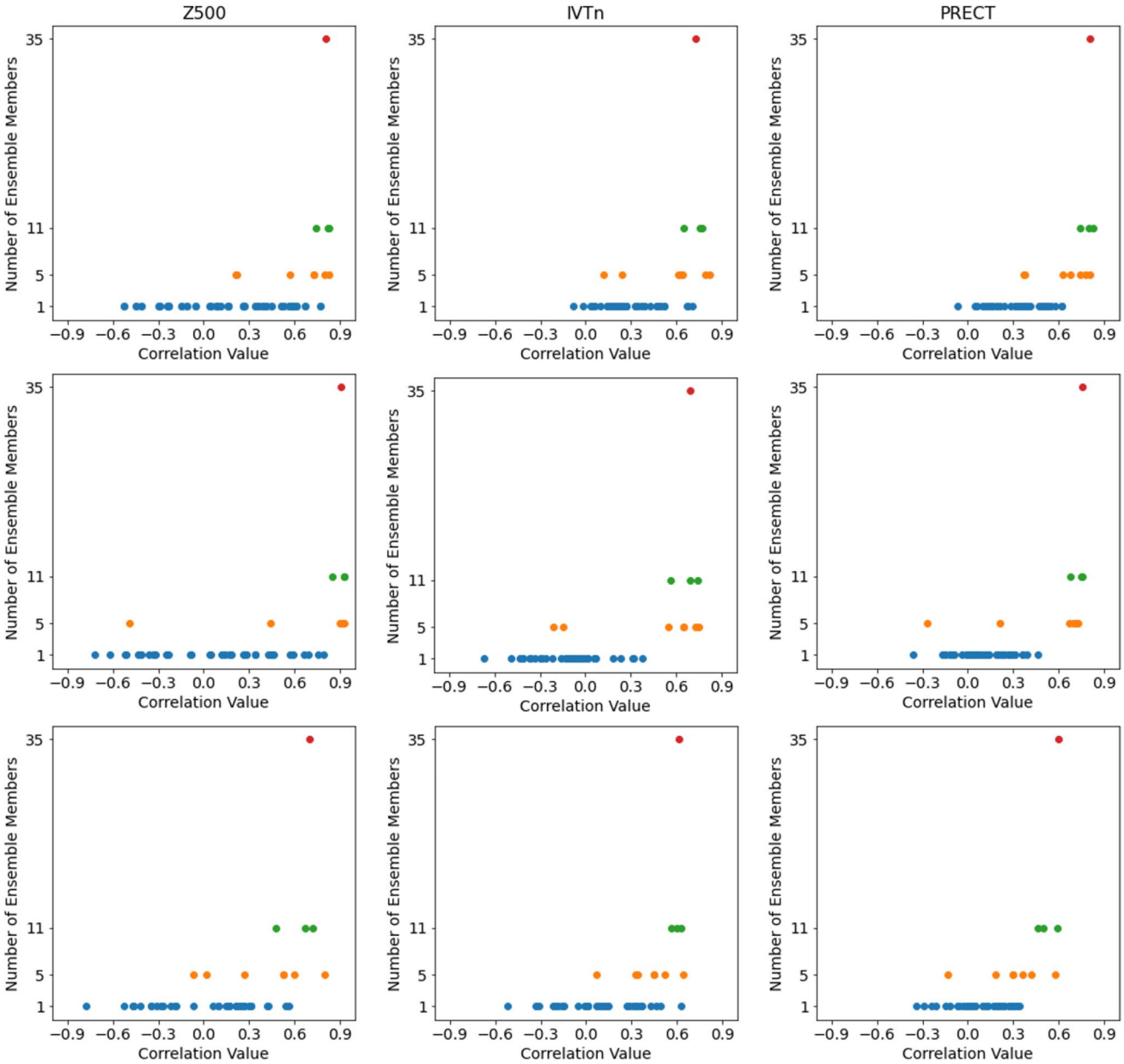

Fig. 12 Scatterplots of spatial correlation values between ERA5 and groups of 1, 5, 11, and all 35 ensemble members, respectively, from CESM1 LENS associated with January plots for Z500 (first column), IVTn (second column), and PRECT (third column). Rows in each column represent the first, second, and third linear modes, respectively

approximates real-word observations reasonably well when a large sample size is considered, and thus can be useful for additional analyses related to understanding future precipitation conditions.

