Journal of Image and Signal Processing 图像与信号处理, 2022, 11(4), 183-190
Published Online October 2022 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2022.114018

Hans汉斯

基于局部敏感哈希的天气形势场图像相似预报

闫　旭[1]，段　勇[1]，才奎志[2]，胡光亮[1]

[1]沈阳工业大学信息科学与工程学院，辽宁　沈阳
[2]辽宁省气象局气象灾害监测预警中心，辽宁　沈阳

收稿日期：2022年8月29日；录用日期：2022年9月19日；发布日期：2022年10月9日

摘　要

气象相似预报是指从海量高维历史天气数据样本中查找与某时刻天气数据样本最相似的个例，并以此为依据制作天气预报。由于搜索空间巨大，现有的相似预报方法存在检索速度慢的问题。借鉴图像检索的思想，本文提出了一种基于局部敏感哈希的天气形势场相似预报方法，利用局部敏感哈希函数将大量高维的历史形势场图像进行哈希散列，建立历史形势场的索引结构，并以相似离度作为相似性准则衡量样本之间的相异程度，得到相似个例排序，从而减少计算量。实验结果表明，该方法能够对高维海量的历史形势场进行有效搜索，在保证相似预报准确率的基础上，降低了相似预报的时间复杂度，提高了相似预报的检索效率。

关键词

相似预报，天气形势场，局部敏感哈希，相似离度

Similarity Prediction of Weather Situation Field Image Based on Locality Sensitive Hashing

Xu Yan[1], Yong Duan[1], Kuizhi Cai[2], Guangliang Hu[1]

[1]School of Information Science and Engineering, Shenyang University of Technology, Shenyang Liaoning
[2]Meteorological Disaster Monitoring and Warning Center, Liaoning Meteorological Bureau, Shenyang Liaoning

Received: Aug. 29th, 2022; accepted: Sep. 19th, 2022; published: Oct. 9th, 2022

Abstract

Meteorological similarity prediction refers to finding the individual cases that are most similar to the weather data samples at a certain time from a large number of high dimensional historical

文章引用: 闫旭, 段勇, 才奎志, 胡光亮. 基于局部敏感哈希的天气形势场图像相似预报[J]. 图像与信号处理, 2022, 11(4): 183-190. DOI: 10.12677/jisp.2022.114018

weather data samples and making weather forecasts based on them. Because of the huge search space, the current similarity prediction method has the problem of slow retrieval speed. Using the idea of image retrieval as a reference, this paper proposes a similarity prediction method for the weather situation field based on locality sensitive hashing. The method hashes a large number of high-dimensional historical situation field images with locality sensitive hashing function to establish the index structure of the historical situation field, and measures the degree of dissimilarity between samples by using similarity divergence as a similarity criterion to get the order of similar cases, thus reducing the amount of calculation. The experimental results show that the proposed method can effectively search the high-dimensional massive historical situation field, reduce the time complexity of similarity prediction while ensuring the accuracy of similarity prediction, and improve the retrieval efficiency of similarity prediction.

Keywords

Similarity Prediction, Weather Situation Field, Locality Sensitive Hashing, Similarity Dispersion

1. 引言

随着天气预报业务系统的不断发展与完善，天气预报技术已取得了突出的成果，而目前统计预报在天气预报领域是强有力的技术方案，相似预报方法则是统计预报的方法之一，是在天气预报业务中被广泛使用的方法[1]。所谓相似预报就是将某时刻的大气状态与过去出现过的所有历史大气状态逐个比较，从历史大气状态中找出若干最相似的个例，并以此为依据制作预报结果，其中未来某时刻的大气状态一般是由数值预报产品构造出的[2] [3] [4]。相似预报的效果优劣取决于是否选择了恰当的相似度匹配方法，国内外许多专家学者对相似度匹配方法进行了长期深入的研究，常用的相似匹配算法有相关系数[5]、相似系数[6]、欧氏距离[7]、海明距离[8]和相似离度[9]，其中相似离度作为考虑较为全面的方法被广泛用于相似预报之中。

