Supporting Information for "The System for Classification of Low-Pressure Systems (SyCLoPS): An All-in-One Objective Framework for Large-scale Datasets"

Yushan Han[1], Paul A. Ullrich[1,2]

[1]Department of Land, Air and Water Resources, University of California, Davis, Davis, CA, USA

[2]Division of Physical and Life Sciences, Lawrence Livermore National Laboratory, Livermore, CA, USA

Contents of this file

Text S1. Track Matching in the Objective ERA5 LPS dataset

1. **TC track matching**: A match between the ERA5 LPS dataset and IB-TC is realized when they lie within 2° GCD of each other at the same timestamp. 98% of the 3510 IB-TC tracks are matched, which forms the matched TC dataset. Additionally, we obtain the matched tropical LPS dataset, of which 96% of the 4495 tracks are matched, by comparing it to main-type tracks in 1979-2021 IBTrACS without the 34 knots wind speed filter. This non-filtered version of IBTrACS is also used for the LPS label POS computation described in Sec. 3 and the selection of the tropical/subtropical cluster discussed in Text S2.

2. **MS track matching**: Detected ERA5 LPS tracks having at least 5 nodes lying within 5° GCD of a record in the daily-frequency Sikka dataset on the same date are considered matches. 88% of the 350 Sikka tracks are matched, which forms the matched MS dataset.

3. **STLC track matching**: A match between the ERA5 LPS dataset and the ERA5 manually-tracked Mediterranean cyclone dataset is realized when they lie within 2° GCD of each other at the same timestamp. 98% of the 129 Mediterranean cyclone tracks, which consist of mainly STLCs and SCs, are matched. Then, potential STLCs are chosen by the method described in Section 2, forming the matched STLC dataset.

4. **PL track matching**: We first round the timestamps in the STARS archive to the nearest 3-hourly timestamps. Detected ERA5 tracks having the most nodes lying within 3° GCD of a record in the STARS dataset at the same timestamp are considered matches. 63% of the 186 STARS tracks are matched, which forms the matched PL dataset.

5. **EW track matching**: Detected ERA5 LPS tracks having the most nodes lying within 5° GCD of a record in the the objectively-tracked EW dataset at the same timestamp are considered matches. 44% of the 1396 easterly wave tracks are matched, which forms the matched EW dataset.

Text S2. Selection of the Tropical/Subtropical Cluster and Sensitivity Test

To make the two clusters more alike in terms of their latitudinal locations, warm cores, and underlying SSTs, we first select a set of candidate LPS tracks that are more likely to be largely non-tropical/subtropical and have a more stable warm core for the subtropical cluster:

1. The track is within the year range of 1979-2021 and cannot be a match to any tracks in the non-filterd IBTrACS.

2. The track must begin poleward of 20° latitude from the equator to demonstrate a potential non-tropical origin.

3. The track must have at least 5 time steps within 25° to 50° latitudes from the equator.

4. The track must establish at least 5 time steps with both upper- and lower-level warm cores (UPTKCC<0 $m^2\,s^{-2}$ and LOTKCC<0 $m^2\,s^{-2}$).

Tracks in the non-filtered IBTrACS are considered candidates for the tropical cluster. Next, we choose LPS nodes from both candidate track sets that meet the subtropical latitudinal requirement of 20° to 45° from the equator (a range of latitudes that frequently host both TCs and SCs), the warm-core requirement of UPTKCC<0 $m^2\,s^{-2}$ and LOTKCC<0 $m^2\,s^{-2}$, and a potential underlying SST (data from ERA5) requirement to only include subtropical cluster's LPSs over a comparable SST to that of the tropical cluster. According to observations, almost all TCs with a maximum wind speed over 35 knots in IBTrACS are over 288 K SSTs, and over 95% of those TCs are over 293 K SSTs (Stansfield & Reed, 2021). We find that the final classification could be sensitive to the SST requirement chosen. Hence, we select a range of minimum SST requirements from 288 K to 295 K with an interval of 0.5 K to perform a sensitivity test on the optimal thresholds needed for the decision trees.

As shown in Fig. S2, DPSH is not sensitive to the SST requirement, and its threshold is stable at around 10 $\mathrm{m\,s^{-1}}$. After 293.5 K, the decision trees with a depth of 2 only split on DPSH (RH100 becomes indecisive) as the sample size of the subtropical cluster decreases more quickly, making the two clusters more imbalanced. This could mean that a larger proportion of tropical LPS nodes are falsely included in the subtropical cluster. For the SST range of 288-293.5 K, sample sizes of the two clusters are comparable, ranging between 50 and 80 thousand. RH100 thresholds (shown with the blue line) within this range are averaged at 20% and they are also most stable around the 20% level after being rounded off to the nearest 5% as indicated by the gray line. Hence, RH100=20% and DPSH=10 $\mathrm{m\,s^{-1}}$ is considered the optimal combination to separate the two clusters. Accuracy scores for these decisions range from 77% to 80%. We use the threshold of SST $\geq$ 291 K to produce an example result for Fig. 3 in Sec. 3. The dashed line shows the decision tree with a depth of 1 when only RH100 is considered. The threshold of RH100 in this situation is more stable at about the 50-55% levels, with a roughly 74% accuracy.