4.4 CESM1 LENS: multiple linear regression using climate indices to predict linear modes

As discussed in Ning and Bradley (2014), several modes of climate variability have important influences across the Northeast, and these influences manifest in certain LSMPs that act to suppress or enhance precipitation. For example, during years with high (above one standard deviation) NAO, there are significant high sea-level pressure anomalies over the northern part of the northeast United States, which can act to block winter storms from passing through and thus reduce wintertime precipitation (Ning et al. 2012b). Previous studies have also shown that the PNA pattern varies considerably with differences in tropical Pacific SST anomalies (Straus and Shukla 2002), and the PNA can, in turn,

influence conditions over the Northeast. Overall, the three teleconnection patterns discussed above correspond to different placement and orientation of high pressure over North America, which can lead to different moisture transport patterns over the Northeast (Straus and Shukla 2000; 2002).

Following the methodology discussed in Sect. 3.3, we can assess how well our linear modes and precipitation data can be predicted by the linear combination (MLR) of some of these teleconnection patterns. Generally speaking, the NAO, PNA, and ENSO serve as statistically significant predictors (p-value < 0.01) in our MLR model. The predictability of our linear modes by these teleconnection patterns is given in the last column of Table 5, where the three rows for each month in that column represent the MLR R-squared value for the first, second and third linear modes, respectively. Our discussion here focuses on the winter and late fall periods, as these periods generally have the highest R-squared values from the MLR. This is perhaps not surprising, as precipitation variability in the late spring and summer seasons is less significantly related to the large-scale circulation than during the fall and winter periods (Ning et al. 2012a). In December, January, February, March, October, and November, the second and third modes are reasonably predicted by some of the teleconnection patterns (Table 5). The result for March is not surprising, as winter conditions can persist into March over this region (Kunkel and Angel 1999; Bradbury et al. 2003). Recall that the second linear mode during the winter period is associated with extratropical cyclone activity south of the Northeast, and the third mode is associated with vapor transport from the Gulf of Mexico. Our results here are in agreement with those of previous studies such as Straus and Shukla (2002), Ning and Bradley (2014), and Risser et al. (2021) regarding the role of the NAO, PNA, and ENSO in influencing wintertime precipitation totals over the Northeast by modulating cyclone development and the 3D circulation.

The teleconnection patterns are not good predictors of $\widehat{\mathbf{p}}^0$ (second column of Table 5), as R-squared values are generally small. This result is in agreement with previous studies such as Bradbury et al. (2002a) and Archambault et al. (2008) that did not find robust direct correlations between total winter precipitation and individual teleconnection patterns. As we have shown, different teleconnection patterns can have important influences on different precipitation modes that are decomposed from the total precipitation field. Additionally, it appears that many of these teleconnection patterns act only to influence the 3D circulation (i.e., zonal and meridional flow, trough/ridge location, etc..) and not necessarily the moisture processes that are most important for precipitation development. As discussed earlier in Sect. 4 and reiterated here, these teleconnection patterns seem to manifest themselves via the large-scale circulation. This is confirmed by including the indices for the teleconnection patterns as predictors along with the LOD selected time series in the MLR models to predict $\widehat{\mathbf{p}}^0$ (third column of Table 5). The addition of the climate modes does not lead to a substantial increase in the predictability of $\widehat{\mathbf{p}}^0$ (this predictability is given in the first column of Table 5, using the first three linear orthogonal modes), and this reinforces the point that their influence is already encapsulated in our extracted modes. This point is further reinforced by generating composites for the Z500 field associated with ± 1 standard deviation of our linear orthogonal mode time series, and then comparing those composites with corresponding ones associated with ± 1 standard deviation for the time series for NAO, Nino3.4, and PNA. This comparison is quantified by calculating the spatial correlation across the composites in the domain of 25–60 ° N and 60–100 ° W (tabulated spatial correlation values given in Sect. 3 of the Online Resource). Focusing on the winter period (December–March), there is excellent correlation between the linear orthogonal mode composites and the climate mode composites (magnitude of correlation values in the range of around 0.60–0.90). Generally, it seems as though the teleconnection patterns drive the large-scale circulation in which our linear orthogonal modes are embedded, but the circulation itself does not necessarily drive precipitation. Instead, the large-scale circulation acts to generate environments that are conducive (or not conducive) for precipitation to occur.

5 Summary and discussion

The present study aims to build upon the existing literature related to assessing the large-scale drivers associated with monthly precipitation variability in the Northeast via the use of a novel linear orthogonal decomposition method.

The results of our analysis are summarized in part here:

1) During the winter (Figs. 1, 2, and 3) and fall periods (available in the Online Resource), the dominant, large-scale atmospheric drivers of monthly precipitation variability in the Northeast are integrated vapor transport northward along the Atlantic seaboard (first linear mode), extratropical cyclone activity (second linear mode) and associated frontal systems, and integrated vapor transport from the Gulf of Mexico (third linear mode). In the summer season (Figs. 7 and 8), while vapor transport along the Atlantic seaboard is still important, extratropical cyclone activity is reduced, and the third linear mode is difficult to physically interpret, although we believe it may be related to MCSs. As constructed, these modes are linearly independent from one another, and so multiple such modes over a given month can contribute to precipitation totals for that month.