但在长期的气象观测和研究中，也产生了大量高维的历史数据，因此，在用不同的方法进行相似预报时都会遇到同一个问题，即在计算大量高维矩阵的相似度过程中需要较大的时间复杂度和空间复杂度，这样将不能满足用户高效检索的需求。考虑到相似预报使用的形势场是二维格点数据，有与单通道图像相同的结构，所以可以借鉴图像检索的思想，构建检索索引结构，从而加快对天气形势场的检索速度。局部敏感哈希算法[10] (Locality Sensitive Hashing, LSH)就是在处理高维海量数据时用于构建索引结构，从而解决维数灾难问题的算法，本文将相似离度作为相似性准则衡量样本之间的相似度，结合局部敏感哈希算法设计索引结构，给出一种基于局部敏感哈希的天气形势场相似预报方法，以期实现对大量高维历史数据的高效、准确地索引与搜索。

2. 气象数据资料及可视化图像

本文以辽宁省为主要研究区域，所使用的预报资料是欧洲气象中心提供的全国范围内 1980 年到 2019 年逐日 3 小时的 ERA-5 再分析资料，均包括 6 个基本量：高度场、温度场、湿度场、经度方向风、纬度方向风、海平面气压场，所有形势场都以二维格点矩阵的形式存储在文件中，每个形势场样本都与图像

的单个通道上的像素矩阵有相似结构，这样就可以将相似预报问题当作图像检索问题处理，每个形势场矩阵均可以可视化为如图 1(a)、图 1(b)所示弱纹理、无线条的天气图。

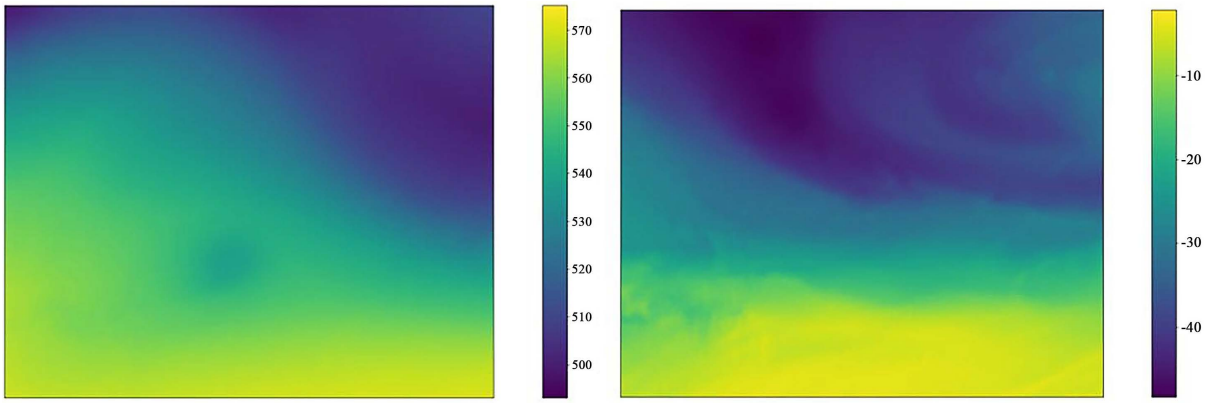

Figure 1. Visualization of height field and temperature field
图 1. 高度场与温度场的可视化图像

为了完成对地面实况数据的预测，验证预报效果，气象数据资料也需要包括与形势场数据对应的实况资料，具体包括 1980 年到 2019 年逐日 3 小时的降水量、能见度、风向风速等地面实况数据。根据以上资料实现对辽宁省范围内 62 个国家气象站的 6 小时降水量、能见度、风速等实况数据的预报。

3. 基于局部敏感哈希的天气形势场相似预报方法

基于前文所述的气象数据，本文采用如图 2 所示的相似预报流程。首先在历史场数据集合中寻找与预报场最为相似的历史个例；再将该历史个例对应的地面实况资料作为该次预报的结论。