Text S3. Quasi-stationary Track Condition

We calculate three additional track parameters included in the "Additional_Track_Info.csv", which are track linearity, track spread, and track inland ratio. Track spread is the standard deviation of the distance between each node in a track and the first track node; track linearity is the Pearson correlation coefficient of all the nodes within a track on the latitude-longitude coordinate; and track inland ratio is the percentage of nodes located within 1° GCD of any inland areas, which are defined as those grid points with surface geopotential exceeding 150 $m^2\,s^{-2}$. One may expect a typical quasi-stationary LPS to be meandering and bouncing around some topographical features within a limited region. Hence, the linearity of those tracks should be low, the deviation of track nodes from the first detected track node should be rather small, and they reside mostly inland or close to shorelines. We select a set of quasi-permanent LPSs near Colombia in South America as being representative of quasi-stationary tracks. They exist year-round, potentially due to the positive feedback between rainforest evapotranspiration and mesoscale convection (Poveda et al., 2014). Drawing from the intuition mentioned above, we optimize the three quasi-stationary thresholds of the three parameters by maximizing the filtering of the Colombia LPSs and minimizing the overlap between the matched TC tracks and the selected quasi-stationary tracks. The final decision for the quasi-stationary track condition is that a track needs to have a track linearity lower than 0.55, a track spread smaller than 3° GCD, and a track inland ratio greater than 65%.

Text S4. Alternatives to PMX200 thresholds

In regional models, one may assume the region covers mostly polar regions or the subtropics, so the PMX200 thresholds are unnecessary. If the model domain does not cover very high latitudes (i.e., 65° or higher), in the event that the polar jet/front is further poleward of the domain limit, one may use the following alternatives when PMX200 of an LPS is lower than $30 \mathrm{m\,s^{-1}}$ in the SC condition and $25\ \mathrm{m\,s^{-1}}$ in separating PLs:

DPSH and T850 are considered alternatives to PMX200. The decision tree classifier selects the PMX200 alternative thresholds for the SC condition based on two clusters of 100 thousand node samples, each chosen using the PMX200>$30\ \mathrm{m\,s^{-1}}$ or PMX200$\leq 30\ \mathrm{m\,s^{-1}}$ threshold in the SC condition. The result shows that DPSH>$14\ \mathrm{m\,s^{-1}}$ and T850>273 K is the best combination, giving a 77% accuracy. If temperature thresholds are deemed to be avoided under global warming scenarios, DPSH>$12\ \mathrm{m\,s^{-1}}$ alone is also acceptable with a 72% accuracy.

The alternatives to the PMX200 threshold for distinguishing PLs following the TLC condition are found in a similar way as in the SC condition. Results show that DPSH<$11\ \mathrm{m\,s^{-1}}$ and T850<273 K is the best combination, giving an 80% accuracy. Similarly, DPSH<$11\ \mathrm{m\,s^{-1}}$ alone is also acceptable with a 77% accuracy.

Text S5. Justification of Parameter Specifications

It's near-impossible to additionally optimize every parameter selection process (i.e., determining the optimized GCD used in some parameters), as it will require significantly larger time complexity to compute. Most of the chosen parameter specifications are derived directly or indirectly from previous studies. Here is a brief justification of some of the chosen parameters:

1. MSLCC, UPTKCC: These are the same parameters used in the previous TE's TC tracking algorithms (Zarzycki & Ullrich, 2017), which have been optimized with respect to the GCD distance. The 1° GCD offset allowance in UPTKCC to search for a thickness

maximum is also used in Zarzycki & Ullrich (2017) based on observational intuition. They also found that the TC detection result is relatively insensitive to this offset value.

2. CMSLCC: The 2.0° GCD specification in INMSLCC is chosen to represent the "core" of cyclones. For example, Weatherford and Gray (1988) defined the TC outer core as the 1° - 2.5° GCD region from the TC center. The region outside of the core may be regarded as the outer region/wind field (i.e., within the 5.5° GCD as specified in MSLCC).

3. VOR500, RH100, RHAG850: The 2.5° GCD specification is chosen in accordance with CMSLCC with a 0.5° GCD buffer given that they are upper-level parameters.

4. MIDTKCC, LOTKCC, Z500CC: Rather than being assessed for a particular value, these three parameters are computed only to confirm if they are non-zero. Specifically, they indicate whether there are warm cores or a 500 hPa closed circulation close to the core of an LPS. Hence, a rather small GCD of 3.5° is used with a 1.0° offset allowance. Since we only need to know if a low-level warm core is present, it should be acceptable–though untested–to replace the 925 hPa geopotential needed for LOTKCC with 850 hPa geopotential if some datasets or models use a vertical coordinate system with missing values below the surface.

5. DPSH: The 10° GCD specification is chosen to be reasonably large to reflect the local large-scale environment.

6. UDF850: The 5.5° GCD specification is chosen to be the same as MSLCC.

7. PMX200: The 1.0° offset allowance is chosen to be the same as for the others.

Text S6. TE Command Lines and Instructions

This is an example of TE commands for generating all parameters in the ERA5-

based catalogs. The commands below detect LPS centers (nodes) and output parameters for classification or reference purposes. A detailed TE documentation can be found at: `https://climate.ucdavis.edu/tempestextremes.php`.