Table 5 Multiple linear regression results related to teleconnection patterns for CESM1 LENS

CESM1 LENS MLR Results Including Teleconnection Patterns				
	Predictability w/ LOD Modes	Predictability w/ Teleconnection Patterns	Predictability w/ LOD Modes + Teleconnection Patterns	Predictability of LOD Modes Using Teleconnection Patterns
January	0.624	0.109	0.635	0.116
				0.160
				0.117
February	0.628	0.087	0.630	0.023
				0.253
				0.332
March	0.619	0.071	0.624	0.023
				0.267
				0.330
April	0.580	0.054	0.591	0.015
				0.117
				0.132
May	0.600	0.043	0.602	0.025
				0.088
				0.036
June	0.643	0.019	0.646	0.013
				0.059
				0.013
July	0.646	0.053	0.650	0.032
				0.011
				0.104
August	0.565	0.035	0.569	0.016
				0.033
				0.216
September	0.527	0.010	0.532	0.048
				0.041
				0.027
October	0.603	0.032	0.607	0.049
				0.030
				0.174
November	0.572	0.055	0.577	0.024
				0.193
				0.231
December	0.599	0.062	0.607	0.079
				0.163
				0.231

First column is the predictability of precipitation using the first three linear modes from the linear orthogonal decomposition method only. Second column is the predictability of precipitation using only the teleconnection patterns. Third column is the combined predictability of precipitation using both the first three linear modes from the linear orthogonal decomposition method and the teleconnection patterns. Fourth column is the predictability of the linear modes from the linear orthogonal decomposition method using only the teleconnection patterns. Rows in the fourth column are the predictability (R-squared value) for the first, second, and third linear modes, respectively. Listed R-squared values reflect significance at the 99% confidence level

2) MLR models, based on using the LOD modes as predictors, are able to predict monthly, region-averaged precipitation totals reasonably well, with R-squared values in the range of around 0.54 to 0.65 for CESM1 LENS (Table 1), and higher values in the range of 0.58 to 0.88 for the ERA5 (Table 3) and NOAA-CIRES-DOE 20CRv3 (Table 4) reanalysis products. Predictability is generally best during the late fall / winter and early to mid-summer periods, and is generally lowest during April–May and August–September for CESM1 LENS,

due in part to the seasonal cycle of precipitation in the region and the diluted influence of sub-monthly time scale systems in the monthly averaging. Overall, the cap on linear predictability may be related to the inherent, complex variability of regional precipitation that cannot be fully encapsulated by linear prediction methods.

3) CESM1 LENS compares reasonably well with reanalysis products both in terms of their linear orthogonal modes, and also in the projection of those modes onto the atmospheric fields we have considered (Fig. 11). The first three linear orthogonal modes (or two, in the summer period) are very similar between the three datasets we have analyzed, with significant divergence only occurring with the fourth linear mode across all three datasets. With the inherent variability associated with the fourth linear mode across all three datasets, we conclude that the first three linear modes are likely the most important for understanding the processes associated with monthly precipitation in the Northeast.
4) The large-scale atmospheric drivers of Northeast precipitation, particularly those associated with the second and third linear orthogonal modes, are linked in some capacity with the NAO, ENSO, and PNA (Table 5). We can conclude that the mechanisms by which these teleconnection patterns influence monthly precipitation totals over the region are through modification of the intensity and frequency of the linear orthogonal modes. That is, the teleconnection patterns influence the linear orthogonal modes, which in turn influence the monthly precipitation totals.

Our results are in general agreement with the existing literature related to understanding monthly precipitation variability in the Northeast. While we have not specifically examined smaller-scale phenomena (besides ETCs) that can contribute to this variability, our linear modes indicate that the modulation of precipitation totals in the region are directly linked with the large-scale circulation. In favorable large-scale environments, phenomena such as ETCs, MCSs, and localized convection are likely to occur more frequently, resulting in greater monthly precipitation totals. By contrast, if the large-scale environment is not conducive to the development of these systems, then monthly precipitation totals on average would be reduced. In the context of the drivers within these large-scale environments, as the analysis of our linear modes indicates, there seems to be some seasonal differences in the manifestation of certain drivers. While the three linear modes for winter and fall are robust and are easily interpreted physically by examining our figures, it is more difficult to gain physical insight into the processes associated with the third linear mode in the spring and summer periods. We hypothesize that these seasonal differences may be due in part to the difference in jet stream behavior, as the jet stream is stronger and more easily identified in the cold seasons, and as such the large-scale circulation is also more clearly defined. Additionally, these seasonal differences occur across the three datasets we have analyzed, thus ruling out any product-specific deficiency. We have proposed several mechanisms to explain this third mode in the spring and summer periods, but since this mode contributes little to the precipitation variability in these seasons, and tends to vary in structure between the data products, it is unclear if it conveys any physical meaning about the upstream drivers of monthly precipitation.