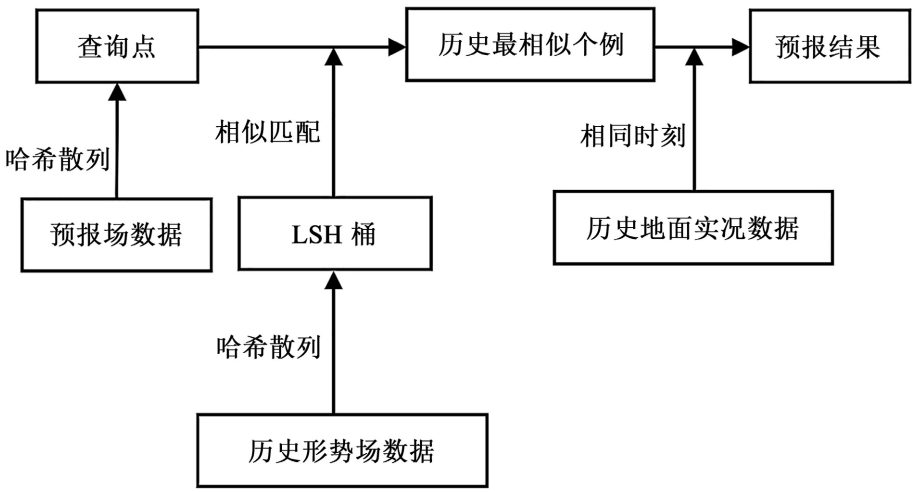

Figure 2. The retrieval process
图 2. 检索流程

从以上相似预报的流程中可以看出，相似预报的关键是找到最佳的相似匹配方法。本文首先使用局部敏感哈希函数对所有历史形势场数据进行散列并保存为局部敏感哈希(LSH)，使散列后的历史场依然保

持原来的位置关系，再对预报场数据进行同样的散列形成查询点，并在 LSH 桶中查询候选点，对候选点进行相似离度的计算与排序，找出离查询点最近的十个历史最相似个例，根据最相似个例对应的历史地面实况数据制作预报结果。

3.1. 相似时间与空间

选取适当的相似时间与空间是做好相似预报的前提条件。在相似时间方面，历年同一时段的天气形势、地面实况是相似的，因此选取历史每年与预报时刻同期前后一个月的时间为查找范围[11]，为检索过程的高效、准确打下基础，如图 3 所示。

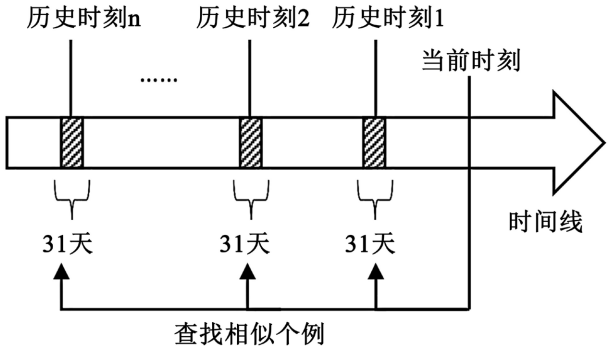

Figure 3. Time search range
图 3. 时间查找范围

在相似空间方面，形势场的对比范围过大会造成计算时间的浪费、增大误差，对比范围过小会漏掉关键信息。本文最终确定了天气系统大尺度的形势场范围为 110˚E~135˚E、32˚N~52˚N、中尺度的形势场范围为 115˚E~131˚E、35˚N~48˚N、小尺度的形势场范围为 119˚E~127˚E、39˚N~45˚N，如图 4 所示。