`latname` and `lonname` in the commands only need to be specified when the latitude and longitude variables in the given dataset use different names other than the standard "lat" and "lon". Specify `logdir` to store temporary log files in the desired folder.

Pointing to your TempestExtremes directory by: `TEMPESTEXTREMESDIR=...`

Define your input and output files (or file lists), for example:

```
inputfile=ERA5_lps_in.txt; outputfile=ERA5_lps_out.txt
```

Parameters are calculated under `outputcmd` using variables in ERA5 in the following order, separated by semi-columns: MSLP;WS;CMSLCC;MSLCC;DPSH;UPTKCC;MIDTKCC; LOTKCC;Z500CC;VO500;RH100;RHAG850;T850;Z850;ZS;UDF850;PMX200.

The input file includes a list of files containing all the required ERA5 variables listed in Table S1 for each time frame (i.e., per day, month, or year) in a txt file. TE can parallel files by each time frame in the list when computing.

The `DetectNodes` command lines start here:

```
$TEMPESTEXTREMESDIR/DetectNodes
  --in_data_list $inputfile --out_file_list $outputfile
  --searchbymin MSL --closedcontourcmd "MSL,10,5.5,0"
  --mergedist 6.0
  --outputcmd "MSL,min,0;_VECMAG(VAR_10U,VAR_10V),max,2.0;
```

```
      MSL,posclosedcontour,2.0,0;MSL,posclosedcontour,5.5,0;
      _DIFF(_VECMAG(U(200hPa),V(200hPa)),_VECMAG(U(850hPa),V(850hPa))),avg,10.0;
      _DIFF(Z(300hPa),Z(500hPa)),negclosedcontour,6.5,1.0;
      _DIFF(Z(500hPa),Z(700hPa)),negclosedcontour,3.5,1.0;
      _DIFF(Z(700hPa),Z(925hPa)),negclosedcontour,3.5,1.0;
      Z(500hPa),posclosedcontour,3.5,1.0;VO(500hPa),avg,2.5;
      R(100hPa),max,2.5;R(850hPa),avg,2.5;
      T(850hPa),max,0.0;Z(850hPa),min,0;ZS,min,0;
      U(850hPa),posminusnegwtarea,5.5;_VECMAG(U(200hPa),V(200hPa)),maxpoleward,1.0"
  --timefilter "3hr" --latname "latitude" --lonname "longitude" --logdir "./TE_log"
```

Next, detected nodes are stitched in consecutive time with parameters' name formatted using `StitchNodes`. The output of it is a single txt file.

```
$TEMPESTEXTREMESDIR/StitchNodes
  --in_list $inputfile --out $outputfile
  --in_fmt "lon,lat,MSLP;WS;CMSLCC;MSLCC;DPSH;UPTKCC;MIDTKCC;LOTKCC;Z500CC;VO500;
      RH100;RHAG850;T850;Z850;ZS;UDF850;PMX200"
  --range 4.0 --mintime "18h" --maxgap "12h" --threshold "MSLCC,>=,100.0,5"
```

Now, to detect LPS size blobs for calculating LPS size, we first need to use `VariableProcessor` to produce smoothed cyclonic relative vorticity from 850 hPa U and V. The `_CURL{8,3}` operator is used below to smooth the vorticity field by evaluating the curl of the wind field using 8 equiangular points at a distance of 3° GCD.

```
$TEMPESTEXTREMESDIR/VariableProcessor --in_data_list $inputfile
  --out_data_list "ERA5_smoothed_850RV.txt" --var "_CURL{8,3}(U(850hPa),V(850hPa))"
  --varout "Vorticity" --latname latitude --lonname longitude
  --timefilter "3hr"
$TEMPESTEXTREMESDIR/VariableProcessor --in_data_list "ERA5_smoothed_850RV.txt"
  --out_data_list "ERA5_smoothed_cyclonic_850RV.txt"
  --var "_COND(_LAT(),Vorticity,_PROD(Vorticity,-1))" --varout "Cyclonic_Vorticity"
  --latname latitude --lonname longitude
```

#Now, detect LPS size blobs for calculating LPS size using `DetectBlobs`. This command outputs detected features marked by binary mask:

```
$TEMPESTEXTREMESDIR/DetectBlobs
  --in_data_list $inputfile --out_list $outputfile
  --thresholdcmd "((Cyclonic_Vorticity,>=,2e-5,0) &
      (_VECMAG(U(925hPa),V(925hPa)),>=,12.0,0)) | (Cyclonic_Vorticity,>=,4e-5,0)"
  --geofiltercmd "area,>=,1e4km2" --tagvar "object_id"
  --latname latitude --lonname longitude --timefilter "3hr" --logdir "./TE_log"
```

Lastly, derive properties of blobs for LPS node pairing using `BlobStats`. The command will output a list of blobs with their unique ID and properties. One can opt to calculate blobs' IKE by `sumvar` in the command (slow if single-threaded):

```
$TEMPESTEXTREMESDIR/BlobStats --in_list $inputfile --out_file $outputfile
  --findblobs --var "object_id"
```

```
  --out "centlon,centlat,minlat,maxlat,minlon,maxlon,area"
  # --sumvar "_SUM(_POW(U(925hPa),2),_POW(V(925hPa),2))"
  --out_fulltime --latname latitude --lonname longitude
```