The findings presented here also provide several opportunities for further work. The applicability of our linear orthogonal decomposition technique to both reanalysis data and model output provides confidence in the reliability of CESM1 LENS in replicating observations. Given this reliability, CESM1 LENS data can be used, and our analysis can be extended, to examine precipitation characteristics and the associated large-scale drivers in the Northeast in the near-future and far-future. Such a study could be a valuable resource for future, regional water management practices in the face of ongoing anthropogenic climate change. Furthermore, our methodology can be reasonably applied to other regions of the world as well. Doing so can provide insight into the precipitation characteristics of those regions, as well as the LSMPs that modulate those characteristics. Finally, as touched on briefly in our discussion, our analysis in this study is linked to linear predictability—it is nonetheless relevant to understand how much additional predictability could be afforded if a nonlinear approach, such as a machine learning model, were employed.

Appendix A

Linear orthogonal decomposition (LOD)

We begin with local precipitation time series vector $\mathbf{p}$ with components $p_k = p(t_k)$ and associated 2D field time series $\mathbf{F}_i(\mathbf{x}_j)$ with components $\mathbf{F}_i(t_k,\mathbf{x}_j)$. Here t_k and $\mathbf{x}_j$ denote temporal and spatial coordinates, respectively. In this study, time is discretized at the monthly time scale, but this method can be applied to other time scales.

The first step involves removal of a reference field (here taken to be the climatological mean) along the time axis in both precipitation and 2D fields:

$$\hat{\mathbf{p}}^0 = \mathbf{p} - \mathrm{E}\left[\mathbf{p}\right], \quad \hat{\mathbf{F}}_i(t_k, \mathbf{x}_j) = \mathbf{F}_i(\boldsymbol{X}_j) - \mathrm{E}[\mathbf{F}_i(\boldsymbol{X}_j)]$$

If an obvious and undesired trend is present in the time series, it can be removed by using a time-dependent linear field instead. Here we can interpret the quantity $\hat{\mathbf{p}}^n$ as the component of the precipitation time series that is

unpredicted at stage n (this is the whole precipitation time series at stage 0). To proceed, we perform an iterative process analogous to a Gram-Schmidt decomposition. Specifically, we iterate over the following three steps ($n=1,\ldots,N$), where N is the termination point:

1. Calculate correlation between $\widehat{\mathbf{p}}^n$ and $\widehat{\mathbf{F}}_i(\mathbf{x}_j)$ at each point (the value of the correlation field should be between -1 and 1):
   $c^n(\mathbf{x}_j)=\frac{\widehat{\mathbf{p}}^{n-1}\cdot\widehat{\mathbf{F}}_i(\mathbf{x}_j)}{\mathrm{StdDev}[\widehat{\mathbf{p}}^{n-1}]\cdot\mathrm{StdDev}[\widehat{\mathbf{F}}_i(\mathbf{x}_j)]}$
2. Identify the point of maximum absolute correlation x^n in field i^n. Define the predicted component of the precipitation field from this point as:
   $\mathbf{d}\widehat{\mathbf{p}}^n = c^n(\mathbf{x}^n)[\widehat{\mathbf{F}}_{i^n}(\mathbf{x}^n) - \sum_{m=1}^{n-1}\frac{\widehat{\mathbf{F}}_{i^n}(\mathbf{x}^n)\cdot\mathbf{d}\widehat{\mathbf{p}}^m}{\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^m}\mathbf{d}\widehat{\mathbf{p}}^m]$.
   Note that $\mathbf{d}\widehat{\mathbf{p}}^n$ here is multiplied by the correlation by convention.
3. Calculate the remaining unpredicted component of the precipitation:

$$\widehat{\mathbf{p}}^n = \widehat{\mathbf{p}}^0 - \sum_{m=1}^{n}\frac{\widehat{\mathbf{p}}^0\cdot\mathbf{d}\widehat{\mathbf{p}}^m}{\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^m}\mathbf{d}\widehat{\mathbf{p}}^m = \widehat{\mathbf{p}}^{n-1} - \frac{\widehat{\mathbf{p}}^0\cdot\mathbf{d}\widehat{\mathbf{p}}^n}{\mathbf{d}\widehat{\mathbf{p}}^n\cdot\mathbf{d}\widehat{\mathbf{p}}^n}\mathbf{d}\widehat{\mathbf{p}}^n$$

As this is an orthogonal decomposition, the following properties should hold for each step of the process:

(a) $\mathrm{E}[\mathbf{d}\widehat{\mathbf{p}}^n]=0$ and $\mathrm{E}[\widehat{\mathbf{p}}^n]=0$
(b) $\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^n=0$ for $n\neq m$
(c) $\widehat{\mathbf{p}}^n\cdot\mathbf{d}\widehat{\mathbf{p}}^m=0$ for $m<n$
(c) $||\widehat{\mathbf{p}}^m||^2 > ||\widehat{\mathbf{p}}^n||^2$ for $m<n$

By orthogonality of $\mathbf{d}\widehat{\mathbf{p}}^n$ and using $\mathrm{E}[\mathbf{d}\widehat{\mathbf{p}}^n]=0$, we have.
$\mathrm{Var}[\sum_{m=1}^{n}\frac{\widehat{\mathbf{p}}^0\cdot\mathbf{d}\widehat{\mathbf{p}}^m}{\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^m}\mathbf{d}\widehat{\mathbf{p}}^m]=\sum_{m=1}^{n}\left(\frac{\widehat{\mathbf{p}}^0\cdot\mathbf{d}\widehat{\mathbf{p}}^m}{\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^m}\right)^2\mathrm{Var}[\mathbf{d}\widehat{\mathbf{p}}^m]$
$=\frac{1}{N-1}\sum_{m=1}^{n}\frac{(\widehat{\mathbf{p}}^0\cdot\mathbf{d}\widehat{\mathbf{p}}^m)^2}{\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^m}$
where N is the number of elements in $\mathbf{d}\widehat{\mathbf{p}}^m$.
Using $\mathrm{E}[\mathbf{d}\widehat{\mathbf{p}}^n]=0$, we have.
$\mathrm{Var}[\widehat{\mathbf{p}}^0]=\frac{1}{N-1}(\widehat{\mathbf{p}}^0\cdot\widehat{\mathbf{p}}^0)$
Combining these equations, the total fraction of variance (*fov*) explained at stage n of the process is given by:
$Fov=\sum_{m=1}^{n}fov^m$ where $fov^m=\frac{(\widehat{\mathbf{p}}^0\cdot\mathbf{d}\widehat{\mathbf{p}}^m)^2}{(\widehat{\mathbf{p}}^0\cdot\widehat{\mathbf{p}}^0)(\mathbf{d}\widehat{\mathbf{p}}^m\cdot\mathbf{d}\widehat{\mathbf{p}}^m)}$
The total fraction of variance is equal to the R-squared provided by an MLR model acting either on time series $\{\mathbf{d}\widehat{\mathbf{p}}^1,\ldots,\mathbf{d}\widehat{\mathbf{p}}^n\}$ or $\{\widehat{\mathbf{F}}_{i^1}(\mathbf{x}^1),\ldots,\widehat{\mathbf{F}}_{i^n}(\mathbf{x}^n)\}$.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00382-022-06289-y.

Acknowledgements This research was supported by the RGMA program area(s) in the U.S. Department of Energy's Office of Biological and Environmental Research as part of the multi-program, collaborative Integrated Coastal Modeling (ICoM) project. Paul A. Ullrich was supported by Department of Energy Office of Science award number DE-SC0016605, "A Framework for Improving Analysis and Modeling of Earth System and Intersectoral Dynamics at Regional Scales (HyperFACETS)." Paul A. Ullrich and Richard Grotjahn are also supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, hatch project under California Agricultural Experiment Station project accession nos. 1016611 and 1010971, respectively. We would also like to thank the two anonymous reviewers for their helpful comments regarding the revision of this manuscript.

Author contributions RS conducted most of the data analysis for this study, and is the primary author of the manuscript. Dr. PU assisted with some of the data analysis, in particular the use of the TempestExtremes software package, and also provided feedback, comments, and editing of the manuscript and guidance on research tasks. Dr. RG likewise provided feedback, comments, and editing of the manuscript, and guidance on research tasks.

Funding This research was supported by the RGMA program area(s) in the U.S. Department of Energy's Office of Biological and Environmental Research as part of the multi-program, collaborative Integrated Coastal Modeling (ICoM) project. Paul A. Ullrich was supported by Department of Energy Office of Science award number DE-SC0016605, "A Framework for Improving Analysis and Modeling of Earth System and Intersectoral Dynamics at Regional Scales (HyperFACETS)." Paul A. Ullrich and Richard Grotjahn are also supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, hatch project under California Agricultural Experiment Station project accession nos. 1016611 and 1010971, respectively.

Data availability The CESM1 LENS data are available at https://www.earthsystemgrid.org/dataset/ucar.cgd.ccsm4.CESM1_CAM5_BGC_LE.html. ERA5 reanalysis data can be accessed and downloaded from Copernicus Climate Change services at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5&type=dataset. NOAA-CIRES-DOE 20CRv3 data can be accessed and downloaded at https://psl.noaa.gov/data/gridded/data.20thC_ReanV3.html.