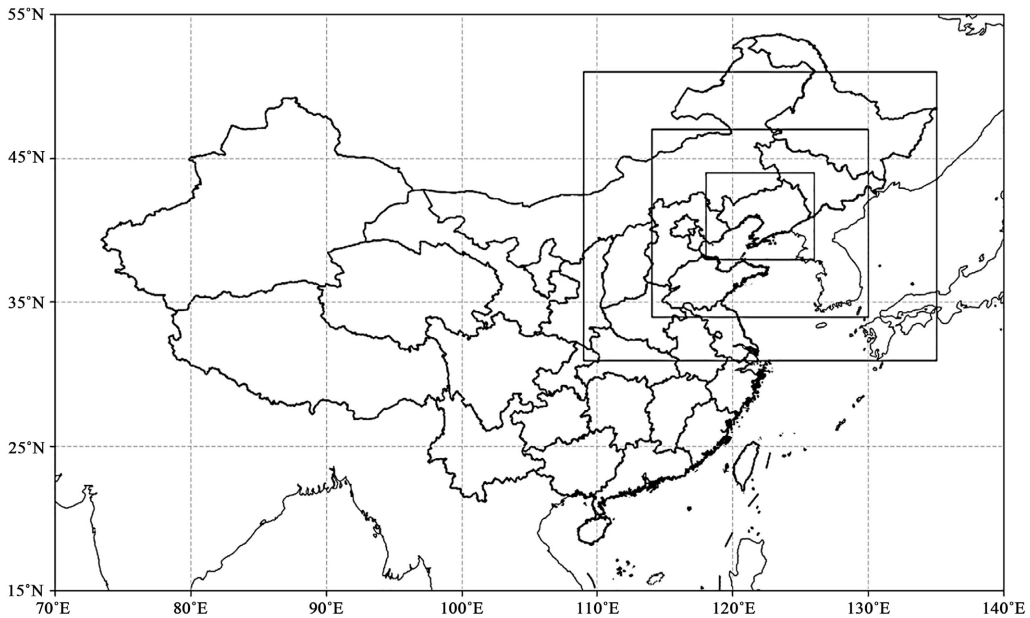

Figure 4. The situation field range of weather systems at large, medium and small scales
图 4. 天气系统在大、中、小尺度的形势场范围

3.2. 基于相似离度的相似性度量机制

相似预报方法中的相似离度算法，综合考虑了形势场之间的“形”与“值”两方面的相似[9]，是一种考虑因素较为全面的相似匹配方案，因此本文选取相似离度算法作为衡量样本之间相似度的指标。

将 m 维的形势场数据样本 i 中的每个格点按从上到下，从左到右的顺序标记为 $H_i(1), H_i(2), \cdots, H_i(m)$。计算样本 i 、样本 j 之间的相似离度，首先需要计算两个样本在 $k(1 \le k \le m)$ 位置上的“值”的差异 $H_{ij}(k)$：

$$H_{ij}(k) = H_i(k) - H_j(k) \tag{1}$$

样本 i 、样本 j 之间整体的“值”的 D_{ij} 通过公式(2)计算得到。

$$D_{ij} = \frac{1}{m}\sum_{k=1}^{m}\left|H_{ij}(k)\right| \tag{2}$$

可以使用平均离散程度来表示样本 i 、样本 j 之间的“形”的差异 R_{ij}：

$$R_{ij} = \frac{1}{m}\sum_{k=1}^{m}\left|H_{ij}(k) - E_{ij}\right| \tag{3}$$

其中， E_{ij} 通过公式(4)计算得到。

$$E_{ij} = \frac{1}{m}\sum_{k=1}^{m}H_{ij}(k) \tag{4}$$

综合考虑样本 i 、样本 j 之间“形”与“值”的相似性，将相似离度定义为公式(5)。

$$C_{ij} = \frac{\alpha R_{ij} + \beta D_{ij}}{\alpha + \beta} \tag{5}$$

其中 α 、 β 分别为“形”与“值”对相似程度的重要程度，即权重参数。

3.3. 基于局部敏感哈希的天气形势场相似预报方法

在当前相似预报不断发展的背景下，历史样本规模增长迅速、数量庞大、维度较高，导致传统的穷举搜索所有数据的方法计算耗时较高、检索速度较慢、效率较低，只使用相似离度作为相似性准则衡量形势场样本之间的相异程度将无法满足用户需求。在保证准确率的同时快速完成检索任务成为了新的需求，为此，本文提出了基于局部敏感哈希的天气形势场相似预报方法，算法过滤了多数与预报形势场样本不相似的历史形势场样本，极大的缩小了搜索范围，减少了冗余的相似度计算，进而实现快速检索。

局部敏感哈希算法的思路是，选取合适的局部敏感哈希函数，将相似时间之内的所有历史形势场进行哈希散列并存储，让相邻位置的形势场以较大概率映射到同一个哈希值，使散列后的历史形势场依然保持原来的位置关系，在检索时只需要根据索引表生成候选集，顺序计算检索样本与所有候选样本之间的相似离度即可，不须将预报场与每个历史场一一对比，以此节约大量的计算时间。局部敏感哈希的定义[12]如下：