Optionally, one can use `StitchBlobs` to give each blob in the output of `DetectBlobs` an ID identical to those output by `BlobStats`. It can then be combined with the blob-node pairing process documented in "LOWSIZE_pair_cal.py" to generate blobs with new designated blob IDs according to their corresponded LPS classes. For example, all blobs associated with EX-labeled nodes can be tagged with an ID of 1.

```
TEMPESTEXTREMESDIR/StitchBlobs
  --in_list $outputfile_from_DetectBlobs --out_list $outputfile
  --var "object_id" --tagonly --latname latitude --lonname longitude
```

Then, one may calculate the IKE at each grid point contained within each size blob that are tagged "1" by something like:

```
$TEMPESTEXTREMESDIR/VariableProcessor
  --in_data_list $inputfile --out_list $outputfile
  --var "_PROD(_EQUALS(object_id,1),
      _PROD(_SUM(_POW(U(925hPa),2),_POW(V(925hPa),2)),0.5),_AREA())"
  --varout "ike_lps" --latname latitude --lonname longitude
```

This printed TE command lines are also available via Zenodo in the shell script:

The TE shell script: "TE_commands.sh"

The following files are also provided via Zenodo:

The classifier: "SyCLoPS_classifier.py"

Blob-node pairing and LOWSIZE calculator: "LOWSIZE_pair_cal.py"

Example uses of the classified catalog: "SyCLoPS_examples.py"

The input LPS catalog: "SyCLoPS_input.parquet"

The output classified LPS catalog: "SyCLoPS_classified.parquet"

The additional track information file: "Additional_Track_info.csv"

The labeled size blobs of each year: "size_blob_1979_2022.tar.gz"

The labeled precipitation blobs of each year: "preci_blob_1979_2022.tar.gz"

The size and precipitation blob tag file each has 44 compressed nc files for each year. Decompress each nc file before using them to optimize computation performance. Blobs are tagged with five different labels (ID numbers from 1-5). Blobs associated (paired) with 1 = TC nodes in TC tracks; 2 = TD and TLO (TLO(ML), TD(MD), TLO, and TD) nodes in MS tracks; 3 = STLC nodes in STLC tracks; 4 = PL nodes in PL tracks; and 5 = other LPS nodes. Users may alter this ID system using the `StitchBlob`'s outputs. For example, blobs associated with all nodes with `Tropical_Flag`=1 in TC tracks can be under one ID, and blobs paired with all nodes with `Tropical_Flag`=1 in MS but not TC tracks can be under another ID.

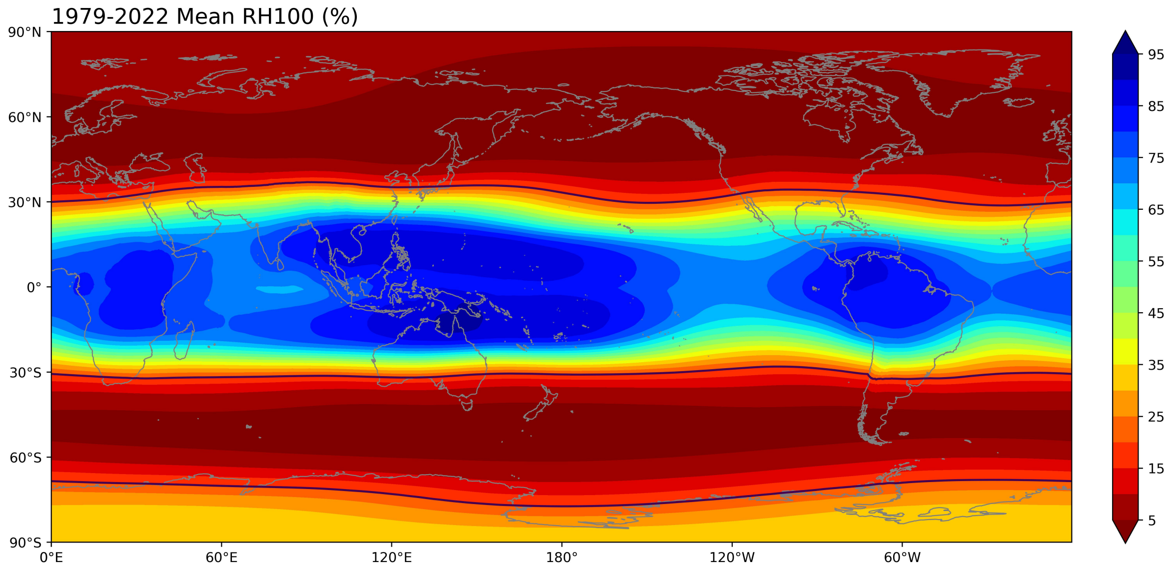

Figure S1. 1979-2022 mean RH100 at each grid point. The black solid-line contour is the 20% RH100 threshold we choose for the tropical condition.