Software availability Data analysis for this study was conducted using the Anaconda software distribution of the Python programming language and its included packages and modules. Anaconda is available for download at https://anaconda.com/. Maps created as part of this study were generated using the Cartopy Python package and its associated dependencies. Documentation and instructions for installation of Cartopy can be found at https://scitools.org.uk/cartopy/docs/latest/.

Declarations

Conflict of interest The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Agel L, Barlow M, Qian J-H, Colby F, Douglas E, Eichler T (2015) Climatology of daily precipitation and extreme precipitation events in the northeast United States. J Hydrometeorol 16:2537–2557. https://doi.org/10.1175/JHM-D-14-0147.1

Agel L, Barlow M, Feldstein SB, Gutowski WJ (2017) Identification of large-scale meteorological patterns associated with extreme precipitation in the US northeast. Clim Dyn. https://doi.org/10.1007/s00382-017-3724-8

Agel L, Barlow M, Colby F et al (2019) Dynamical analysis of extreme precipitation in the US northeast based on large-scale meteorological patterns. Clim Dyn 52:1739–1760. https://doi.org/10.1007/s00382-018-4223-2

Archambault HM, Bosart LF, Keyser D, Aiyyer AR (2008) Influence of large-scale flow regimes on cool-season precipitation in the northeastern United States. Mon Weather Rev 136:2945–2963. https://doi.org/10.1175/2007MWR2308.1

Barlow M (2011) Influence of hurricane-related activity on North American extreme precipitation. Geophys Res Lett 38:L04705. https://doi.org/10.1029/2010GL046258

Barnston AG, Livezey RE (1987) Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. Mon Weather Rev 115:1083–1126. https://doi.org/10.1175/1520-0493(1987)115%3c1083:CSAPOL%3e2.0.CO;2

Bosart LF, Pagnotti V, Lettau B (1973) Climatological aspects of eastern United States back-door cold frontal passages. Mon Wea Rev 101:627–635. https://doi.org/10.1175/1520-0493(1973)101,0627:CAOEUS.2.3.CO;2

Bradbury JA, Dingman SL, Keim BD (2002a) New England drought and relations with larger scale atmospheric circulation patterns. J Am Water Resour Assoc 38:1287–1299. https://doi.org/10.1111/j.1752-1688.2002.tb04348.x

Bradbury JA, Keim BD, Wake CP (2002b) U.S. east coast trough indices at 500 hPa and New England winter climate variability. J Clim 15:3509–3517. https://doi.org/10.1175/1520-0442(2002)015%3c3509:USECTI%3e2.0.CO;2

Bradbury JA, Keim BD, Wake CP (2003) The influence of regional storm tracking and teleconnections on winter precipitation in the Northeastern United States. Ann Assoc Am Geogr 93:544–556. https://doi.org/10.1111/1467-8306.9303002

Cai M, van den Dool HM (1991) Low-frequency waves and traveling storm tracks. Part I: Barotropic component. J Atmos Sci 48:1420–1436. https://doi.org/10.1175/1520-0469(1991)048%3C1420:LFWATS%3E2.0.CO;2

Catto JL, Pfahl S (2013) The importance of fronts for extreme precipitation. J Geophys Res Atmos 118(10791–10):801. https://doi.org/10.1002/jgrd.50852

Catto JL, Pfahl S, Jakob C, Berry G, Nicholls N (2012) Relating global precipitation to atmospheric fronts. Geophys Res Lett 39:L10805. https://doi.org/10.1029/2012GL051736

DeGaetano AT (2009) Time-dependent changes in extreme- precipitation return-period amounts in the continental United States. J Appl Meteor Climatol 48:2086–2099. https://doi.org/10.1175/2009JAMC2179.1

Feng Z et al (2019) Spatiotemporal characteristics and large-scale environments of mesoscale convective systems east of the Rocky Mountains. J Clim 32(21):7303–7328. https://doi.org/10.1175/JCLI-D-19-0137.1

Grotjahn R et al (2016) North American extreme temperature events and related large scale meteorological patterns: a review of statistical methods, dynamics, modeling, and trends. Clim Dyn 46:1151–1184. https://doi.org/10.1007/s00382-015-2638-6

Harman JR (1991) Synoptic Climatology of the Westerlies: Process and Patterns. Association of American Geographers, 80 pp.