一个映射函数族 $H = \{h : S \to U\}$，对原始数据 m 、n 做映射，得到对应的哈希值 $h(m)$ 、$h(n)$，$d(m,n)$ 为 m 、n 之间的距离，表示两个对象之间的相异程度，$P[h(m) = h(n)]$ 表示 m 、n 的哈希值相等的概率，如果满足条件(6)和条件(7)则认为这样的一族哈希函数 $H = \{h : S \to U\}$ 是 (r_1, r_2, p_1, p_2) 敏感的。

$$\text{若} d(m,n) < r_1, \text{ 则} P[h(m) = h(n)] \ge p_1 \tag{6}$$

$$\text{若} d(m,n) > r_2, \text{ 则} P[h(m) = h(n)] \le p_2 \tag{7}$$

其中，r_1、r_2为距离常量，p_1、p_2为概率常量。

由以上定义可知，当两个形势场样本足够相似时大概率会映射到同一个哈希值，如果不相似映射到同一哈希值的概率就足够小，构建历史形势场的局部敏感哈希索引的过程如图 5 所示。

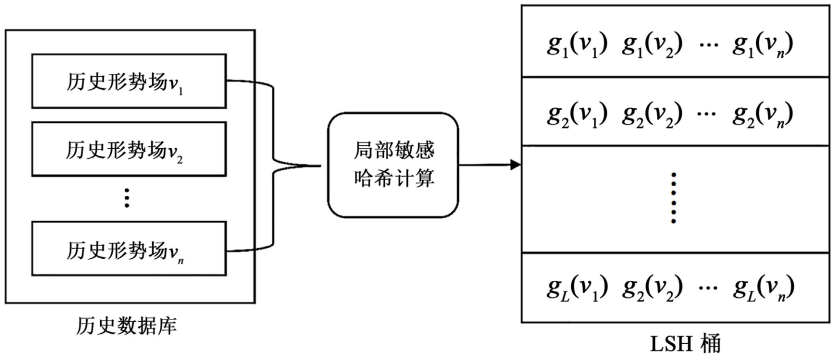

Figure 5. Indexing
图 5. 建立索引

将每个历史形势场数据都串接为一维形势场向量v，构建k个哈希原子函数，定义一个函数组$\zeta=\{g:S\to U^k\}$，每个形势场向量v的$g(v)$如公式(8)所示，从ζ函数组中独立、随机的选取L个函数$g_1,g_2,\cdots,g_L$，对所有历史形势场向量进行哈希散列，计算$g_1(v),g_2(v),\cdots,g_L(v)$，生成索引表，并存入 LSH 桶中。

$$g(v)=(h_1(v),h_2(v),\cdots,h_k(v)) \quad h_i(v)\in H \tag{8}$$

对于预报形势场q，计算q的L个局部敏感哈希值$g_1(q),g_2(q),\cdots,g_L(q)$，生成查询点，并在 LSH 桶中搜索，两个节点之间只要有一个对应位置的哈希值相同，就认为这两节点对应的向量是相等或相似的，从 LSH 桶中取出与查询点邻近的所有节点后，得到候选集，候选集中保存着所有可能结果，最后使用相似离度计算预报形势场与候选集中每个历史形势场之间的相似距离，顺序搜索相似个例，得到排序结果，搜索过程如图 6 所示。