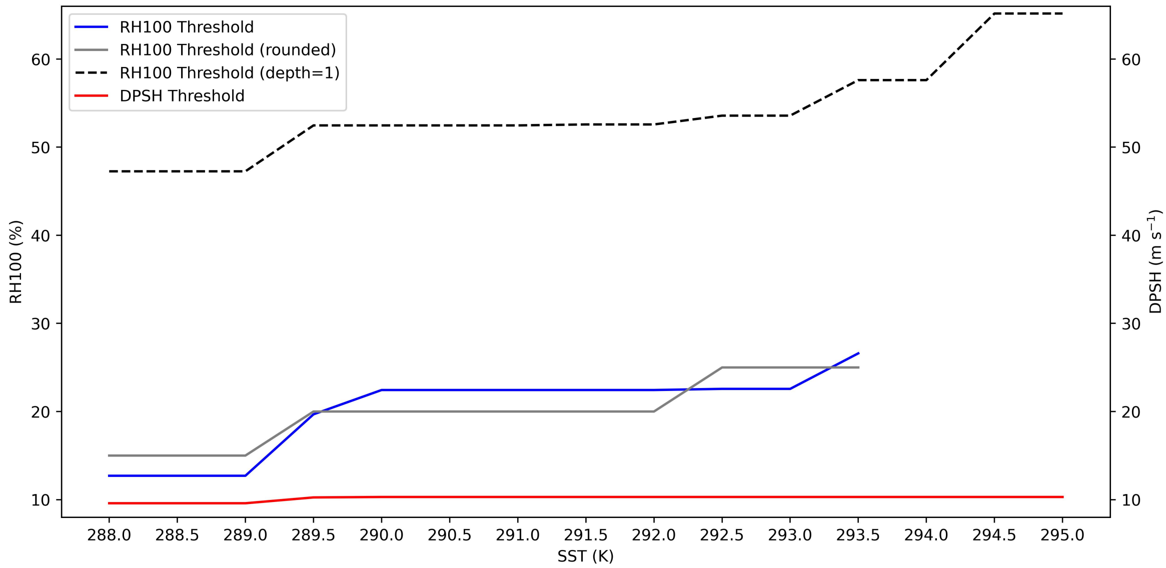

Figure S2. Sensitivity of RH100 and DPSH thresholds to the minimum SST requirement used in the decision tree classifier. See Text S2 for details.

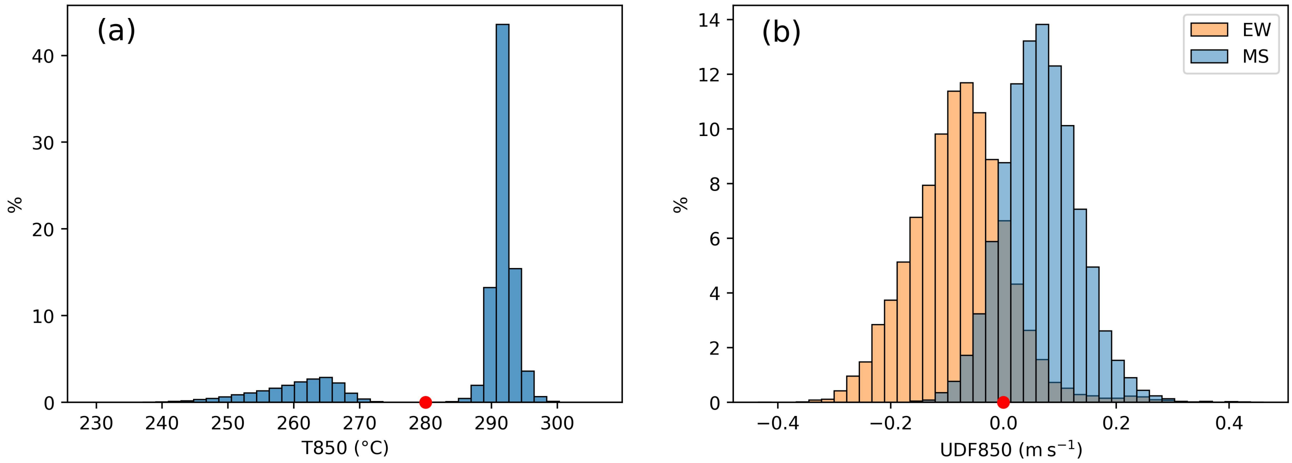

Figure S3. (a) T850 distribution of potential tropical systems that already satisfy the RH100 and DPSH thresholds of the tropical condition in the workflow, and (b) percentage distribution of UDF850 for the matched MS dataset (blue) and EW dataset (orange). Red dots indicate the thresholds we choose for MS classification.