Hartley S, Keables MJ (1998) Synoptic associations of winter climate and snowfall variability in New England, USA, 1950–1992. Int J Climatol 18:281–298. https://doi.org/10.1002/(SICI)1097-0088(19980315)18:3%3c281::AID-JOC245%3e3.0.CO;2-F

Hawcroft MK, Shaffrey LC, Hodges KI, Dacre HF (2012) How much Northern Hemisphere precipitation is associated with extratropical cyclones? Geophys Res Lett 39:L24809. https://doi.org/10.1029/2012GL053866

Hayhoe K, Wake CP, Huntington TG et al (2007) Past and future changes in climate and hydrological indicators in the US Northeast. Clim Dyn 28:381–407. https://doi.org/10.1007/s00382-006-0187-8

Hersbach H and Coauthors (2020) The ERA5 global reanalysis. Quart. J. Roy. Meteor. Soc., https://doi.org/10.1002/qj.3803

Hoskins BJ, Hodges KI (2002) New perspectives on the northern hemisphere winter storm tracks. J Atmos Sci 59:1041–1061. https://doi.org/10.1175/1520-0469(2002)059%3c1041:npotnh%3e2.0.co;2

Howarth ME, Thorncroft CD, Bosart LH (2019) Changes in extreme precipitation in the Northeast United States: 1979–2014. J Hydrometeorol 20:673–689. https://doi.org/10.1175/JHM-D-18-0155.1

Hurrell JW (1995) Decadal trends in the North Atlantic Oscillation: Regional temperature and precipitation. Science 269:676–679. https://doi.org/10.1126/science.269.5224.676

Kay JE, Deser C, Phillips A, Mai A, Hannay C, Strand G, Arblaster J, Bates S, Danabasoglu G, Edwards J, Holland M, Kushner P, Lamarque J-F, Lawrence D, Lindsay K, Middleton A, Munoz E, Neale R, Oleson K, Polvani L, Vertenstein M (2015) The community earth system model (CESM1) large ensemble project: a community resource for studying climate change in the presence of internal climate variability. Bull Am Meteor Soc 96:1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kocin PJ and Uccellini LW (2004) Northeast Snowstorms. Meteor. Monogr., No. 54, Amer. Meteor. Soc., 818 pp. doi:https://doi.org/10.1007/978-1-878220-32-5

Kunkel KE, Angel JR (1999) Relationship of ENSO to snowfall and related cyclone activity in the contiguous United States. J Geophys Res 104:19425–19434. https://doi.org/10.1029/1999JD900010

Kunkel KE, Andsager K, Easterling DR (1999) Long-term trends in extreme precipitation events over the conterminous United States and Canada. J Climate 12:2515–2527. https://doi.org/10.1175/1520-0442(1999)012,2515:LTTIEP.2.0.CO;2

Kunkel KE, Easterling DR, Kristovich DAR, Gleason B, Stoecker L, Smith R (2012) Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous United States. J Hydrometeorol 13:1131–1141. https://doi.org/10.1175/jhm-d-11-0108.1

Kunkel KE et al (2013) Regional climate trends and scenarios for the U.S. National Climate Assessment. Part 1. Climate of the Northeast U.S., NOAA Tech. Rep. NESDIS 142–1. Available online at http://www.nesdis.noaa.gov/technical_reports/NOAA_NESDIS_Tech_Report_142–1-Climate_of_the_Northeast_U.S.pdf

Lau NC (1988) Variability of the observed midlatitude storm tracks in relation to low-frequency changes in the circulation pattern. J Atmos Sci 45:2718–2743. https://doi.org/10.1175/1520-0469(1988)045%3C2718:VOTOMS%3E2.0.CO;2

Leathers DJ, Yarnal B, Palecki MA (1991) The Pacific/North American teleconnection pattern and United States climate. Part I: regional temperature and precipitation associations. J Clim 4:517–528. https://doi.org/10.1175/1520-0442(1991)004%3c0517:TPATPA%3e2.0.CO;2

Lott N and Ross T (2006) Tracking and evaluating U. S. billion-dollar weather disasters, 1980–2005. AMS Forum: Environ- mental Risk and Impacts on Society: Successes and Chal- lenges, Atlanta, GA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/Annual2006/techprogram/paper_100686.htm.]

Melillo, JM, Richmond TC, and Yohe GW (2014) Climate Change Impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program, 841 pp., doi:https://doi.org/10.7930/J0Z31WJ2.

