

CAMELSH – Visão geral do dataset, variáveis e estrutura

O que é o CAMELSH

- **CAMELSH** (Catchment Attributes and Hourly HydroMeteorology) é um dataset de grande amostra em escala horária para os EUA contíguos (CONUS).
- Integra séries horárias meteorológicas (NLDAS-2 e, opcionalmente, ERA5-Land), atributos e limites de bacias (GAGES-II, HydroATLAS) e observações horárias de vazão para milhares de bacias.
- Período: ~1980–2024 (horário).
- Propósito: estudos hidrológicos e de ML (alerta de cheias, previsão de vazão, comparação de modelos, etc.).

Estrutura (pastas principais)

- Hourly2/
 - NetCDF por estação (ex.: 01011000_hourly.nc).
 - Variáveis observadas: streamflow (Q), water_level (quando disponível), time.
- NLDAS-2 forcings/
 - Folder 1/
 - timeseries/ : NetCDF por gauge (ex.: 01011000.nc) com forçantes meteorológicas horários agregados por bacia.
 - attributes/ : arquivos CSV de atributos (GAGES-II e HydroATLAS) e info.csv com metadados por estação.
 - shapefiles/ : polígonos de bacias (WGS84) para junção espacial e visualização.
 - Folder 2/ (simulados/forecast para comparação):
 - NWM/ : séries do National Water Model (v3.0 e 2.1) em locais de estações USGS.
 - CNRFC_nc/ : estágios (nível) de previsão/arquivo do California-Nevada River Forecast Center.

Variáveis (séries horárias)

- Saída/Alvo (observado):
 - streamflow (Q, m³/s), water_level (quando disponível), e time / DateTime.
- Entradas meteorológicas típicas (NLDAS-2, nomes mais comuns vistos em Folder 1/timeseries):
 - Rainf (precipitação horária)
 - Tair (temperatura do ar a 2 m)
 - Qair (umidade específica a 2 m)
 - PSurf (pressão de superfície)
 - Wind_E, Wind_N (componentes do vento a 10 m)
 - SWdown, LWdown (radiação de onda curta e longa descendente)
 - CAPE (energia potencial convectiva)
 - CRainf_frac (fração de precipitação convectiva)
 - PotEvap (evaporação potencial)

Observações:

- A convenção de nomes pode variar levemente entre arquivos (ex.: PSurf vs. Psurf); verifique as chaves do NetCDF.
- Em alguns casos, as séries de chuva podem usar outros nomes (total_precipitation, precipitation).

Atributos de bacia (estáticos)

- **GAGES-II** (~439 atributos): clima, geologia, hidrologia, morfologia, paisagem, nutrientes, solo, topografia, influência antrópica (barragens, população, etc.).
- **HydroATLAS** (~195 atributos): hidrologia, fisiografia, climatologia, solos/geologia, cobertura do solo, influência antrópica.
- Finalidade: fornecer preditores estáticos para modelos (regionalização, generalização entre bacias, condicionamento físico).

Distribuição e cobertura

- Bacias: ~9.008 no total; com **vazão observada** horária em um subconjunto grande (na sua cópia local, Hourly2 contém milhares de estações).
- Tempo: séries horárias ao longo de ~45 anos (1980–2024), permitindo múltiplos ciclos úmidos/secos.
- Variáveis meteorológicas: 11 do NLDAS-2 (horárias), com cobertura completa sobre as bacias do CONUS.

Conteúdo dos diferentes arquivos e objetivo

- NetCDF em Hourly2/ : observações hidrológicas por bacia (Q, nível). Objetivo: alvo de modelos (regressão) e rótulo para alerta (classificação 2013 ex.: Flood=1 se Q excede limiar).
- NetCDF em NLDAS-2 forcings/Folder 1/timeseries/ : forçantes meteorológicas agregadas por polígono de bacia. Objetivo: preditores de entrada nos modelos.
- attributes/*.csv : atributos estáticos para análise e como features adicionais (ex.: área, declividade, clima médio, uso do solo, etc.).
- shapefiles/*.shp : geometrias para recorte, junção espacial e visualização.
- Folder 2/NWM e Folder 2/CNRFC_nc : séries simuladas/forecast (baseline/benchmark para comparação ou composição de ensembles; não são obrigatórias para começar).

Uso recomendado em ML

- **Alerta (classificação) 2013 vai/não vai alagar:**
 - Rótulo (alvo): derivado de streamflow (ex.: Flood=1 se Q > limiar p99/p99.5), opcionalmente com duração mínima (ex.: 22656 201312 h) para evitar falsos positivos.
 - Entradas: Rainf e derivados (acumulados 1/3/7 dias), demais variáveis meteorológicas, e atributos de bacia.
 - Métricas: AUC-PR, F1, precisão/recall.
- **Previsão (regressão) 2013 Q/nível em t+394:**

- Alvo: `streamflow` (e/ou `water_level`) em $t+\Delta t$.
- Entradas: janelas históricas das variáveis meteorológicas e, se desejado, autoregressão de Q .
- Métricas: NSE, RMSE, correlação (r).

Janelas temporais e splits

- **Janelas (janela deslizante):** entrada usa histórico de H horas (ex.: $H=72$ h), alvo em $t+\Delta t$ (ex.: $\Delta t=6$ h). Pode-se prever 1 passo ($K=1$) ou sequência futura.
- **Splits temporais:** treino/validação/teste por blocos no tempo, sem embaralhar, evitando vazamento do futuro.
- **Tratamento de faltantes:** mascarar NaNs, preencher apenas lacunas curtas (ex.: 3 h), descartar janelas com lacunas longas ou cruzando fronteiras de split.

Exemplos de carregamento (Python)

```
import xarray as xr

# Observações hidrológicas (Hourly2)
ds_q = xr.open_dataset('Hourly2/01011000_hourly.nc')
q = ds_q['streamflow'].values # e opcionalmente ds_q['water_level']
t_q = ds_q['time'].values

# Forçantes meteorológicas (Folder 1/timeseries)
ds_m = xr.open_dataset('NLDAS-2 forcings/Folder 1/timeseries/01011000.nc')
rain = ds_m['Rainf'].values
tair = ds_m['Tair'].values
```

Observações finais

- Para iniciar, **NLDAS-2** (Folder 1) cobre bem a modelagem no CONUS; **ERA5-Land** é opcional para comparativos.
- O Folder 2 (NWM/CNRFC) é útil como referência/baseline, mas não é requisito para a primeira fase.
- A definição de evento (limiar/duração) deve ser ajustada ao objetivo (evitar excesso de eventos e falsos positivos).