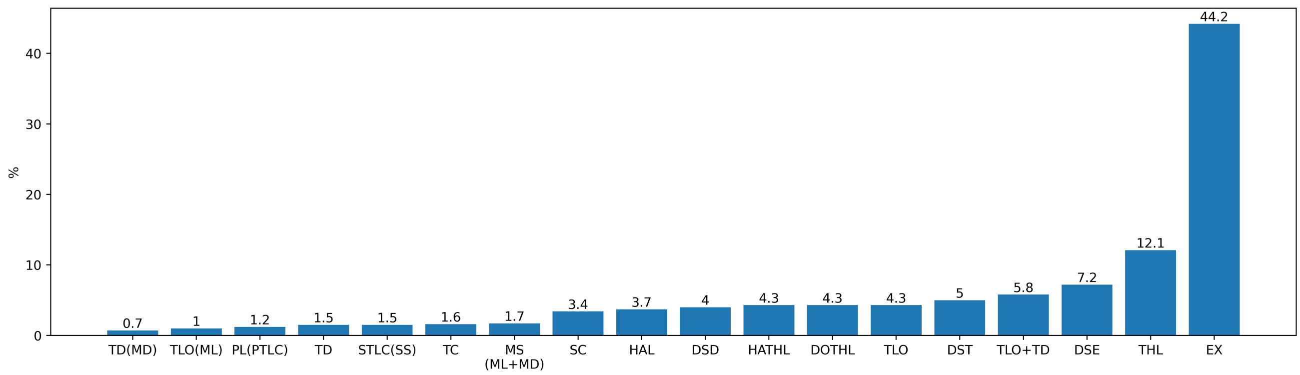

Figure S4. Frequencies (in percentages) of each type of LPS labels in our classified LPS dataset.

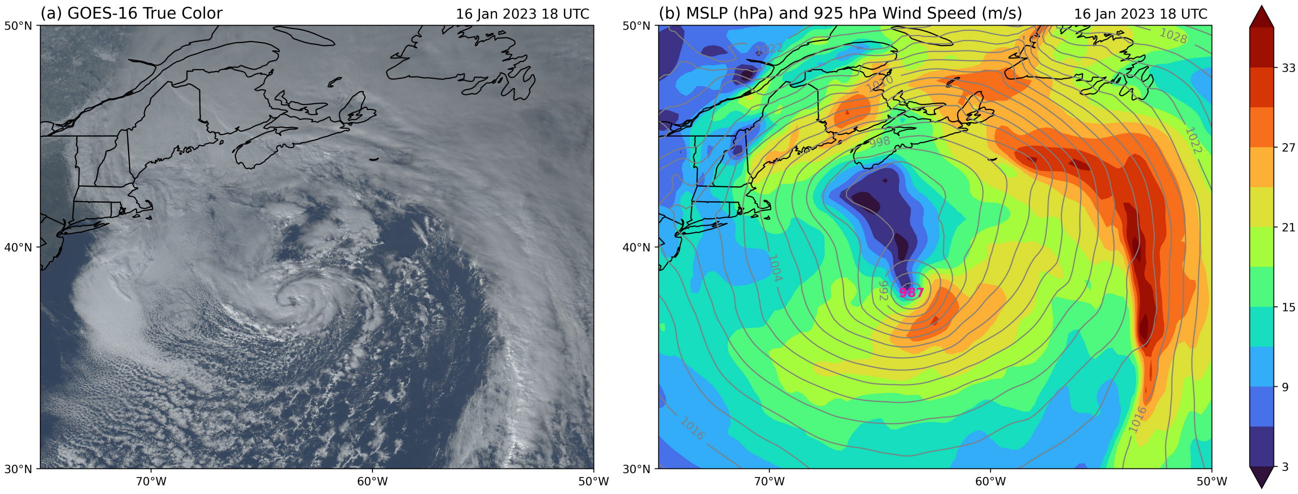

Figure S5. An example of an "embedded" TLC on Jan 16, 2023 in the North Atlantic. This storm was later recognized as a notable off-season subtropical storm in NHC's 2023 best track, indicating its potential to become a "real" TC. (a) shows its visible satellite image near peak intensity. A developed eyewall structure can be clearly seen. Its 925 hPa wind and MSLP field during the same time are shown in (b). We can see that its 987 hPa core (the TLC portion) was embedded within a synoptic-scale cyclonic circulation, where winds of near-gale scale were widespread. A weaker MSLP minima was also developing to the north of the TLC at this time, kicking off a twin-cyclone system.