Milrad SM, Atallah EH, Gyakum JR, Dookhie G (2014) Synoptic typing and precursors of heavy warm-season precipitation events at Montreal, Québec. Weather Forecast 29:419–444. https://doi.org/10.1175/WAF-D-13-00030.1

Namias J (1966) Nature and possible causes of the northeastern United States drought during 1962–1965. Mon Weather Rev 94:543–554. https://doi.org/10.1175/1520-0493(1966)094%3c0543:NAPCOT%3e2.3.CO;2

Ning L, Bradley RS (2014) Winter precipitation variability and corresponding teleconnections over the northeastern United States. J Geophys Res Atmos 119:7931–7945. https://doi.org/10.1002/2014JD021591

Ning L, Mann ME, Crane R, Wagener T (2012a) Probabilistic projections of climate change for the mid-Atlantic region of the United States—Validation of precipitation downscaling during the historical era. J Clim 25:509–526. https://doi.org/10.1175/2011JCLI4091.1

Ning L, Mann ME, Crane R, Wagener T, Najjar RG, Singh R (2012b) Probabilistic projections of anthropogenic climate change impacts on precipitation for the mid-Atlantic region of the United States. J Clim 25:5273–5291. https://doi.org/10.1175/JCLI-D-11-00565.1

Pfahl S, Sprenger M (2016) On the relationship between extratropical cyclone precipitation and intensity. Geophys Res Lett 43:1752–1758. https://doi.org/10.1002/2016GL068018

Pfahl S, Wernli H (2012) Quantifying the relevance of cyclones for precipitation extremes. J Climate 25:6770–6780. https://doi.org/10.1175/JCLI-D-11-00705.1

Pryor SC, Howe JA, Kunkel KE (2009) How spatially coherent and statistically robust are temporal changes in extreme precipitation in the contiguous USA? Int J Climatol 29:31–45. https://doi.org/10.1002/joc.1696

Risser MD, Wehner MF, O'Brien JP et al (2021) Quantifying the influence of natural climate variability on in situ measurements of seasonal total and extreme daily precipitation. Clim Dyn 56:3205–3230. https://doi.org/10.1007/s00382-021-05638-7

Ropelewski CF, Halpert MS (1986) North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). Mon Weather Rev 114:2352–2361. https://doi.org/10.1175/1520-0493(1986)114%3c2352:NAPATP%3e2.0.CO;2

Schumacher RS, Johnson RH (2005) Organization and environmental properties of extreme-rain-producing mesoscale convective systems. Mon Weather Rev 133:961–976. https://doi.org/10.1175/MWR2899.1

Seager R, Pederson N, Kushnir Y, Nakamura J, Jurburg S (2012) The 1960s drought and the subsequent shift to a wetter climate in the Catskill Mountains Region of the New York City watershed. J Clim 25:6721–6742. https://doi.org/10.1175/JCLI-D-11-00518.1

Slivinski LC, Compo GP, Whitaker JS, Sardeshmukh PD, Giese BS, McColl C, Allan R, Yin X, Vose R, Titchner H, Kennedy J, Spencer LJ, Ashcroft L, Brönnimann S, Brunet M, Camuffo D, Cornes R, Cram TA, Crouthamel R, Domínguez-Castro F, Freeman JE, Gergis J, Hawkins E, Jones PD, Jourdain S, Kaplan A, Kubota H, Le Blancq F, Lee T, Lorrey A, Luterbacher J, Maugeri M, Mock CJ, Moore GK, Przybylak R, Pudmenzky C, Reason C, Slonosky VC, Smith C, Tinz B, Trewin B, Valente MA, Wang XL, Wilkinson C, Wood K, Wyszyński P (2019) Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. Q J R Meteorol Soc (accepted). https://doi.org/10.1002/qj.3598

Straus DM, Shukla J (2000) Distinguishing between the SST-forced variability and internal variability in mid latitudes: Analysis of observations and GCM simulations. Q J R Meteorol Soc 126:2323–2350. https://doi.org/10.1002/qj.49712656716

Straus DM, Shukla J (2002) Does ENSO force the PNA? J Clim 15:2340–2358. https://doi.org/10.1175/1520-0442(2002)015%3c2340:DEFTP%3e2.0.CO;2

Suriano ZJ, Benjamin AE, Leathers DJ, Schroeter D, Corradina V (2018) Northeast United States growing season moisture conditions: associations with a North American mid-tropospheric wave train. Int J Climatol 38:5542–5550. https://doi.org/10.1002/joc.5746

Trenberth KE (1997) The definition of El Niño. Bull Am Meteorol Soc 78:2771–2777. https://doi.org/10.1175/1520-0477(1997)078%3c2771:TDOENO%3e2.0.CO;2

Trenberth KE, Dai A, Rasmussen RM, Parsons DB (2003) The changing character of precipitation. Bull Amer Meteor Soc 84:1205–1217. https://doi.org/10.1175/BAMS-84-9-1205

Ullrich PA, Zarzycki CM (2017) TempestExtremes: a framework for scale-insensitive pointwise feature tracking on unstructured grids. Geosci Model Dev 10(1069–1090):2017. https://doi.org/10.5194/gmd-10-1069-2017

Wallace JM, Gutzler DS (1981) Teleconnections in the geopotential height field during the Northern Hemisphere winter. Mon Weather Rev 109:784–812. https://doi.org/10.1175/1520-0493(1981)109%3c0784:TITGHF%3e2.0.CO;2