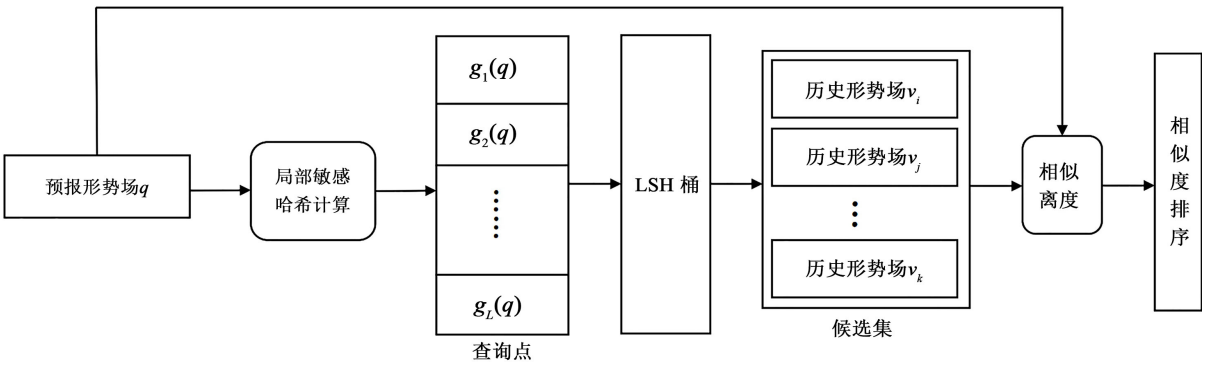

Figure 6. The search process
图 6. 搜索过程

4. 实验结果及分析

本次实验将 2019 年的样本作为预报样本，将 1980 年至 2018 年的样本作为历史样本，对辽宁省 62

个站点的降水情况做预报。这些历史形势场数据维度为 82,091，每次预报的相似时间之内的历史样本数量为 1170 左右，使用局部敏感哈希将历史形势场数据散列为一个 LSH 桶，得到候选集后，针对候选数据计算相似离度，主要的实验步骤如下：

- Step 1：构建柯西分布下的具有局部敏感性的哈希函数；
- Step 2：根据给定参数，选择 $g_1, g_2, \cdots, g_L$；
- Step 3：将预报场的相似时间之内的所有历史形势场作为训练数据，使用构建的局部敏感哈希函数对训练数据进行哈希散列，形成一个 LSH 桶，并保存；
- Step 4：使用同样的局部敏感哈希函数对预报场进行哈希散列，将其作为查询点，在 Step 3 中的 LSH 桶中搜索查询点的全部邻近点，得到候选集，再对候选集内的候选数据进行顺序的相似离度计算与排序；
- Step 5：利用距离排序计算地面实况，制作预报结果。

局部敏感哈希的天气形势场相似预报方法的优势在于，在保证与相似离度算法具有相似的准确率的条件下提高检索效率，降低时间复杂度，节约用户的检索时间。为了验证方法的可用性，在相同的软硬件环境下进行了以下对比试验，首先使用本文方法对多个预报样本进行预报，再使用相似离度法对同样的预报样本进行预报，两次实验的检索量相同，获取到的检索时间如表 1 所示，从表 1 可以看出相对于相似离度，本文算法检索时间更短，平均检索时间有所减少。

Table 1. Retrieval time and average retrieval time of samples (s)
表 1. 检索时间及样本平均检索时间(s)

预报方法	检索时间	样本平均检索时间
本文方法	228.015	0.194
相似离度	371.513	0.317

以 2019 年 1 月 16 日 2:00 的高度场、2019 年 8 月 14 日 11:00 的温度场检索结果为例，叠加了等值线的检索结果如图 7、图 8 所示，可以看出基于局部敏感哈希的天气形势场相似预报方法能够在一定程度上保证形势场等值线走向、数值分布的相似性。