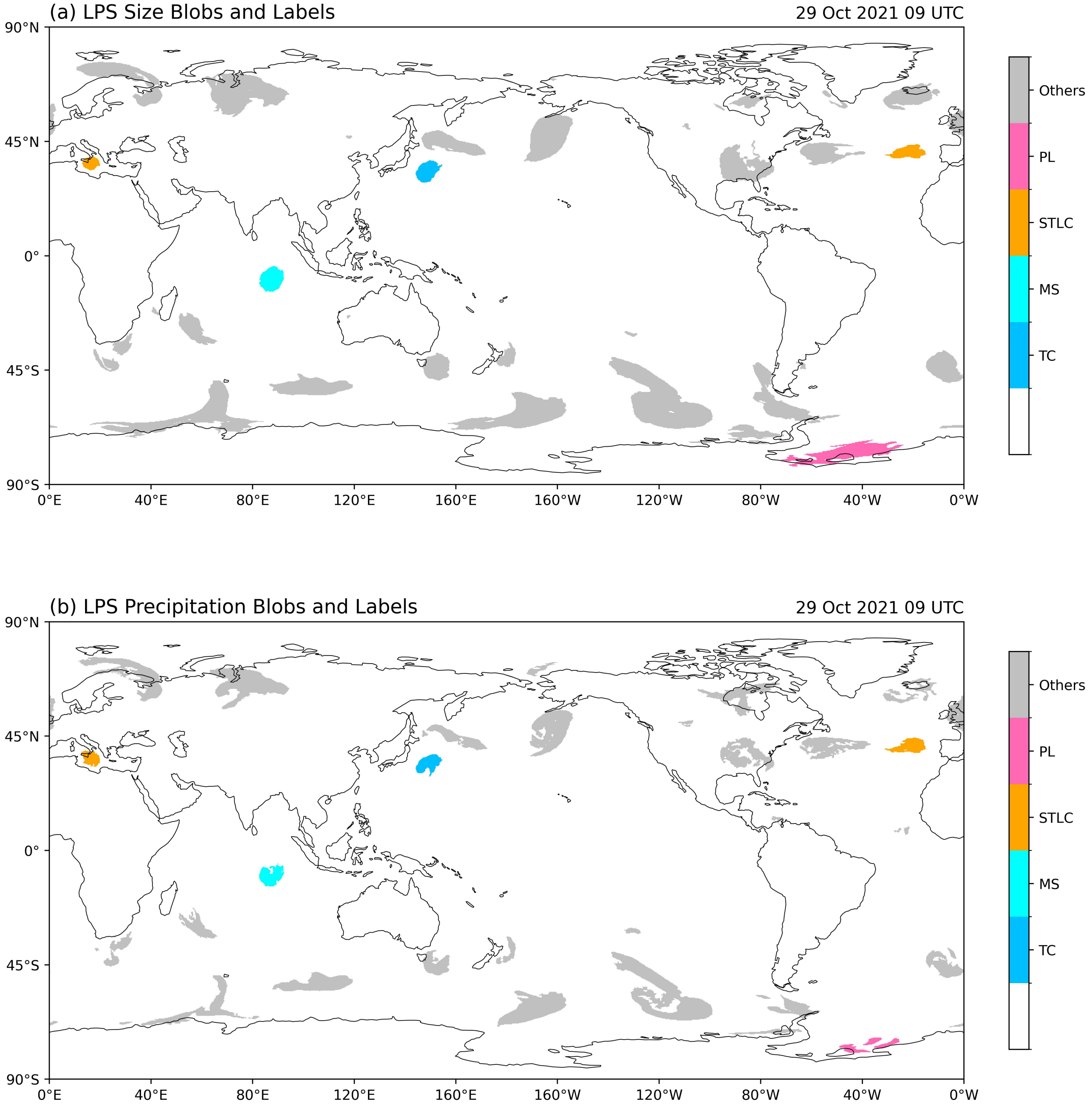

Figure S6. (a) An illustration of LPS size blobs and (b) precipitation blobs with LPS labels when they are paired with labeled LPS nodes and tracks. We choose a time slice that contains all four types of high-impact LPSs, which include 2021 Typhoon ”Malou” and 2021 Cyclone “Appolo,” a.k.a. Medicane (Mediterranean hurricane) “Nearchus.” This labeled ERA5 LPS size and precipitation blob files are available via Zenodo (see text S6 for details).

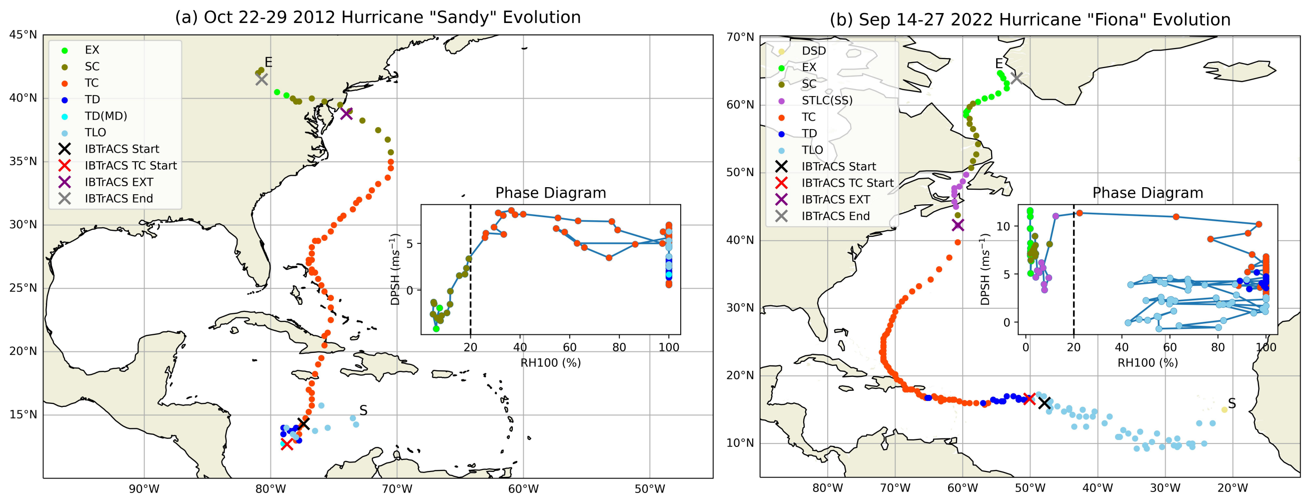

Figure S7. As in Fig. 10 of Sec. 6, but for two interesting TC cases in the North Atlantic basin.

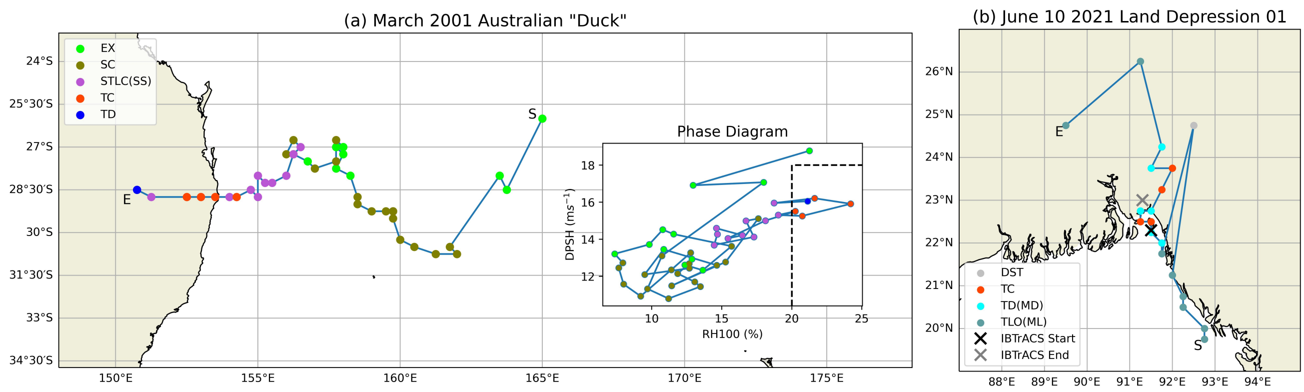

Figure S8. As in Fig. 10 of Sec. 6, but for two marginal TC cases. (a) is a classic tropical transition case happened near Australia (the "Duck"), and (b) shows an example of a South Asia monsoon system recorded as a land depression by IMD in IBTrACS. They both labeled "TC" four times by our classification, but are not recorded as TCs in IBTrACS, potentially because they are too transient to be realised by the agencies and that they are close to or over land as TCs.

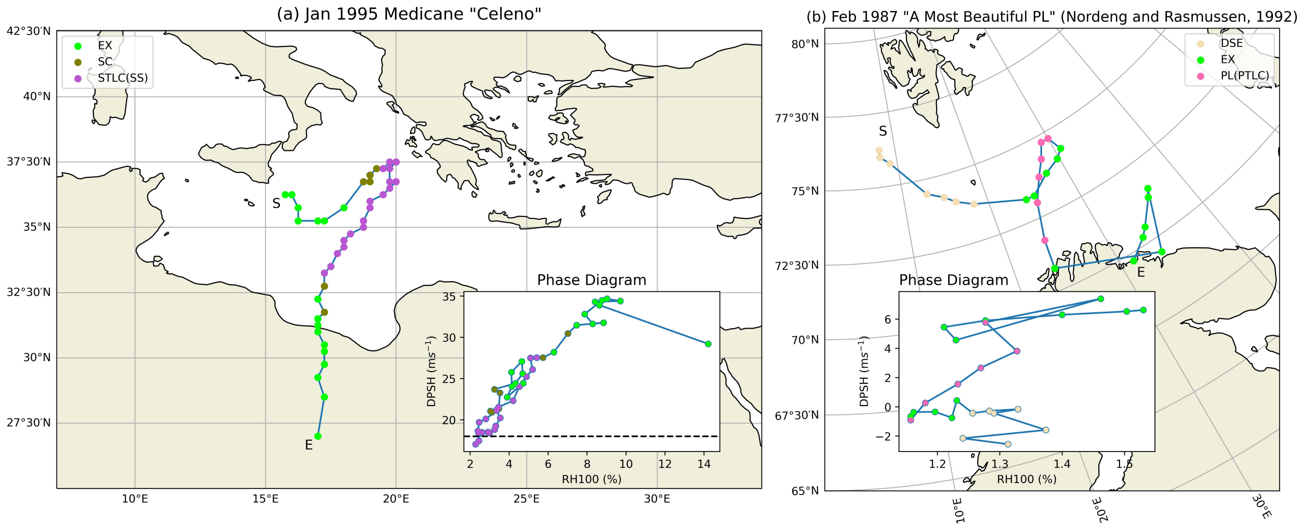

Figure S9. As in Fig. 10 of Sec. 6, but for two TLC cases. (a) is a well-studied 1995 medicane, as documented in Pytharoulis et al. (2000), Emanuel (2005), and others. (b) is a polar low described as “a most beautiful polar low” in Nordeng and Rasmussen (1992), as shown by its beautiful eye on satellite images before making landfall.

Table S1. Variables Needed for Classification

Variable Name	Pressure Level (hPa)
U-component Wind (U)	925, 850, 200
V-component Wind (V)	925, 850, 200
Temperature (T)	850
Relative Humidity (R)[a]	850, 100
Mean Sea Level Pressure (MSLP)	Sea Level
Geopotential (Z)	Surface, 925, 700, 500, 300
Relative Vorticity (VO)[b]	500

[a] Specific humidity can be converted to R with additional temperature information
[b] VO can also be computed by U and V if not directly available.