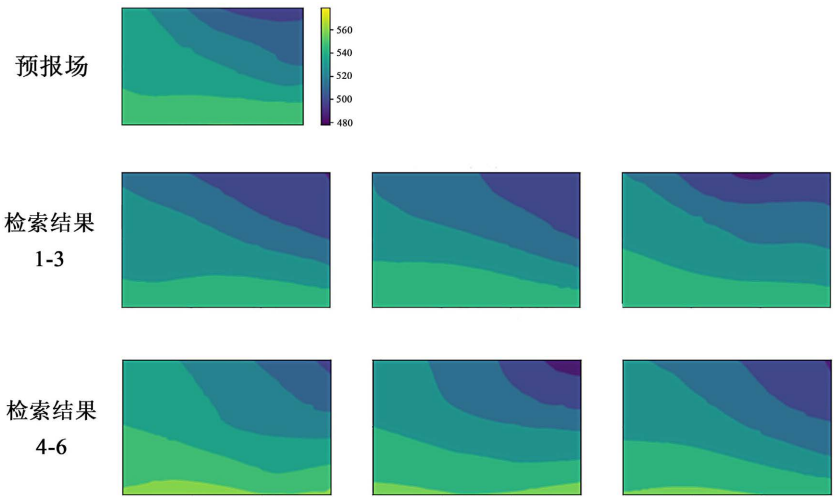

Figure 7. Similar results of 500 hPa height field
图 7. 500 hPa 高度场相似结果

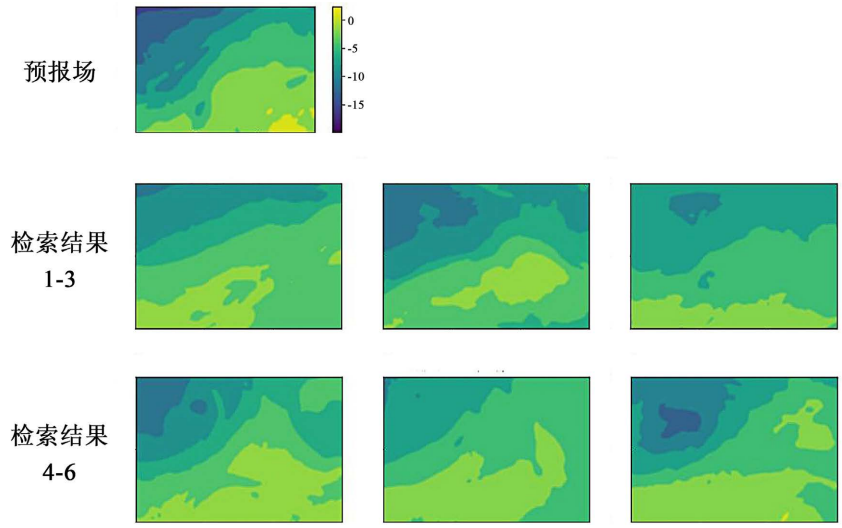

Figure 8. Similar results of 500 hPa humidity field
图 8. 500 hPa 湿度场相似结果

5. 结论

本文将局部敏感哈希与相似离度结合起来，提出了一种基于局部敏感哈希的天气形势场相似预报方法，并使用该方法对多级降水进行试预报，验证了本文提出的方法的有效性。通过一系列理论分析和实验研究可以确定局部敏感哈希算法在天气形势场搜索过程中有较好的性能，在处理高维天气形势场时有较低的时间复杂度，故该方法对预报效率的提高具有重要意义，有较高的使用价值。

参考文献

[1] 张延亭. 逐步引进因子场作相似预报[J]. 气象, 2000, 26(3): 22-27.
[2] 刘勇. 综合相似预报法在短期暴雨预报中的应用[J]. 气象, 1996, 22(10): 31-34.
[3] 马玉坤, 赵中军, 王玉国, 等. 环渤海地区云量的动力过程相似预报方法[J]. 兰州大学学报(自然科学版), 2011, 47(4): 38-43.
[4] 许炳南, 周颖. 贵州春季冰雹短期预报的高空温压场相似法[J]. 高原气象, 2003, 22(4): 426-430.
[5] 李南声. 用网格点相关系数法选相似形势场——对天气图经验预报的模拟[J]. 气象, 1982(1): 16-17.
[6] 阎惠芳, 李社宗, 黄跃青, 等. 常用相似性判据的检验和综合相似系数的使用[J]. 气象科技, 2003, 3(4): 211-215.
[7] 王明洁, 张志秀, 白人海. 相似方法在夏季降水预报中的应用[J]. 黑龙江气象, 2000(3): 20-21.
[8] 罗阳, 聂新旺, 王广山. 几种统计相似方法的适用性比较[J]. 气象, 2011, 37(11): 1443-1447.
[9] 李开乐. 相似离度及其技术[J]. 气象学报, 1986, 44(2): 176-183.
[10] 董小灵. 局部特征的局部敏感哈希专利二值化图像检索[J]. 电视技术, 2022, 46(5): 54-60+66.
[11] 刘金科. 基于气象历史数据的中期降水相似预报研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2021.
[12] 吴家皋, 王永荣, 邹志强, 等. 局部敏感哈希图像检索参数优化方法[J]. 计算机技术与发展, 2020, 30(1): 32-37.